Intelligent Methods for Signal Processing and Communications

Baiona (Vigo), Spain, June 1996
Edited by D. Docampo, A.R. Figueiras and F. Pérez

COLECCIÓN: CONGRESOS

SERVICIO DE PUBLICACIÓNS

UNIVERSIDADE DE VIGO

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	23 February 1997	Conference Proceedings

4. TITLE AND SUBTITLE

Intelligent Methods for Signal Processing and Communications

5. FUNDING NUMBERS

F6170896W0195

6. AUTHOR(S)

Conference Committee

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

ETSI Telecom-Universidad de Vigo
Vigo 36200
Spain

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

EOARD
PSC 802 BOX 14
FPO 09499-0200

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

CSP 96-1048

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

A

13. ABSTRACT (Maximum 200 words)

The Final Proceedings for Workshop on Intelligent Methods in Signal Processing and Communications, 24 June 1996 - 26 June 1996

The Topics covered include: signal processing and communications

14. SUBJECT TERMS

Communications

15. NUMBER OF PAGES

231

16. PRICE CODE

N/A

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Universidad de Vigo

Intelligent Methods for Signal Processing and Communications

Baiona (Vigo), Spain, June 1996

Edited by D. Docampo, A.R. Figueiras and F. Pérez

COLECCION: CONGRESOS

INTELLIGENT METHODS FOR SIGNAL PROCESSING AND COMMUNICATIONS

Baiona Workshop

Edición:
Servicio de Publicacións da Universidade de Vigo, 1996

Imprime:
Gamesal, Torrecedeira, 92. VIGO

IS.B.N.: *84-8158-043-0*

Depósito Legal: *VG-724-96*

Fourth Baiona Workshop on Intelligent Methods for Signal Processing and Communications

Based on the proceedings of a conference organized by the University of Vigo, held at Baiona in June 1996.

Edited by

D. DOCAMPO
Universidad de Vigo

A.R. FIGUEIRAS
Universidad Carlos III de Madrid

F. PÉREZ
Universidad de Vigo

UNIVERSIDAD DE VIGO · 1996

FOREWORD

These Proceedings include the contributed papers presented at the Fourth Baiona Workshop on Intelligent Methods for Signal Processing and Communications, held in Baiona, Spain, June 24–26, 1996.

Six technical sessions were organized around the following topics:

I Neural Networks and Non-linear Modeling.

II Image Processing.

III Array Processing and Channel Identification.

IV Adaptive Systems in Communications.

V Emergent Techniques.

VI Hardware and Software Implementation.

We hope that the arrangements we made have benefited from our previous experience in organizing such kind of events; as in past occasions, the good atmosphere at the Conference was again facilitated by the interest of the contributions, the active involvement of all the participants, and the nice place where we met. Thank to all the authors, since they provided the essential material, creating the conditions for repeating the event, getting even higher quality and participation.

One of the main objectives of this workshop was to serve as a starting point from which further applied research initiatives can appear. Therefore, the involvement of the different social and economic organizations is of paramount importance throughout the process. We very much appreciate the understanding and support of the following Spanish institutions:

- CICYT: Acción Especial TIC95-1340-E
- Xunta de Galicia
- Universidad de Vigo
- FEUGA
- IEEE-Rama Española

We also wish to thank the United States Air Force European Office of Aerospace Research and Development and the Office of Naval Research, Europe, for their contribution to the success of the conference.

We finally express again our invitation to researchers both from the Universities and Research Centers and from the industrial arena to participate in these events since we truly believe that they will appreciate the interest of being active at these kind of Workshops, in which the state-of-the-art is revised, the latest applications discussed, round tables about hot topics maintained, international cooperation fostered, and interesting contacts made.

Domingo Docampo
Aníbal R. Figueiras
Fernando Pérez

CONTENTS

Neural Networks and Non-linear Modeling

Image Processing

Array Processing and Channel Identification

Adaptive Systems in Communications

Emergent Techniques

Hardware and Software Implementation

Neural Networks and Non-linear Modeling

EFFICIENT REGULARIZED RBF PREDICTION FOR CELP-TYPE SPEECH CODERS

Fernando Díaz-de-María and José L. Sancho-Gómez

ETSI Telecomunicación - Universidad Politécnica de Madrid
Ciudad Universitaria, 28040 Madrid, Spain
Ph: 341.549.5700, Ext. 249; Fax: 341.336.7350
emails: fdiaz@gtts.ssr.upm.es and jlsancho@gtts.ssr.upm.es

ABSTRACT

In this paper we present a new speech coder whose main characteristic is the use of a nonlinear predictor consisting in a cascade of a RBF (Radial Basis Function) network and a linear filter. The most significant disadvantage of this coder in comparison to CELP algorithms is the increase in computational cost. Here, we study the possibility of diminishing noticeably this drawback by reducing the nonlinear analysis frame length from which the predictor is estimated. Some simulations reveal how the computational cost can be reduced by a factor of three without compromising performance.

1. INTRODUCTION

Neural Networks (NN) constitute a well established technique for signal processing due to their outstanding properties: nonlinearity, learning from examples, generalization ability, parallelism, fault tolerance, easy VLSI implementability, etc. The use of NN usually involves two phases: learning and retrieving. The learning phase is computationally expensive, but in most cases is carried out in an off-line way; while the retrieving phase can be performed efficiently in real time with an appropriate hardware implementation.

Speech signals exhibit both nonlinearity and nonstationary. NN are inherently nonlinear; however, their adaptation ability to the time varying characteristics of the input in an on-line manner is strongly limited by the computational burden associated to their learning. This problem has already been evidentiated in a recent work [1].

Our work focuses on the application of adaptive neural network-based prediction to speech coding. In the previously cited paper, the application to speech coding was considered, and a new modular and recurrent network efficiently operating on a sample to sample basis was presented. Such a network predictor was reported to work properly in waveform coding algorithms at high bit rates [2]. Our investigation centers on coding schemes working at medium bit rates, which operate on a block to block basis (this is also valid for low bit rates-operation schemes). In our previous works [3][4] we proposed a cascade of a Radial Basis Functions (RBF) network and a linear filter as a nonlinear predictor for CELP-type coders; nevertheless, in that scheme quality was improved compromising computational cost, which should be reduced to get practical coders. Other related works [5][6] do not address the problem of computational complexity.

In this paper we present a new version of our previous work directed to reduce its complexity by showing the *performance - complexity* balance and proposing a way to improve it. As it will be shown, the computational effort of training in this block-processing approach is proportional to the block (hereafter, frame) length. As a result, this frame length, which in linear schemes has only proportional impact in the correlation coefficient calculation and in the buffering delay (provided that it is small enough to avoid averaged results and large enough to include a representative number of samples), plays an important role in computational efficiency of nonlinear approaches.

The paper is organized as follows: in Section 2, a reduced analysis frame length is proposed to improve the efficiency of network-based block-adaptive predictors for speech coding, and its benefits and limitations are discussed. In Section 3 the reasons for using a cascade of RBF Network and a linear filter as predictor are explained. An overview of a CELP-type coder including this type of predictors is given in Section 4; also here, some simulations are presented which reveal that the proposed approach significantly reduces the computational complexity without compromising performance. Finally, in Section 5 conclusions about this investigation are given and ongoing work is outlined.

2. EFFICIENCY CONSIDERATIONS FOR NETWORK-BASED BLOCK-ADAPTIVE PREDICTORS

The cost function used to train a predictive network usually contains a principal term of the form

$$\sum_{n=0}^{N-1} (x_{n+1} - f(\mathbf{x}_n))^2 \qquad (1)$$

where $\mathbf{x}_n$ is the vector of previous samples and x_{n+1} is the sample to be predicted, $f(\cdot)$ is the predictive mapping, and N is the frame length. The dimension of the vector $\mathbf{x}_n$ is the predictor order, i.e., the number of previous samples used for prediction. As equation (1) shows, the squared difference between the actual value and its prediction is evaluated for every sample of the frame. Usually, these cost functions are minimized by a gradient descent algorithm which is iterated through the training set a number of times (epochs). The gradient expressions will involve the term corresponding to each sample; all these terms should be evaluated in a sequential way. Therefore, the training cost will be proportional to N.

In a parallel hardware implementation of the algorithm, the frame size, if it is large, renders the most important contribution to the final computational cost in the learning phase, since each algorithm's iteration through the frame cannot be processed in parallel.

In this work we study the possibility of reducing the frame length from which the predictor is trained, while preserving good generalization properties and, eventually, good performance. However, there can be found some constraints when reducing the frame length: a very short frame means a very reduced training set, and very likely not large enough to reveal the regularities of the signal; from other perspective, a reduced training set leads to poor generalization.

Another obvious way of reducing the computational cost of the learning phase is optimizing as much as possible the number of epochs used. The reason for this is that computational resources should not be wasted in improving a convergence degree from which further refinements have not significant effects on the prediction (coding) performance.

3. REGULARIZED RBF-BASED HYBRID PREDICTORS

The RBF network is a single-layer network which computes the formula

$$f(\mathbf{x}) = \sum_{i=0}^{M-1} c_i G(\|\mathbf{x} - \mathbf{t}_i\|) \qquad (2)$$

where $\{G(\cdot)\}$ are RBF, $\{\mathbf{t}_i\}$ are the RBF centers, $\{c_i\}$ are the weights of the linear combination, and M is the number of RBF used. We use Gaussian RBF

$$G(x) = exp\left(-\frac{x^2}{\sigma^2}\right), \qquad (3)$$

being σ its variance or width.

We choose the RBF network for the nonlinear prediction task for two main reasons:

- the computational cost of its training is very small compared to other types of networks.
- it yields a regularized solution to the prediction problem. This means that we seek an smooth solution, which offers good predictions in the regions where training data is not available. A compromise exits between smoothness and closeness to the data, which is controlled through a regularization parameter. Specifically, the cost function to minimize takes the form

$$H[f] = \sum_{n=0}^{N-1} (x_{n+1} - f(\mathbf{x}_n))^2 + \lambda \|Pf\|^2 \qquad (4)$$

  where $\|Pf\|^2$ is a functional of the solution that introduces the smoothness constraint (in general, this functional embeds the "a priori" knowledge of f), and λ is the regularization parameter.

We train the RBF network in two stages. First, initial values for the centers and the variance are obtained by means of a fast procedure, and the output weights are determined via pseudoinverse [7]. Second, this solution is further refined by means of some epochs of a gradient descent method.

The proposed predictor is the de-coupled serial configuration shown in Figure 1. A comprehensive discussion about this configuration has been previously reported [4]. Here, we will briefly summarize the two main reasons to choose this configuration:

- to complement the linear prediction capabilities with a nonlinear contribution, instead of removing the linear basis and building a new global nonlinear solution.
- it can be easily applied to analysis-by-synthesis coders in a suboptimal way (by removing the nonlinear part from the excitation selection procedure), still providing good results.

Some experiments were performed to design the optimum value for the regularization parameter revealing

that it is convenient to switch dynamically the predictor between two possible states: with or without network (*network on - network off*) [4]. In the first state, the predictor is as shown in Figure 1 and λ takes a value around 10 (our speech signals amplitudes range is [-32768,32767]). For the second state, corresponding to a high value of λ, the network is disabled, and the predictor remains linear.

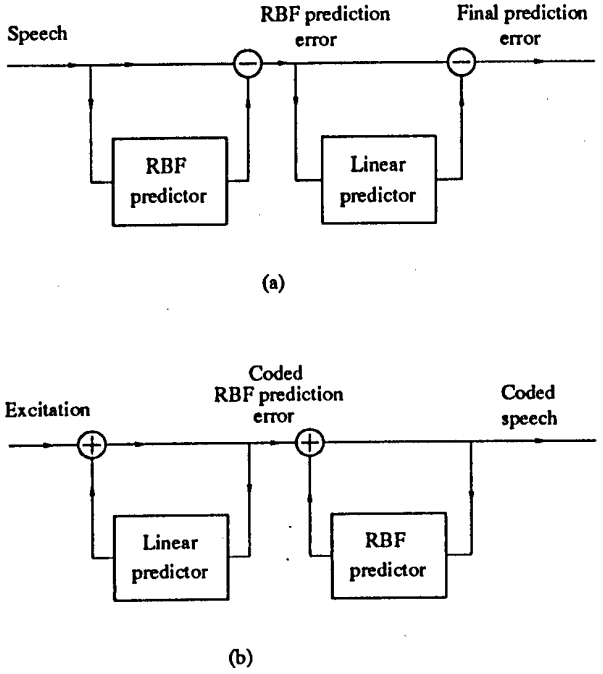


Figure 1: The predictor: de-coupled serial configuration of RBF and linear predictors. (a) Analysis system. (b) Synthesis system.

4. DESIGN OF AN EFFICIENT CENP CODER

The suggested optimization of the hybrid predictor adaption (reduction of the analysis frame length) is investigated for a low delay CELP-type coder, which we call CENP (Code-Excited Nonlinear Predictive) coder. Firstly, we present a brief description of the CENP coder; afterwards, we show the experiments carried out to improve the *performance - efficiency* balance by means of reducing the frame length; finally we draw our conclusions.

4.1. OVERVIEW OF THE PRELIMINARY CENP CODER

Figure 2 shows the block diagram of the CENP, whose main characteristics are next described. The long term predictor is carried out by means of an adaptive codebook. The stochastic codebook contains 1024 vectors. Both predictor and excitation adaptations are performed once every 2.5 ms. The prediction adaption is backward; thus the predictor parameters have not to be transmitted since they are updated from coded speech, also available in the decoder. As a result, there is no increase in the bit rate due to the inclusion of the nonlinear prediction; on the other hand, quality improvements are realized only at the cost of the computational effort required for network training. The final bit rate is in the range 8–16 kb/s, depending on the quantization scheme and the predictor and excitation updating rates.

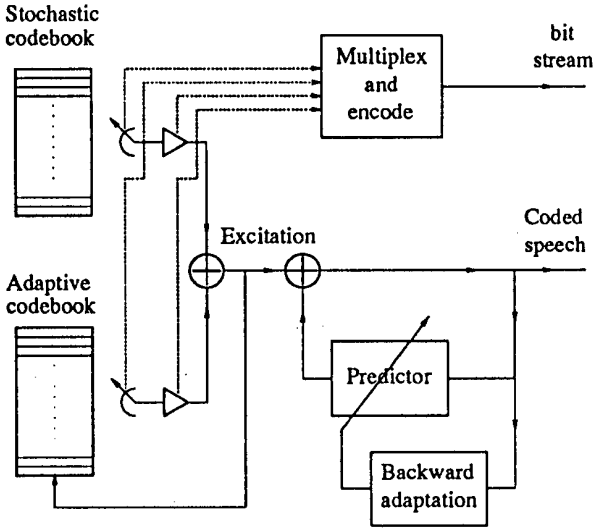


Figure 2: Block diagram of the CENP coder.

The excitation is selected as in typical CELP coders; however, in our case, the objective of the search procedure is that the output of the linear filter matches the nonlinear prediction residual (see Figure 1(b)). The perceptual weighting of the quantization error has not been included yet, leaving this issue for further work.

The decision to include or not the network is based on the results of a simplified synthesis. This synthesis is carried out in both cases (network-on and network-off) and the one which produces a coded speech closer to the original will be the chosen one. The simplified procedure is as follows: for the network-off case, the adaptive contribution (delay and gain) is determined considering all possible delays; for the network-on case,

the delay is sought only around the delay obtained in network-off case. The stochastic contribution is calculated for both cases using only a 12% of the whole stochastic codebook. For the final synthesis, the search in the stochastic codebook is completed for the winner option. Therefore, the decision is performed at the expense of an increase (around the 12 %) in computational cost of the excitation selection. Obviously, this decision procedure is suboptimum, and there is room for investigating new methods.

4.2. INCREASING THE EFFICIENCY

Initially, the procedure suggested to optimize the learning consisted in starting each frame training with the final network parameters of the previous frame [4]. This approach has an important advantage: when the predictor is frequently updated, the low variation, frame to frame, of the input signal is easily tracked, since the training starts nearer its objective. However, we have also found a significant disadvantage: sometimes, when there are rapid changes of the characteristics of the signal from one frame to the next, the training is trapped in a local minimum and it takes several frames to escape from it.

To solve this problem it is necessary to endow the training algorithm with enough agility to avoid being trapped in a local minimum during several frames by reinitiating the training every frame. After this change, the performance of the CENP has been evaluated using different number of epochs in the second stage of the training. Results evidentiate that only two epochs are enough to reach a level of performance very close to that obtained using ten or even more epochs. Therefore, the initial solution from which the gradient descent algorithm starts is very close to an acceptable solution; and even though it performs worse than the first procedure with respect to tracking slow variations, the second stage (gradient descent algorithm) might be completely eliminated, resulting in important computational savings.

The network size has been designed to maximize performance, resulting in an optimum size of 4 centers of dimension 4. Thus, the network only uses the 4 previous samples to predict the present one. This fact benefits our approach, since the analysis frame length may be shorter than that used for a 10th order prediction. It is important to remark that this length is only shortened with the purpose of training the RBF, while it is maintained to adapt the linear predictor in cascade, which has 10 taps.

We explore the possibility of reducing the nonlinear analysis frame length (it should be noted that this reduction only affects the network training). To that end, we code a moderate data base of speech signals (4 sentences of about 2 s., pronounced by four speakers; then, a total of 16 sentences) using different frame lengths. In order to avoid tying down the results to either a particular quantization scheme or a suboptimum procedure to dynamically decide if the network is on or off, the simulations are carried out without quantizing and the decision about enabling or not the network is made from the whole synthesis with both options. In these conditions, we have found that performance is maintained for frame lengths from 180 to 100 samples (the sampling rate being 8 KHz); under 100 samples, the performance begins to decrease. However, the degradation is acceptable for frame lengths of 80 and even 60 samples; specifically, for 60 samples the segmental SNR diminishes around 0.05 dB, while the computational effort is 3 times smaller than the corresponding to the preliminary length of 180 samples.

With such a reduction of the computational cost, we estimate that the network training takes 4 times as much as the computation of the coefficients of a linear predictor, providing that an appropriate hardware implementation exploits the parallelism of the network. We believe that this increase in the computational cost is acceptable, since the LPC analysis consumes only about 2% of the typical DSP computation for a CELP coding-decoding system.

Figure 3 shows the results corresponding to frame lengths of 100, 80, 60, 40, and 20 samples, for several values of the regularization parameter λ around 10. The performance reached by a CELP appears in solid line as a reference. The only difference between this CELP and our CENP is the predictor: the first one employs a linear predictor of 10 taps, while our approach uses the hybrid predictor described before. As we previously noted, the performances achieved for frame lengths larger than or equal to 60 are very similar; as a result, a 60 sample frame-length seems a good choice. Regarding the influence of λ, Figure 3 evidentiates that the best selection is $\lambda = 10$, since the performance decreases as this parameter is increased.

The performances achieved by the CENP coder working at 10800 bps. for frame lengths of 180 and 60 samples are shown in Table 1, in terms of segmental SNR. The performance obtained by a CELP working at the same bit rate is also shown for comparison. As it can be seen, any of the CENP coder versions shows an increase of around 0.4 dB in terms of segmental SNR over the results obtained using a CELP coder.

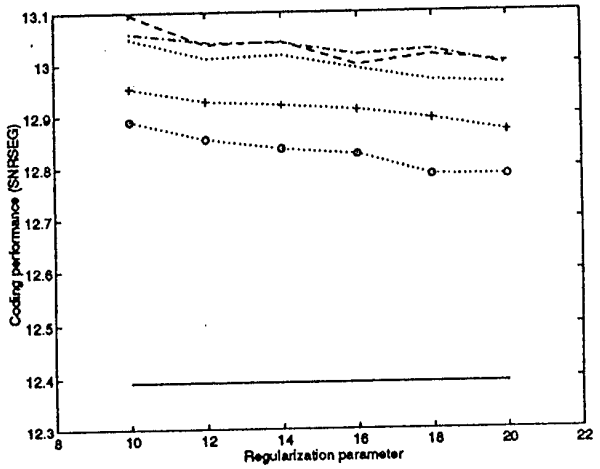


Figure 3: Coding performance vs. regularization parameter for different nonlinear analysis frame lengths (in samples): 100 (– –), 80 (–·), 60 (···), 40 (··+··), and 20 (··o··).

	CENP		CELP
Frame length	180	60	
SNRSEG (dB)	12.60	12.56	12.17

Table 1: Performances of the CENP coder at 10800 bps. for frame lengths of 180 and 60 samples. The performance reached by a CELP at the same bit rate is also shown for comparison.

5. CONCLUSIONS AND FURTHER WORK

Starting from a preliminary CENP coder using a nonlinear frame length of 180 samples, we have studied the *performance – complexity* balance stated by varying the frame length. The experiments carried out show how the frame length can be reduced by a factor of three (down to 60 samples) without compromising performance. This reduction also implies lowering down the initial computational cost accordingly.

We have also modified the initial training algorithm (which started every frame from the parameters corresponding to the previous frame) by reinitiating it every frame. Surprisingly, we have found that only two epochs of the gradient descent algorithm are enough to reach an acceptable solution. This result suggests the possibility of completely eliminating the second stage of the network training. In such a case, the reduction of the frame length will still be very valuable for optimizing the first stage.

Although the computational cost is still about 4 times above that of the linear procedure, these results are encouraging as we consider that the proposed coder can be successfully implemented in real-time with a minor improvement: as mentioned above, refining the first stage of the network training and skipping the gradient stage may results in significant savings.

The inclusion of the network in the analysis by synthesis procedure is necessary to properly incorporate a perceptual weighting filtering. With this extension, a meaningful subjective quality test can be carried out. A preliminary way could consist on obtaining the gain associated to each vector by the suboptimum procedure currently used, i.e., without taking into account the nonlinear part of the synthesis system. Then, using these gains, the excitation vector that leads to a coded speech closer to the actual one is selected.

6. REFERENCES

[1] S. Haykin and L. Li: "Nonlinear Adaptive Prediction of Nonstationary Signals", *IEEE Trans. on Signal Processing*, 43, pp. 526–535, Feb. 1995.

[2] L. Li: "Nonlinear Adaptive Prediction of Nonstationary Signals and its Application to Speech Communications", *Ph.D. Thesis*, McMaster University, Canada, 1994.

[3] F. Díaz-de-María and A. R. Figueiras-Vidal: "Radial Basis Functions for Nonlinear Prediction of Speech in Analysis-by-Synthesis Coders", Proc. 1995 IEEE Workshop on Nonlinear Signal and Image Processing, vol. I, pp. 66–69, Halkidiki, Greece, 1995.

[4] F. Díaz-de-María and A. R. Figueiras-Vidal: "Improving CELP Coders by Backward Adaptive Nonlinear Prediction", submitted to *The International Journal of Adaptive Control and Signal Processing*, Special Issue on Adaptive Techniques in Speech Processing.

[5] L. Wu, M. Niranjan and F. Fallside: "Fully Vector-Quantized Network-Based Code-Excited Nonlinear Predictive Speech Coding", *IEEE Trans. on Speech and Audio Processing*, 2, pp. 482–489, Oct. 1994.

[6] J. Thyssen, H. Nielsen and S. D. Hansen: "Non-Linear Short-Term Prediction in Speech Coding", Proc. ICASSP-94, vol. I, pp. 185–188, Adelaide, Australia, 1994.

[7] Moody, J. and C. Darken, 'Fast-learning in Networks of Locally-tuned Processing Units', *Neural Computation*, 1, 281–294 (1989).

BOUNDARY METHODS FOR DISTRIBUTION ANALYSIS

José Luis Sancho
Aníbal Figueiras-Vidal

E.T.S.I. de Telecomunicación
Universidad Politécnica de Madrid
Ciudad Universitaria
28040 MADRID España

Batu Ulug
William Pierson
Stanley C. Ahalt

Department of Electrical Engineering
The Ohio State University
Columbus, OH 43210

ABSTRACT

In this paper we introduce the use of Boundary Methods (BMs) for distribution analysis. We view these methods as tools which can be used to extract useful information from sample distributions. We believe that the information thus extracted has utility for a number of applications, but in particular we discuss the use of boundary methods for determining the suitability of a particular feature set for pattern classification. We present results which establish the correspondence of BMs and the probability of error (Pe) for normal distributions.

1. INTRODUCTION

For many investigations of physical processes, scientists and engineers must use samples drawn from the process in order to construct algorithms which model or monitor the underlying process. For example, in the telecommunications industry applications such as equalization (e.g. echo-cancellation), source-coding (e.g., video-coding using vector quantization), and detection (e.g., CDMA decorrelators) require that samples of transmitted signals be analyzed to formulate appropriate signal processing algorithms. For problems such as these we believe that the distribution analysis methods we describe will offer significant advantages in designing and fielding robust and efficient algorithms, particularly those in which classification plays a dominant role in the processing.

(*) Support for this work was provided by the AMOS Research consortium. Graduate student support was provided to Bill Pierson through the DoD Palace Knight Program. Stan Ahalt is currently on sabbatical leave at Universidad Politécnica de Madrid and gratefully acknowledges the support of the Dirección General de Investigación Cientifica y Técnica, Ministerio de Educacion y Ciencia of España.

If an investigator has a reasonably complete understanding of the physical process, a mathematical model can be constructed and the samples can be used to estimate the parameters of the model. If the number of samples are sufficient in number with respect to the dimensionality and the statistics of the problem, then the needed pdf's can be estimated, and an optimal Bayesian classifier can be constructed [1, 2].

However, in many practical situations, there are problems with this approach. First, constructing a model can be time consuming, and verification of the model can be problematic. Second, as the dimensionality of the data increases, exponentially larger numbers of samples are required to accurately estimate the class conditional probabilities. It is often either physically impossible or financially prohibitive to obtain the needed data. Thus, accurate estimates of the probability density functions can not be obtained – which implies that the Bayes error cannot be accurately estimated. Third, determining the estimates for the prior probabilities becomes especially difficult when the number of classes is large – which is common for many practical applications. One approximation is to assume a uniform distribution for the prior class probabilities [1] which simplifies the analysis. However, for optimal performance, the class probabilities need to be estimated accurately in order to apply Bayesian analysis. Consequently, for these reasons, investigators must turn to other alternatives for many practical problems.

For those cases in which the process cannot be readily modeled, either supervised or unsupervised learning can be employed. In either case, the learning process can be viewed as distribution analysis. If only unlabeled data is available, e.g., when the number of classes is unknown, unsupervised learning techniques, usually based on clustering, are used to discover a model which captures the structure of the data in the data-space.

Clusters thus formed can be evaluated, e.g., using Indices of Partitional Validity (IPVs) [3, 4], in an attempt to measure how well the clusters capture the structure of the data. Usually these indices use some combination of measures which quantify the compactness and isolation of each of the discovered clusters. While these methods have proven to be useful [5, 4], they require the use of an explicit distance metric. The choice of this distance metric can have a significant impact on the reliability or utility of the analysis.

When labeled data is available, as we assume here, it is standard practice to use mixture decomposition techniques to allocate each pattern to a particular cluster, and then estimate the cluster parameters. These techniques generally require that the number of clusters be known and adopt a density model which assumes that the clusters are multivariate normal. Mixture decomposition techniques then focus on 1) assigning each data sample to the correct cluster, and 2) estimating the mean and covariance matrices of each cluster. A particularly complete discussion of these techniques can be found in [4].

2.. MOTIVATION

Our research is focused on determining measures we believe are of significant importance to designers of pattern classifiers. In particular we are interested in:

- Classifier Independent Discriminant Measures (CIDM)s which yield a numeric result which indicates how separable two classes, or the features derived from two classes are. In particular, we are searching for CIDMs which can be directly related to probability of error (Pe) or other pertinent measures.
- Feature-Set Evaluation (FSE) techniques which, given alternative ways of deriving feature sets from observations, order those sets by classification - fitness. Of course, this measure of fitness should also be related to Pe or other pertinent measures.
- Sample-Pruning (SP) techniques which support the development of classifiers which are:
  - quickly constructed,
  - execution-efficient,
  - generalized, and
  - robust.

We observe that a CIDM would ideally analyze all data contained in the sample population and yield one value to denote how separable the two classes are. On the other hand, FSEs analyze at least two sample populations and yield two values which can be compared to determine which of the two populations consist of better features. Finally, SP techniques operate on one population and yield another population which is always a subset of the original population.

Further, to clarify the third point above, we observe that a) in order to quickly construct a classifier, we need to minimize the use of samples that provide little useful classification information, b) classifiers that are execution-efficient are those constructed such that a small number of parameters need to be evaluated in order to reach a classification decision, c) generalized classifiers reliably estimate the classification mapping using noisy training samples, and d) robust classifiers reliably estimate the classification mapping when the noise process which affects the training samples differs from the noise process which affects the testing samples.

While we do not claim to have solved any of the above problems, we believe the distribution analysis technique we discuss here, BM, does have a significant benefit in meeting the objectives of FSE and SP. In this paper, however, we restrict our discussion to the use of BM for FSE. We also note that the statistics community has an extensive literature on the use of linear methods, particularly principal component analysis. However, we are not aware of more directly related FSE work in the statistics community. Similar work in the Neural Networks and pattern recognition communities also rely on identification of individual features. An excellent review of some of these techniques can be found in [6].

3. GENERAL DESCRIPTION OF THE BOUNDARY METHOD

This method exploits the distributions in a controlled way in order to extract useful information about how the distributions are composed relative to each other. Suppose we have a hypothetical population consisting of two distributions drawn from two classes, as shown in Fig.(1)(a).

We enclose the classes with a boundary, according to some criteria. In the case shown in Fig.(1)(b) we have drawn the enclosing boundaries as closed interpolated splines with approximately 15 control knots. Generally, some criteria is used in forming the boundaries so that obvious outliers are excluded and relatively compact boundaries are formed.

While the boundaries shown are quite complicated, it is reasonably easy to specify and manipulate compli-

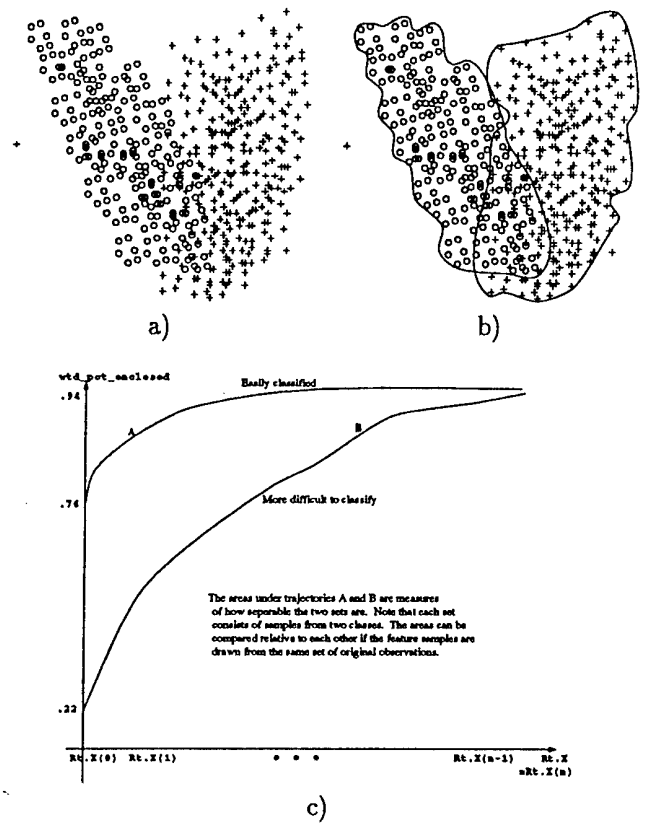


Figure 1: a) A population of two distributions drawn from two classes; b) Distributions enclosed by arbitrary boundaries; c) Trajectory of samples enclosed by increasing boundaries.

cated boundaries using methods such as interpolated splines. Indeed, the boundaries can be determined in any way that is computationally feasible, and can be of arbitrary shape - with suitable modifications to the basic algorithm. In the examples presented later we use elliptical boundaries since they are very simple to specify and manipulate.

We then collapse the boundaries until the boundaries are just touching. For complicated boundaries, the amount of shrinking necessary to effect tangential, or "just-touching" boundaries can be relatively difficult calculate and there can be multiple possible solutions. However, for many boundaries - such as the one shown - a straight-forward search process over the scale of the boundary, e.g., about the center point, can be used to establish the tangential boundaries. In any case we always establish a canonical starting point at which the boundaries enclose the most samples from each class, but without overlap of the volumes enclosed within the boundaries.

We choose to use tangential boundaries for the following reasons. First, shrinking the boundaries until they are tangential establishes a specific point to begin our calculations and allows us to normalize our results, as explained later. Second, the tangential boundaries enclose subsets of the class samples that, we believe, most reliably represent the class distributions. Finally, as briefly discussed later, tangential elliptical boundaries have a direct relationship to Fisher's LDA projection axis.

At this point we have established the two values to be used as the end-points in our final calculations: the original enclosure size, labeled Rt.X(n); and a minimum, non-overlapping size, labeled Rt.X(0).

We now begin to grow the boundaries. We grow the boundaries gradually for a number of steps, say n, that is sufficient to obtain the desired trajectory, as described below. We refer to the area under the trajectory we form as the Trajectory Area (TA). As the boundaries grow, the number of samples that are enclosed from each class increases, and the number of samples that are enclosed in the region common to both boundaries increases. We keep track, via a count, of how the number of samples in either the overlap region, or within both boundaries, increases, regardless of class. We observe, however, that we can weight the samples such that the closer the samples are to the boundary the more they contribute to the count.

Once we have expanded the boundaries out to their original position, we have captured a measure of how the samples are distributed in the space as recorded in the count. We now plot the trajectory, as shown in Fig.(1)(c). In Fig.(1)(c) we have plotted two hypothetical trajectories for two different populations of distributions of two classes.

We claim that areas under these trajectories is a measure of how the samples are distributed in the following sense. The areas for two different sample populations can be quantitatively compared to one another, and the relative ordering of the areas is invariant with respect to linear transformations applied to the data distributions.

For population A the samples that are enclosed in the minimal-boundary, Rt.X(0) are relatively compact (76% are enclosed when the ellipses are tangential) and the number of enclosed samples quickly asymptotes to the number of samples enclosed by the final (i.e., the original) boundary, thus the classes are, to some degree, isolated.

In contrast, population B has a trajectory that indicates that the distributions are not very compact (only 22% are enclosed by the tangential boundaries) and the trajectory indicates that the two classes are fairly interspersed, as the trajectory climbs slowly to the final

value - hence the classes are not well isolated.

By comparing the areas (TA) under the two trajectories we have a qualitative measure of how good the two distributions are for classification. Population A, which has the bigger area, will be more easily classified than Population B. We have, effectively, a method for FSE.

4. RESULTS

In this section we show two simple experimental results using Boundary Methods. The first is shown in Fig.(2), where we have selected two test populations in which one population is more separable than the other.

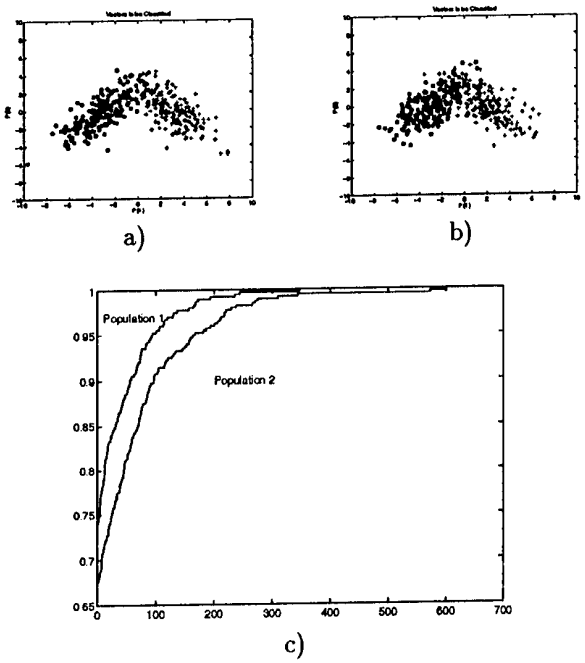


Figure 2: Figures a) and b) show two test populations (200 samples of each class) with similar separability; Figure c) shows the trajectory areas of populations a) and b). We can see they are very similar.

For this population we use elliptical boundaries, and calculate the Trajectory area (TA) using the counts of samples within the overlapping boundaries. The initial boundary for each class is formed by an ellipse, where the size of the ellipse is determined using a Chi-square test and fixing the percentages of enclosed samples. Since we know that the data consists of two classes, we have simply estimated the means and covariances directly. If the data was not unimodal we could either use a single boundary (which would result in a different trajectory) or optionally employ unsupervised techniques to determine the number of clusters-per-class to use when estimating these parameters.

The use of ellipses is attractive for many reasons. First, the assumption is reasonably satisfied for many realistic cases [6]. Second, the assumption that the data is distributed normally simplifies analysis, giving rise to ellipsoidal boundaries because Gaussian distributions have elliptical constant-density contours. Quadratic forms such as ellipses lend themselves to formal analysis and are closely tied to Bayesian formalisms. Third, ellipses are computationally attractive because they are easily manipulated as they can be specified with a mean, a covariance (matrix), and a volume. Fourth, for ellipses, the amount of shrinking necessary to effect tangential, or "just-touching" boundaries is relatively easy to calculate (although there are multiple possible solutions).

There are a number of ways the ellipses can be collapsed. We typically collapse each boundary so that the constant-density contours of each class, which fixes the volume of each ellipse, are kept equal. A simple search procedure will yield this solution quickly, since only one parameter needs to be varied. However, an alternative is to vary the density contour of each of the class distributions separately in order to determine tangential ellipses with other properties. For example, it is possible to find tangential ellipses which have equal magnitude gradients at the tangent point, and that gradient is equivalent to Fisher's linear discriminant projection vector.

Our first example, shown in Fig.(2), demonstrates how the areas under the trajectories (TA) correctly indicate that the first test population is more readily separated that the second.

Another example of the technique is shown in Fig.(3). For this case the means of the distributions are more widely separated in the data space. Note that in this case the trajectories, shown in Fig.(3)(c) and their associated trajectory areas correctly indicate that the first test population is more readily separated that the second - and that the differences in the separability between the two test populations are more pronounced than in our first example, as is visually apparent in looking at the data.

As can be seen, our preliminary tests indicate that the Boundary Method works well for these simple distributions. We are now working on more extensive tests, as well as more formal analyses.

The second experiment demonstrates the correlation between the trajectory (TA) area and the probability of error (Pe). Fig.(4) shows the correlation between TA and the Pe for a number of distributions with different means and covariance matrices. We show two

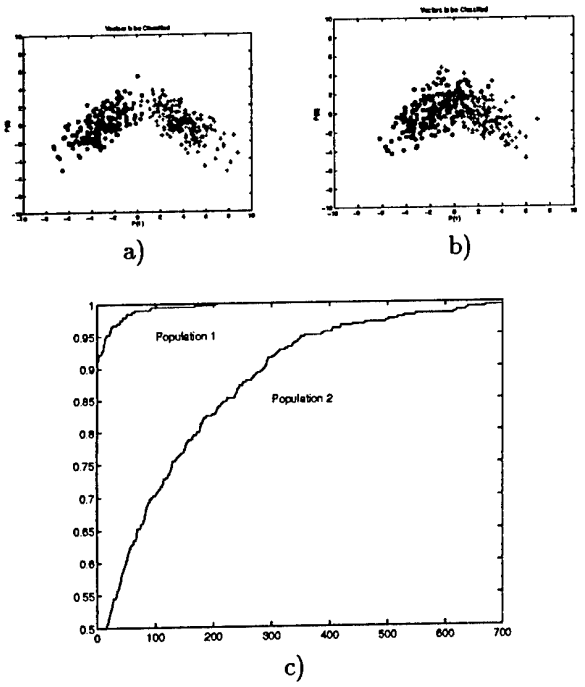


Figure 3: The Figures a) and b) show two test populations (200 samples of each class) with more widely separated means than the Figure 2; the Figure c) show the trajectory areas of populations a) and b). Note that the difference in the areas is greater in this second example, because of the more widely separated means and differing covariances.

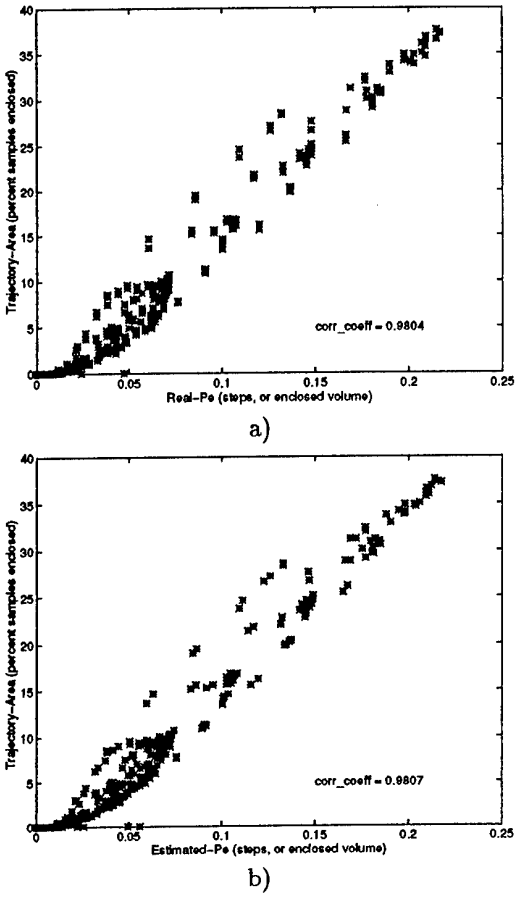


Figure 4: Relationship between the TA and the Pe for a number of distributions with different means and covariance matrices. a) Simulations with actual means and covariance matrices; b) Simulations with estimated means and covariance matrices.

equivalent simulations differing only in that Fig.(4)(a) uses the actual mean and covariance matrices, while Fig.(4)(b) uses estimated means and covariance matrices.

5. CONCLUSIONS

We have presented a new method for Feature Set Evaluation (FSE) which provides information useful in determining how separable one population is with respect to others drawn from the same source. This collection of methods is called Boundary Methods (BM) because arbitrary boundaries can be employed to investigate various separating surfaces. The relationship among the classes is captured in a number called the Trajectory Area (TA). We have given results of simulations using Gaussian distributions and elliptical boundaries which show that the BM-TA has a correlation factor of near one with the Pe.

6. REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic Press, 2nd ed., 1990.

[2] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles.* Reading, Massachusetts: Addison-Wesley Publishing Company.

[3] R. Dubes and A. K. Jain, "Validity Studies in Clustering Methodologies," *Pattern Recognition*, vol. 11, pp. 235–254, 1979.

[4] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data.* Prentice Hall Advanced Reference Series, 1988.

[5] D. Hermann, S. Ahalt, and R. Mitchell, "Clustering and Compression of High-Dimensional Sensor Data," in *SPIE International Symposium on Optical Engineering: Signal Processing, Sensor Fusion, and Target Recognition II*, pp. 286–297, April 1993.

[6] C. Lee and D. A. Landgrebe, "Feature Extraction Based on Decision Boundaries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 388–400, April 1993.

FINITE STATE AUTOMATA GENERALIZATION BY MEANS OF ANNS FOR PHONEME RECOGNITION

J. Santos[1] *and R.J. Duro* [2]

[1]Departamento de Computación
Universidade da Coruña. Facultade de Informática
Elviña s/n. 15403 La Coruña. SPAIN.
Phone:34-81-132552 Fax: 34-81-132736
e–mail: santos@udc.es

[2] Departamento de Ingeniería Industrial
Universidade da Coruña. Escuela Politécnica Superior
Esteiro s/n. 15403 Ferrol (La Coruña). SPAIN.
Phone:34-81-354282 Fax :34-81-354463
e–mail: richard@udc.es

ABSTRACT

In this work we have made use of automata theory in training Artificial Neural Networks (ANNs), with the practical application of phoneme recognition in voice signals. Our aim is to obtain Recurrent ANN structures (RANNs) that make use of the inherent sequentiality of the voice signal and at the same time are easy to train by simultaneously viewing two spaces, the output space and the state space. These spaces correspond to those of a finite state automaton, generalized by the network, that detects a given feature pattern. This methodology has been and is being applied to the construction of a modular speech recognizer for Spanish phonemes. The modularization of the recognition presents the clear advantage of facilitating training in each submodule of the network, and at the same time we are injecting knowledge through this modularization in the global structure of the recognizer.

1. INITIAL WORK

Our starting point has been the use and generation of recurrent ANNs for the processing of binary signals in a purely sequential mode. We have employed recurrent topologies in order to avoid static ontologies for solving problems in signal processing that are intrinsically dynamic. The main reason for processing the signal sequentially is to avoid or minimize the effect of the windowing process, by means of which a set of signal parameters are presented to a network as parallel inputs. The drawback of this solution is the correct selection both, of the representative parameters and of the appropriate window size.

In order to avoid these problems, instead of adopting a preset architecture and topology [1] [2], we have chosen an evolutionary strategy that determines the most adequate network structure for each particular problem. Thus, we have initially used our Genetic Algorithm based application development environment GENIAL [3] for the generation of RANNs that detect bit patterns in binary signals [4].

However, the generalization of the methodology to non binary signals presents the drawbacks of requiring a training set in which the number of analog input-output pairs may be infinite over a continuum of values, and that we must establish in this set a gradation of the similarity between each one of the inputs and the pattern to be recognized in order to establish an equivalent gradation over the response of the network. These two facts enormously complicate the search space, generating multiple local minima and deceptive points so that our genetic algorithm takes too long to find a structure for solving the problem.

This work was supported by The University of A Coruña

2. GENERALIZATION OF FINITE STATE AUTOMATA FOR THE DETECTION OF ANALOG SIGNAL SEQUENCES

Our solution has been to constrain the search space by limiting the possible network topologies. Thus, we have concentrated on structures that simulate finite state automata, automata which detect an input pattern with strict well defined rules (transition and emission matrices) [5]. In other words, we have introduced knowledge by establishing topologies that are appropriate for the increased complexity, but without losing the basic characteristics of recurrent topology and sequential processing of the inputs.

The starting idea is to obtain a non recurrent network for training that is equivalent to the finite state automaton that determines the presence or absence of a given pattern in the sequentially processed inputs. Our equivalent network must generate not only the outputs of the automaton for each input it receives, that is, each emission from the automaton, but also the internal state of the automaton for each transition. In other words, the equivalent network generates values in two different spaces, the output space and the state space, both encoded in one or several nodes of the network. The training set must contain the cases that determine the transition rules of the finite number of states and the emissions produced in each transition. In its use, however, the network is going to be recurrent when it acts over the inputs, as these inputs are made up of the signal values and the coding of the previous state, producing the current outputs and state.

For the training process, we must construct a training surface in a representation in which one of the axis is the state space and the other the output space. Now, depending on how these training surfaces are constructed, we will be able to force different types of generalizations in our networks as a function of what we wish to reflect in our outputs, allowing for a sharp detection or a similarity classification of a pattern through the construction of sharp or smooth areas in the training surface around the pattern to be recognized.

This solution has generalized the concept of finite state automaton to the consideration of a continuum of states and emissions, going from emission and transition matrices to the generation, by means of connectionist systems, of non linear emission and transition functions.

The methodology correctly solves the detection of analog patterns of small lengths. We must take into account that having a non recurrent network in training, this training procedure may be solved with the same genetic methodology as in the previous cases or using well established training algorithms for these topologies such as gradient descent. However, as the length of the patterns to be recognized increase, the two error surfaces become very complicated, as in the case of phoneme recognition, whose solution we present below.

3. USE IN PHONEME RECOGNITION

We have started to use the strategy presented above in the field of automatic speech recognition. In particular, our aim is the recognition at the phonological level, because in voice signals, each phoneme presents a characteristic pattern of parameters that evolve in time. The methodology of training overcomes the problems asociated with clasical training procedures for recurrent neural networks such as the Boltzmann machine [6] and backpropagation for recurrent networks [7].

As a first step, we have applied our methodology to the problem of recognizing the five vocalic phonemes in the Spanish language. We have applied the FFT over fixed size (8 msec) Hamming windows of signals sampled at 8 KHz. By doing this we have not lost the inherent sequentiality in the voice signal, as it is the evolution of the different windows in time what is going to determine the presence or absence of a given phoneme. This way, we may use networks simulating finite state automata, with parallel inputs whose temporal evolution represent the evolution of a given phoneme.

We have carried out the following tests:

1. Use of a RANN that processes current inputs and previous state for the recognition of the five Spanish vocalic phonemes. The network is a perceptron with two hidden layers, 16 input nodes corresponding to the frequency spectrum (between 0 and 4kHz) of each section and 10 output nodes, 5 of them represent the recognition level for each phoneme and the remaining 5 specify the state of each one of the phonemes throughout the evolution (Figure 1).

The training surfaces have been generated using "ramp" functions, both in the input and the output spaces. As pointed out in [8] ramp functions as target functions on the duration of a phoneme determine a linear increase in the estimation of the detection of a phoneme, as the pieces of the signal are sequentially input. In other words, the network is going to be generating a larger level of detection and of the internal state corresponding to a phoneme as it receives correct samples of it. This allows the network to prevent making mistakes when in a given instant it receives distorted input signals, which it will ignore if its internal state level is sufficiently high and compensates for the wrong input. With only 12 samples, belonging to 9 speakers,

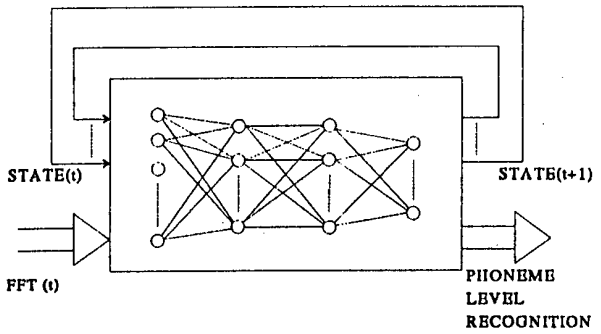


Figure 1: Equivalent neural network of the automaton for the detection of the vocalic phonemes.

6 male and 3 female, in the training set, segmenting by hand the pertinent areas of each phonemes in their intervals of largest energy, we have obtained a recognition level of 100 % in the training set and 97.4 % in the test set.

2. In this solution we have a training procedure that is to rigid in the sense that we are imposing a supervised training scheme on the evolution of both, the outputs and the state. We are also imposing a linear ramp function in this evolution. This selection is not critical in the case of the output space, as we could have chosen other evolution functions, which in the end would only specify the recognition levels during the evolution and would not actually affect it. But it does seem a very restrictive criteria on the evolution of the state space, which is fedback as input in the next step and consequently influences subsequent outputs and states. It would probably be more convenient to try to obtain some type of less supervised training for the state variables, allowing the networks to determine their correct value during evolution.

We have employed a strategy that, even though it is supervised, it is less restrictive regarding the state values as well as the recognition levels for each phoneme:

i) if the largest of the outputs that encodes the state or the recognition levels corresponds to the correct phoneme, then we will not train or we will train less over these outputs. The same for the recognition level nodes.

ii) Otherwise, we will train all of the outputs normally.

Over the same samples as in the previous case, the recognition levels are very similar, 100 % for the training set and 94.1 % for the test set, but this less restrictive supervision in all the output nodes of the network produces a very important consequence. We are now generating output maps that are close to those produced in non supervised training in the sense that we

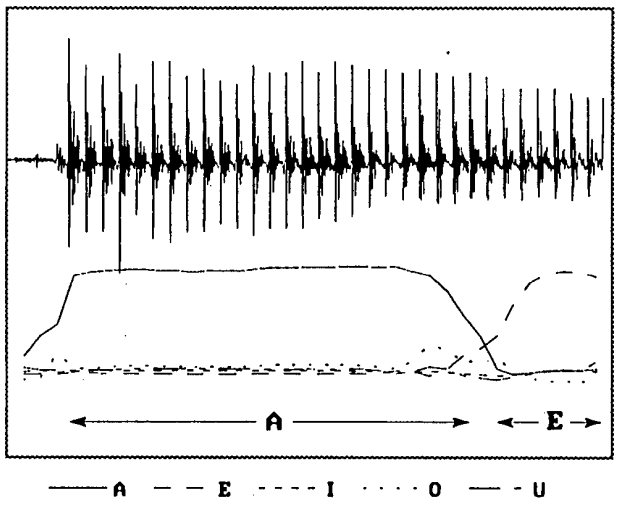


Figure 2: Level recognition in the transition between the phonemes /a/ to /e/ in continuos speech with the second strategy

obtain a clustering of the inputs, generating values for outputs that are similar in those nodes that represent phonemes whose features are similar, in this case whose frequency spectrum is similar. Figure 2 represents the gradual transition from the phoneme /a/ to /e/ with this strategy.

We must now extend this network model to the rest of the phonemes in the Spanish language. In order to do this we have chosen a structure whereby we create specialized recognition submodules for the recognition of different groups of phonemes. Thus we propose the following modules: vocalic, voiced and unvoiced occlusive, fricative, lateral and nasal. The modularization of the recognition presents the clear advantage of facilitating training in each submodule of the network, and at the same time we are injecting knowledge through this modularization in the global structure of the recognizer.

In the case of consonantic phonemes, we find new problems. For instance in unvoiced phonemes such as /f/, /s/ or /z/, the non periodicity of their waveform implies that it is very difficult to obtain a window width that is appropriate for the computation of the FFT, which is not going to be as constant as in other phonemes, specially due to the presence of noise in these signals that have a much lower energy level than vowels. For this reason we have applied a low pass filter over the temporal evolution of the spectra, obtaining results, using the second strategy, of 98.1 % recognition over the training samples and 70 % over the test samples with the same group of speakers.

4. FUTURE WORK

We have applied this methodology to phoneme recognition training the RANNs with samples that were segmented by hand. This permits a sufficiently clear detection of the phonemes when testing using continuous speech samples, as we do not detect the punctual presence of a phoneme through the analysis of a frame, but its evolution throughout the duration of the phoneme. From this point on we must try to include in the training samples the coarticulation effects of each phoneme with its possible neighbors in order to improve the quality of the recognition. On the other hand, the network must detect phonemes independently not only of their temporal lengths, but also of the length of the features whose evolution determines the presence of a phoneme. This is to say, that we must make the automaton more flexible with respect to temporal expansions or contractions in the signal features. That is the case of the voiced stop consonants (/b/, /d/ and /g/), whose characteristics of constriction in the vocal tract, low frequency energy and final explosion, with lack of high frequency energy, or the unvoiced stop phonemes (/p/, /t/ and /k/) characteristics of constriction, frication, aspiration and voiced explosion vary in temporal length, but this variation in an explosive interval that is extremely short with respect to the duration of other phonemes is of the utmost importance for its recognition.

5. CONCLUSIONS

These preliminary results are very promising as they show a path for a simple implementation of speech recognition systems using the inherent sequentiality of speech signals as a resource in order to improve their recognition. In addition, the modularization of the systems permit reducing the ever present complexity of training in these systems, allowing for a better choice of training sets, some control over the sharpness of the detection we desire and a reduction of the problems introduced by windowing. Due to the usual lack of quality of speech signals and the noise levels in normal speaking environments, recognition of individual phonemes in real speech will never be perfect. This implies that it is necessary to generate mechanisms for the interpolation of mistakenly recognized phonemes, which obviously can only be carried out through the introduction of context knowledge in the recognition systems. Their modularization will also help in this process.

6. REFERENCES

[1] Kohonen, T., "The Neural Phonetic Typewriter", *Computer*, Vol 2, N 3, 11-22, 1988.

[2] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang. K., "Phoneme Recognition Using Time Delay Neural Networks". *IEEE Transactions on Acoustics, Speech and Signal Procesing*, Vol ASSP-37, 328-339, March 1989.

[3] Duro, R.J., Santos, J., and Sarmiento, A., "GENIAL: An Evolutionary Recurrent Neural Network Designer and Trainer". *Proceedings of CAST'94 (Fourth International Workshop on Computer Aided Systems Technology)*, Ottawa, Canada, May 1994.

[4] Duro, R.J., and Santos, J., "Sequential Pattern Detecting Recurrent Artificial NNs Designed Using Evolutionary Techniques", *Proceedings of the COST#229 Workshop on Adaptive Systems and Intelligent Approaches, Massively Parallel Computing, and Emergent Techniques in Signal Processing and Communications*, 229-232, Bayona-Vigo, Spain, October 1994.

[5] Santos, J., and Duro, R.J., "Evolutionary Aided Design of Pattern Detecting Recurrent Artificial Neural Networks". *Procc. of EUROCAST'95 (Fifth International Conference on Computer Aided Systems Technology)*, Innsbruck, Austria, May 1995.

[6] Prager, R.W., Harrison, T.D., and Fallside, F., "Boltzmann Machines for Speech Recognition", *Computer Speech and Language* 1, 2-27, 1986.

[7] Jordan, M.I., "Attractor Dynamics and Parallelism in a Connectionist Sequential Machine", *Proceedings of the 1986 Cognitive Science Conference*, L. Erlbaum, Hillsdale, N.J., 531-546, 1986.

[8] Etemad, K., "Phoneme Recognition Based on Multi-resolution and Non-causal Context", *Procc. 1993 IEEE Workshop on Neural Networks for Signal Processing*, 343-352, 1993.

ANALYSIS AND DETECTION OF SPANISH-ACCENTED ENGLISH

R. O. Aréchiga + , R. Jordán + , and Nelson Morgan ++

+ Department of Electrical and Computer Engineering
University of New Mexico

++ International Computer Science Institute
1947 Center St., Suite 600
Berkeley, CA 94704

ABSTRACT

The undesirable variability caused by foreign accent in speaker independent speech recognition systems[6] is claimed to be due to *interference* of the speaker's native language[3, 9]. The pronunciation of particular vowels are the most common and simple type of variation between accents. The Formant structure is the basis for the recognition of most vowel differences[5], and vowel color is mainly determined by the frequencies of the first two formant frequencies. This work explores the nature of accent in general, highlights the differences between Spanish and English, and analyzes the first two formants of the vowels of Spanish-accented English. These are used as features for detection of Spanish accent. The conclusion is that Spanish accent is detectable using formants of vowels as features.

Introduction

Accent, if defined as a manner of pronunciation, is ubiquitous since variability is inherent to pronunciation within and between speakers. In spite of this variability the perceiver discriminates the patterns of the speech sounds, partly by the distinctiveness of the patterns in the artifacts and partly by prediction. Therefore, for communication to occur, the contrasts on which the patterns are based should not be obscured. Some of the linguistic regularities of the patterns could be exploited by a Speech Recognition System (ASR)[2]. Identification of an accent will allow an ASR to improve its performance, by making use of the explicit knowledge about such accent.

Besides carrying the patterns which convey language, the speech sounds accommodate nonlinguistic signs. These signs may be called Indexical Features, some of which are distinctive ways of pronouncing certain vowels or consonants, or of word and sentence stress and intonation patterns.

Almost all speakers of all languages have Regional Indices in their pronunciation. The word *accent* in its popular sense,is usually used to refer to these regional indices. These indices can alternatively be used to assign a different meaning to the words with which they are used[9].

Foreign accent, according to Chreist[3], is the inability of the individual habituated in patterns of his native language to hear the fine differences present in the sound pattern of a second language. In agreement with him, Lado[8] says that the speaker of one language listening to another does not actually hear the foreign language sound units-phonemes. He hears his own. Phonemic differences in the foreign language will be consistently missed by him if there is no similar phonemic differences in his native language.

Hispanics whose native language is Spanish may have problems with some English vowels, consonants, consonant clusters, word and sentence stress and intonation. According to the *interference* theory it can be predicted that accent differences will be either systemic, i.e., caused by a different number of phonemes between the two languages, or realizational, which are related to how the speakers actually pronounce the phonemes. Realizational differences account for most of the distinction between one accent and another, and hence provide most of the identifying characteristics of each accent. Usually, these differences pervade the whole vowel system, and make up the distinguishing characteristics of a particular accent[9].

Duration, another important feature for distinguishing vowels like [UW](two) and [OW](four), has to wait for a more representative sample. The trend so far, as predicted, is that those vowels are longer for native speakers, most probably due to the diphthongization of single vowels, not present in Spanish.

Some American authors claim that there is no variation in length in Spanish vowels; and we could say that this is so compared with the range of variation in length of English vowels, where, for stressed positions, the shorter vowels (IH,UH,EY,and AH) are more central. It would be closer to the truth to say that in Spanish, distinctive differences of duration are not stable[4]. Therefore we can say that in Spanish, vowel duration does not have a contrastive function.

Comparison between Spanish and English

Spanish is one of the languages in which the spelling indicates the sounds of the letters. Consonants and vowels are distributed about equally and are pronounced with relatively the same duration. The five vowels are pure in sound quality, and two or more frequently occur in consecutive arrangement... There are fourteen diphthongs. A few consonant clusters are in the system. Aspiration is not strongly applied to either consonants or vowels. There are nineteen consonant phonemes. [ch] and [f] are lowest in frequency count. Of the vowels, [i] and [u] have the lowest frequency count[10].

English, compared with Spanish, is a relatively unphonetic language. It is not always possible to determine the sound from the spelling. The consonant distribution exceeds that of the vowels and there is a variation in the duration applied as each is pronounced. There are eleven vowels (vowel sounds). Authorities differ in the number of diphthongs. The vowels are not pure in sound quality and there is a tendency to lengthen them into diphthongs, particularly in some regions. Two vowels may occur in consecutive arrangement. The frequency of weak, or unstressed, vowels is typical. There are 25 consonant phonemes, and consonant clusters are typical with many occurring in final position[11].

Spanish does not have the following vowel sounds. They are followed, when predictable, with the vowels Hispanics might substitute them with[7].

IH (bit, hid,..) with [IY](beat, heed).

EY (bait,hayed) ?

AE (bat,had) with [EH](bet, head) or [AA](father, hod).

AX (but,ago) with vowel suggested by spelling.

UH (book,hood) with [UW](boot).

OW (boat) with [AO](bought).

AH (mud) ?

Experiments

The experiments focus on the Formant Structure of Spanish-accented vowels. Some of the vowels pronunciation problems for Hispanics, like UH](bulls), [AH](but), [ER](girls) have been overcome or blurred by the speakers analyzed so far and avoid easy discrimination. The closest vowel is [ER](girls) even though it doesn't exist in Spanish. The most separable vowel turned out to be [AE](ran) which Hispanics will substitute for a vowel close to the Spanish vowel [a](padre, father in Spanish), which is the most frequent vowel in Spanish. The Spanish vowel [a](padre)sits between the English vowels [AA](father) and [AE](ran).

All the vowels were analyzed. Segmentation was done by a combination of a dynamic programming procedure that looks at spectral coefficients, listening to the sounds and looking at the evolution of formants. The data, originally at 16 KHz was bandpass filtered and resampled at 10 KHz, and together with sets of excitation markers, the formants(F1 to F4) were extracted. For each speaker, the means of the first two formants were calculated. Then the overall mean (for all speakers) for each formant was calculated. These last values for the first two formants are the ones reported in tabular and graphical form together with the *typical* values for American English.

Average duration for the words two [UW] and four [OW] are shown below. As predicted, Americans' average is longer.

From these data, the easiest vowel to *detect accent* is [AE] by simply using a Linear Discriminant function, or by measuring the Euclidean distance with two or three formants.

A linear discriminant shows (Figure 2) that in the sample, with the vowel [AE](ran) there is only one Hispanic out of 11 Hispanics classified as American. With the vowel [IH], three are misclassified, followed by [AA](father), etc. The addition of the third formant didn't modify the classifications.

Database

The experiments were done with eleven adult males (natives of Cuba and Peru) from a Database of Spanish-accented English speakers from the Miami area [1], and six adult males from a growing home-based Database[2] of native American speakers. The American English database has been used for contrasting and classification purposes.

[1] provided by Dr. Marc Zissman from MIT

[2] partly with the help of Kevin Campbell, Music Dept. UNM

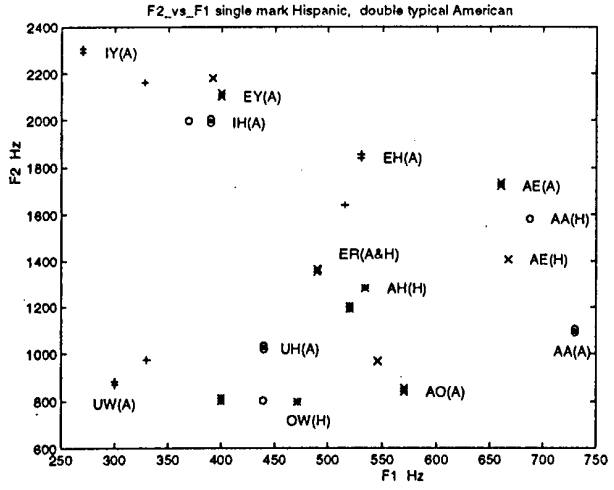


Figure 1: F2 vs F1 for Hispanics and Typical American English

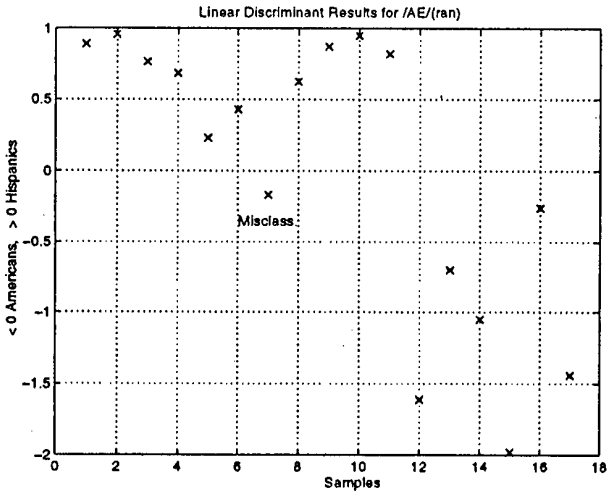


Figure 2: Linear Discriminat for vowel /AE/(ran)

Formant vowel frequencies in Hz. for Spanish-accented English and Std. English				
Vowel	F1(Hisp)	F1(Am)	F2(Hisp)	F2(Am)
[*AA*]	663	730	1584	1090
[*AE*]	670	660	1422	1720
[*AH*]	534	520	1282	1190
[*AO*]	556	570	1121	840
[*IH*]	369	390	1998	1990
[*EH*]	515	530	1641	1840
[*EY*]	392	400	2056	2100
[*IY*]	328	270	2160	2290
[*OW*]	471	450	799	900
[*UH*]	439	442	827	1020
[*UW*]	331	300	983	870
[*ER*]	491	490	1361	1350

Formant vowel frequencies in Hz. for Spanish-accented English and Spanish.				
Vowel	F1(Hisp.)	F2(Hisp.)	F1(Sp.)	F2(Sp.)
[*AA*]	687	1580	725	1300
[*EY*]	392	2056	450	1900
[*IY*]	328	2160	275	2300
[*OW*]	471	799	450	900
[*UW*]	330	974	275	800

Average Duration in seconds		
Word	Hisp.	Am.
two[*UW*]	0.26	0.43
four[*OW*]	0.32	0.39

Conclusions

Spanish-accented English is detectable using formants of vowels as features. The vowel [AE](ran) is the most separable, then IH.

Therefore, to *detect Spanish-accented English from vowels* with highest probability, this study shows that the vowel [AE](ran) is the best choice.

Future Work

From the current experience with the Miami database, the similarities far outweigh the differences. Once the detection is done, the recognition could be done in several ways: A spectral transformation that establishes a correspondence between pairs of 'typical' spectra from two talkers based on their occurrence in the same context speech (for example, vowels embedded in carrier words); or Hidden Markov Models could be trained separately for Hispanics or Multiple Pronunciation Models could be used supported by rules based on the knowledge of the characteristics of the Hispanics.

1. REFERENCES

[1] David Abercrombie, "Elements of General Phonetics", [Aldine Atherton], 1967.

[2] W.J. Barry, C.E. Hoequist and F. J. Nolan, "An approach to the problem of regional accent in automatic speech recognition", *Computer Speech and Language*, 1993.

[3] Fred M. Chreist, "Foreign accent ", [Prentice-Hall Inc.], 1964.

[4] Pierre Delattre, Comparing the phonetic features of English, French, German and Spanish", [George G. Harrap], 1964.

[5] D.B. Fry, "The Physics of Speech", Cambridge University Press , 1979.

[6] J.H.L. Hansen and L. Arslan, "Foreign accent classification using source based prosodic features", *IEEE Proc. Int. Conf. on Ac., Sp., and Sig. Proc.*, vol. 1 (1995).

[7] Joanne Kenworthy, " Teaching English Pronunciation", [Longman], 1987

[8] Robert Lado, "Linguistics Across Cultures", [Ann Arbor - The University of Michigan Press], 1957.

[9] John Laver, "Principles of Phonetics", [Cambridge University Press], 1994.

[10] Tomas Navarro Tomas, "Manual de Pronunciacion Espanola", [Consejo Superior de Investigaciones Cientificas], 1990.

[11] Margaret F. Tucker, " Listening Lessons in connected speech for Puerto Rican College Students for the purpose of improving aural comprehension in English ", [Unpublished M.S. thesis, UNM], 1963.

APPLICATIONS OF CHAOS IN COMMUNICATIONS

Michael Peter Kennedy

Department of Electronic and Electrical Engineering
University College, Dublin 4, IRELAND
Tel: +353-1-706 1963; fax: +353-1-283 0921
e-mail: Peter.Kennedy@ucd.ie

ABSTRACT

This tutorial reviews the state of the art in applications of chaos in broadband communications.

1. INTRODUCTION

The goal of a digital communication system is to convey information from a digital information source to a receiver through a channel as effectively as possible [1]. This is accomplished by mapping the digital information to a sequence of symbols which vary some parameter of an analog electromagnetic wave called the carrier (*modulation*). At the receiver, the signal is demodulated, interpreted, and the information recovered.

The mapping from *baseband* digital information to a *passband* carrier signal may be accompanied by encryption and coding to add end-to-end 'security', data compression, and error-correction capability.

Built-in error-correction capability is required because real channels distort analog signals by a variety of linear and nonlinear mechanisms: attenuation, dispersion, fading, noise, interference, multipath effects, etc.. A *channel encoder* introduces algorithmic redundancy into the transmitted symbol sequence that reduces the probability of incorrect decisions at the receiver.

Modulation is the process by which a symbol is transformed into an analog waveform that is suitable for transmission. Common digital modulation schemes include Phase-Shift-Keying (PSK) and Frequency-Shift-Keying (FSK), where a one-to-one correspondence is established between phases and frequencies, respectively, of a sinusoidal carrier and the symbols.

The *channel* is the physical medium that carries the signal from the transmitter to receiver. Inevitably, the signal becomes corrupted in the channel. Hence, the receiver seldom receives exactly what was transmitted. The role of the *demodulator* is to produce from the corrupted received signal an estimate of the transmitted symbol sequence. The *channel decoder* exploits redundancy in the transmitted sequence to reconstruct the original information. Because of disturbances in real channels, error-free transmission is not possible.

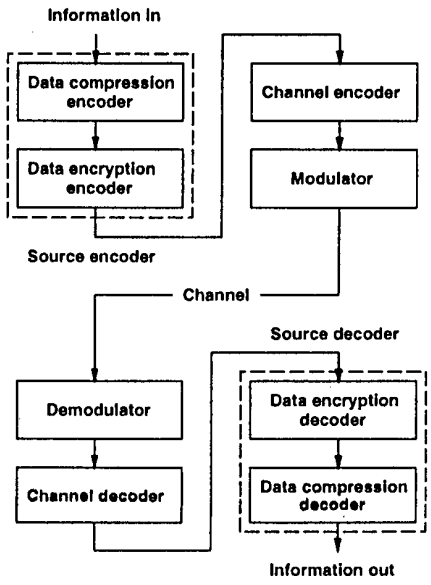


Figure 1: Digital communication system showing source and channel coding, modulation, and channel.

The performance of the communication system is measured in terms of the bit error rate (BER) at the receiver. In general, this depends on the coding scheme, the type of waveform used, transmitter power, channel characteristics, and demodulation scheme. The conventional graphical representation of performance in a linear channel with Additive White Gaussian Noise (AWGN) shows bit error rate versus E_b/N_0, where E_b is the energy per bit and N_0 is the power spectral density of the noise introduced in the channel.

For a given background noise level, the BER may be reduced by increasing the energy associated with

The author acknowledges invaluable discussions with G. Kolumbán, TU Budapest.

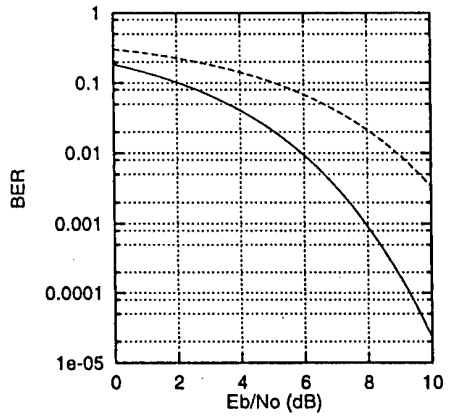


Figure 2: Comparison of the noise performances of two digital modulation schemes: differential PSK (solid) and noncoherent FSK (dashed).

each bit, either by transmitting with higher power or for a longer period per bit. The challenge in digital communications is to achieve a specified BER with minimum energy per bit. A further consideration is *bandwidth efficiency* [1].

Nonlinear dynamics has potential applications in several of the building blocks of a digital communication system: data compression, encryption, and modulation. In this paper, we focus primarily on the application of chaos as a spread spectrum modulation scheme.

2. SPREAD SPECTRUM MODULATION

In spread spectrum modulation, the transmitted signal is spread over a much larger bandwidth than is necessary to transmit the baseband information. Spread spectrum can be used for:

- combatting the effects of interference due to jamming, other users, and multipath effects,
- hiding a signal "in the noise" by transmitting it at low power, and
- achieving message privacy in the presence of eavesdroppers.

Conventional spread spectrum communication systems use *pseudorandom* or *PN* spreading sequences to distribute the energy of the information signal over a wide bandwidth. The transmitted signal appears similar to random noise and is therefore difficult to detect by eavesdroppers. With a synchronized receiver, interferences can be suppressed by despreading. In addition, by using orthogonal pseudorandom spreading sequences, multiple users may communicate simultaneously on the same channel (CDMA).

Spread spectrum techniques are suited for applications in satellite communications (low power spectral density), mobile phones (privacy, high tolerance against multipath effects, multiple users), and military communications (low probability of intercept).

2.1. PSEUDORANDOM VS. CHAOTIC

Pseudorandom (PN) spreading sequences are widely used in spread spectrum communications because their statistics and orthogonality properties are well understood, they are easy to generate, and easy to synchronize. However, the inherent periodicity of a pseudorandom sequence compromises the overall *security* of a spread spectrum communication system. The greater the length of the pseudorandom sequence, the higher is the level of security, but the more difficult it is to establish synchronization at the demodulator.

Figure 3 shows a block diagram of a spread spectrum system using a PN spreader. The modulator spreads the data stream from the channel encoder, as determined by the spreading sequence, and transmits on a sinusoidal carrier using PSK or FSK.

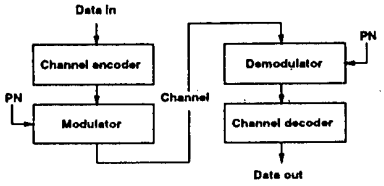


Figure 3: Spread spectrum communication system using a conventional PN spreader.

A pseudorandom sequence generator is a special case of a chaotic system, the difference being that the chaotic system has an infinite number of (analog) states, while the pseudorandom generator has a finite number. A pseudorandom sequence is produced by visiting each state of the system once in a deterministic manner. With only a finite number of states to visit, the output sequence is necessarily periodic. By contrast, a chaotic generator can visit an infinite number of states in a deterministic manner and therefore produce an output sequence which *never* repeats itself.

What are the advantages of using chaos in spread spectrum communication systems? With appropriate modulation and demodulation techniques, the "noise-like" spectral properties of chaotic electronic circuits [2] can be used to provide *simultaneous spreading and modulation* of a transmission. The simplicity of the analog circuits involved could permit extremely high speed, low power implementations.

3. CHAOTIC MODULATION: STATE OF THE ART

Exploratory studies of communicating with chaos have been carried out over the past five years [3]. To date, several techniques have been developed for generating chaotic signals, chaotic modulators and demodulators, and self-synchronizing chaotic receivers.

In the remainder of this paper, we summarize the current state of knowledge in each of these domains and highlight the areas in which improvements are required in order to realize the goal of a practical spread spectrum communication system exploiting chaos.

3.1. GENERATION OF CHAOTIC SPREADING SIGNALS

Two possibilities exist here: the baseband information signal may be spread at an intermediate frequency and up-converted using a conventional mixer and power amplifier, or the spreading may be accomplished directly at the transmission frequency.

Widely-studied circuits such as Chua's oscillator [2] may be used as lowpass chaotic signal generators and their outputs mixed up to the RF transmission band. The principal disadvantage of this approach is that linear wideband circuitry is required both in the mixer and power amplifier stages.

The chaotic analog phase-locked loop (APLL) introduced by Kolumbán [4] and shown in Fig. 4 offers a cost-effective means of directly generating a wideband RF spread spectrum signal at high power.

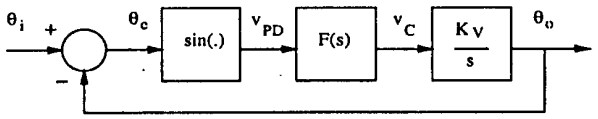


Figure 4: Nonlinear baseband model of the chaotic APLL [4]. $F(s)$ is a second-order LPF.

3.2. CHAOTIC MODULATION AND SPREADING SCHEMES

Over the past three years, five chaos-based modulation and spreading techniques have been developed: chaotic masking, inverse systems, Predictive Poincaré Control (PPC) modulation, Chaos Shift Keying (CSK), and, most recently, differential CSK (DCSK).

Each of the techniques described below has been demonstrated experimentally using discrete components [5] or dedicated circuitry. More recently, prototype spreaders have been realized as integrated circuits [6].

3.2.1. CHAOTIC MASKING

In chaotic masking [7], the information signal $s(t)$ is spread by adding it to the output $y(t)$ of a chaotic system. The resulting signal $s(t) + y(t)$ is modulated and transmitted [3]. Provided that the information signal $s(t)$ is small compared to $x(t)$, an identical chaotic system in the receiver can be made to synchronize with $x(t)$. This permits the receiver to "filter" out the "disturbance" $s(t)$. Thus, $s(t)$ can be retrieved by simply subtracting the output of the receiver's chaotic system from the received signal.

Chaotic masking suffers from the disadvantage that distortion and noise introduced by the channel are indistinguishable from the signal.

3.2.2. INVERSE SYSTEMS

In the inverse system approach [8], the transmitter consists of a chaotic system which is excited by the information signal $s(t)$ and produces a chaotic output $y(t)$. The receiver is an *inverse system*, i.e. one which produces $\tilde{s}(t) = s(t)$ as output when excited by $y(t)$ and started from the same initial condition. If the system is properly designed, $\tilde{s}(t)$ converges to $s(t)$, regardless of the initial conditions. Inverse systems are widely used in digital encryption and spreading schemes.

3.2.3. PREDICTIVE POINCARÉ CONTROL (PPC) MODULATION

In Predictive Poincaré Control (PPC) modulation, symbolic analysis of chaotic systems is used to encode and decode the information [9]. With an appropriate control strategy, the transmitter is forced to follow a prescribed path, in a symbolic sense. On the receiver side, an identical chaotic system will synchronize approximately with the transmitter system. By identifying the symbolic path in the synchronized receiver, the information signal can be retrieved.

Although each of these techniques has been demonstrated in the laboratory, none has been tested experimentally using a noisy communication channel. Indeed, recent simulations [10] suggest that masking and inverse systems perform poorly if the transmitted signal is corrupted by noise. A more robust method is Chaos Shift Keying.

3.2.4. CHAOS SHIFT KEYING (CSK)

In binary Chaos Shift Keying [5]. an information signal is encoded by transmitting one chaotic signal for a "1" and another chaotic signal for a "0". The two chaotic signals come from two different systems (or the same system with different parameters).

Two demodulation schemes are possible: coherent and noncoherent. The coherent receiver contains copies of the systems corresponding to "0" and "1". Depending on the transmitted signal, one of these will synchronize with the incoming signal and the other will desynchronize. By detecting synchronization at the receiver, one may determine which bit is being transmitted.

In the case of non-coherent demodulation, no attempt is made to recover the carrier at the receiver; instead, one simply examines statistical attributes of the received signal.

3.2.5. DIFFERENTIAL CSK (DCSK)

Differential CSK [11] is a development of CSK which exhibits lower sensitivity to channel imperfections than any of the techniques outlined above.

In DCSK, the modulator is a free-running chaotic generator with output $x(t)$. Each binary symbol is encoded for transmission as two bits, the first of which acts as a reference, the second carrying the information. To transmit a "1", $x(t)$ and a one-bit-delayed copy of x are applied to the channel during successive bit periods. A "0" is indicated by transmitting $x(t)$ for the first bit period, and an inverted one-bit-delayed copy of this signal during the next bit period.

The chaotic signal sent via the channel is correlated with the signal received during the previous bit interval and a decision is made based on the output of the correlator. Since both signals presented to the correlator have passed through the same channel, DCSK exhibits robustness in the presence of channel imperfections.

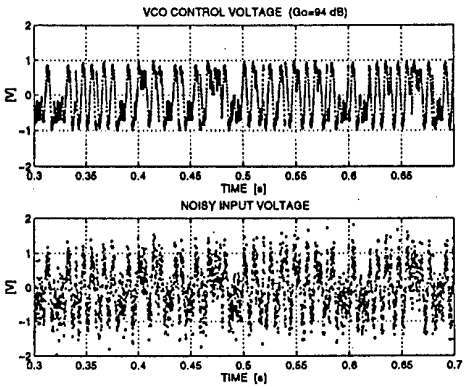
4. DIGITAL DEMODULATION

Digital demodulation refers to the process in the receiver by which the transmitted digital information signal is recovered from the incoming modulated wave. Either noncoherent or coherent techniques may be employed in the demodulator; these typically involve correlation detection or statistical tests in each bit interval, followed by threshold-based decision-making.

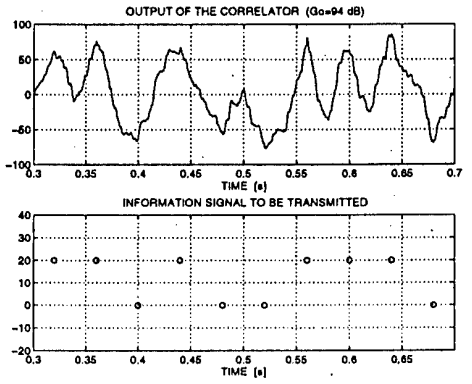
4.1. NONCOHERENT DEMODULATION

Noncoherent demodulation techniques have been proposed for CSK [4, 12] using a transmitter comprised of two chaotic APLLs corresponding to symbols "1" and "0". Statistics of the received signal (mean and standard deviation) are used in decision-making.

In the case of DCSK using APLLs, a noncoherent demodulation scheme has been proposed in which the incoming signal is correlated with a delayed version of itself. The output of the correlator is positive (negative) when a "1" ("0") is transmitted and tends to zero between adjacent bits [11]. Figure 5 shows simulations of this system with additive channel noise. Decisions are made at the peaks of the correlator output.



(a)



(b)

Figure 5: DCSK using APLL chaos. (a) transmitted and noisy received signals; (b) output of correlator and received bit sequence (from [11]).

4.2. COHERENT DEMODULATION BY CHAOS SYNCHRONIZATION

CSK was originally described in terms of *synchronization* of chaotic subsystems in a receiver matched to those in the transmitter [5].

Indeed, it was the observation by Pecora and Carroll at the Naval Research Laboratory in 1990 [13] that two chaotic systems could be synchronized without using an external synchronizing signal which raised the possibility of self-synchronizing coherent demodulators for chaotic transmissions.

While this discovery has generated a great deal of interest in exploiting the properties of chaos for spread spectrum communications, recent studies of currently-known synchronization schemes [10] suggest that they are not sufficiently robust for use in noisy channels. Better synchronization methods must be developed if we are to realize our dream of an efficient chaotic spread spectrum communication system.

5. ADDITIONAL CONSIDERATIONS

In this work, we have adopted the position that security is an add-on feature in a digital communication system which may be implemented by adding encryption/decryption hardware at each end of the system, as shown in Fig. 1. Nevertheless, there exist strong similarities between the concept of discrete-time inverse systems and self-synchronizing stream ciphers [14] which may permit a hybrid approach to chaotic modulation and encryption.

One advantage of using pseudorandom spreading sequences in a spread spectrum system is that multiple users are permitted simultaneous access to the channel provided they use uncorrelated pseudorandom sequences or *codes*. This is called code division multiple access (CDMA). Further work is required to define an equivalent concept of orthogonality for chaotic spreading signals.

6. ENGINEERING CHALLENGES

The field of "Communicating with chaos" presents many challenging problems at the basic, strategic, and applied levels.

The basic system level building blocks from which to construct a practical chaos-based spread spectrum communication system already exist: APLL chaos, CSK and DCSK, noncoherent and coherent demodulators. Nevertheless, further research and development is required in all of these subsystems.

We must characterize completely the dynamics of the APLL and develop design rules for constructing robust, reproducible chaotic transmitters. We must determine the performance of CSK and DCSK modulation schemes for noisy channels. We must analyze the statistical properties of CSK and DCSK transmissions in order to implement simple and robust receivers.

Current proposals for CSK and DCSK receivers using APLL chaos do not exploit the fact that the modulated signal has been produced by a chaotic system. If we can exploit the underlying structure of the transmitted signal, by chaos synchronization or otherwise, improved receiver performance will result.

7. REFERENCES

[1] S.S. Haykin. *Communication Systems.* John Wiley & Sons, Inc., New York, 1994. 3rd edition.

[2] M.P. Kennedy. Basic concepts of nonlinear dynamics and chaos. In C. Toumazou, editor, *Circuits & Systems Tutorials*, chapter 6.1, pages 289–313. IEEE ISCAS'94, London, UK, May 1994.

[3] M. Hasler. Engineering chaos for secure communication systems. *Phil. Trans. R. Soc. Lond. A*, 353(1701), 16 October 1995.

[4] G. Kolumbán and B. Vizvári. Direct symbol generation by PLL for the chaos shift keying modulation. In *Proc. ECCTD'95*, volume 1, pages 483–486, Istanbul, 27–31 August 1995.

[5] H. Dedieu, M.P. Kennedy, and M. Hasler. Chaos shift keying: Modulation and demodulation of a chaotic carrier using self-synchronizing Chua's circuits. *IEEE Trans. Circuits Syst.*, CAS-40(10):634–642, October 1993.

[6] M. Delgado-Restituto, R. López-Ahumada, and A. Rodríguez-Vázquez. Secure communication using CMOS current-mode sampled-data circuits. In *Proc. NDES'95*, pages 237–240, Dublin, 28–29 July 1995.

[7] K.M. Cuomo and A.V. Oppenheim. Circuit implementation of synchronized chaos with applications to communications. *Phys. Rev. Lett.*, 71(1):65–68, 1993.

[8] U. Feldmann, M. Hasler, and W. Schwarz. Communication by chaotic signals: the inverse systems approach. *Int. J. Circuit Theory Appl.*, 24, September 1996. (in press).

[9] J. Schweizer and M.P. Kennedy. Predictive Poincaré control. *Phys. Rev. E*, 52(5):4865–4876, November 1995.

[10] G. Kolumbán, H. Dedieu, J. Schweizer, J. Ennitis, and B. Vizvári. Performance evaluation and comparison of chaos communication systems. In *Proc. NDES'96*, Sevilla, 27–28 June 1996.

[11] G. Kolumbán, B. Vizvári, W. Schwarz, and A. Abel. Differential chaos shift keying: A robust coding for chaotic communication. In *Proc. NDES'96*, Sevilla, 27–28 June 1996.

[12] N. Smyth, C. Crowley, and M.P. Kennedy. Improved receiver for csk spread spectrum communications using analog phase locked loop chaos. In *Proc. NDES'96*, Sevilla, 27–28 June 1996.

[13] L.M. Pecora and T.L. Carroll. Synchronization in chaotic systems. *Phys. Rev. Letters*, 64(8):821—824, 1990.

[14] J. Daemen. *Cipher and Hash Function Design—Strategies based on linear and differential cryptanalysis.* PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium, March 1995.

ERROR BOUNDS IN CONSTRUCTIVE APPROXIMATION

D. Docampo, C.T. Abdallah and D. Hush

Departamento de Tecnologías de las Comunicaciones.
Universidad de Vigo.
36200 Vigo-SPAIN

ABSTRACT

In this paper we study the error bounds of constructive approximations of a given function with elements taken from a prescribed dictionary or subspace.

The paper contributes to clarify some recent convergence results concerning constructive solutions with application in several approximation problems.

1. INTRODUCTION

Continuous functions on compact subsets of $\mathcal{R}^d$ can be uniformly approximated by linear combinations of sigmoidal functions [1] [2]. The error in the approximation is related to the number of functions used (nodes in a neural network). A first convergence result was given by Maurey, as reported in [3]. Recently, some constructive solutions have been reported, where the iterations taking place involve computations in a reduced subset [4] [5].

The problem of constructive approximation can be stated as follows: approximate a given element (function) f in a Hilbert space H by means of an iterative sequence f_n, formed as a convex combination of elements taken from a subset of H, G. It has an enormous impact in establishing convergence results for projection pursuit algorithms [6], neural network training [5] and classification [7].

The paper is organized as follows: Section 2 will state the problem and analyze Barron's [5] and Dingankar's [7] solutions. Section 3 will discuss a framework under which those solutions can be formulated. Section 4 will analyze the limits and bounds of the errors. Finally, Section 5 will close the paper with the conclusions.

2. PRELIMINARIES AND CONSTRUCTIVE RESULTS

Throughout this paper, G will be a subset of a real or complex Hilbert space H, with norm $||.||$, The elements of G, g, are bounded in norm by some positive constant b. $\bar{co}(G)$ will denote the convex closure of G (i.e. the closure of the convex hull of G in H).

2.1. THE APPROXIMATION PROBLEM

The first global bound result, attributed to Maurey, concerning the error in approximating an element of $\bar{co}(G)$ using convex combinations of n points in G, is the following:

Lemma 2.1 *Let f be an element of $\bar{co}(G)$ and c a constant such that $c > b^2 - ||f||^2 = b_f^2$. Then, for each positive integer n there is a point f_n in the convex hull of some n points of G such that:*

$$||f - f_n||^2 \leq \frac{c}{n}$$

The first constructive proof of this lemma was given by Jones [4] and refined by Barron [5]; the proof includes an algorithm to iterate the solution. The result can be stated as follows:

Theorem 2.1 *Let δ be a constant such that $\delta > b_f^2$. Then, for each element f in $\bar{co}(G)$, we can construct an iterative sequence f_n, f_n chosen as a convex combination of the previous iterate f_{n-1} and a $g_n \in G$, $f_n = (1-\lambda)f_{n-1} + \lambda g_n$, such that:*

$$||f - f_n||^2 \leq \frac{\delta}{n}$$

The relation between this problem and the universal approximation property of sigmoidal networks was clearly established in references [4] and [5]; specifically, it has been proven that, under certain mild restrictions, continuous functions on compact subsets of $\mathcal{R}^d$ belong to the convex hull of the set of sigmoidal functions that one hidden layer neural networks can generate. Moreover, since the proofs are constructive, an algorithm to achieve the theoretical bounds is also provided.

This work was supported by ISTEC and CICYT

Other nonlinear approximation techniques have also benefited from the solution to this problem: approximation by hinging hyperplanes [8], projection pursuit regression [9] and radial basis functions [10]. In all these related approximation problems the solution can always be constrained to fall in the closure of the convex hull of a subset of functions (e.g. hinged hyperplanes, ridge functions or radial basis functions in the examples mentioned above).

2.2. CONSTRUCTIVE ALGORITHMS

For the sake of clarity and completeness, we include here the proof of the main Theorem, following [5]. The proof needs to make use of the following Lemma:

Lemma 2.2 *Given $f \in \bar{co}(G)$, for each element of $co(G)$, h, and $\lambda \in [0,1]$:*

$$\inf_{g \in G} ||f-(1-\lambda)h-\lambda g||^2 \leq (1-\lambda)^2||f-h||^2+\lambda^2 b_f^2 \quad (1)$$

The proof of the lemma will be carried out for $f \in co(G)$; it extends to elements in $\bar{co}(G)$ because of the continuity of all the terms involved in the inequalities [11].

Since $f \in co(G)$, there exists a convex combination of elements g^* from G, so that $f = \sum_{k=1}^{m} \alpha_k g_k^*$. Let g^* be a random vector taking values on H so that $P(g^* = g_k^*) = \alpha_k$.

Then $E(g^*) = f$, and

$$\begin{aligned} var(g^*) &= E(||g^* - f||^2) \\ &= E(||g^*||^2) - ||f||^2 \leq b_f^2 \end{aligned}$$

Now, for $\lambda \in [0,1]$ and $d \in H$,

$$\begin{aligned} & E(||\lambda(g^* - f) + (1-\lambda)d||^2) = \\ = \;& \lambda^2 E(||g^* - f||^2) + (1-\lambda)^2||d||^2 \\ \leq \;& \lambda^2 b_f^2 + (1-\lambda)^2||d||^2 \end{aligned}$$

Then,

$$\begin{aligned} & \inf_{g \in G} ||f - (1-\lambda)h - \lambda g||^2 \\ \leq \;& E\left(||(1-\lambda)h + \lambda g^* - f||\right)^2 \\ \leq \;& E\left(||(1-\lambda)(h-f) + \lambda(g^* - f)||\right)^2 \\ \leq \;& (1-\lambda)^2||f-h||^2 + \lambda^2 b_f^2 \end{aligned}$$

which concludes the proof of Lemma 2.2.

We can now prove the main result, using an inductive argument.

At step 1, find g_1 and ϵ_1 so that

$$||f - g_1||^2 \leq \inf_G ||f-g||^2 + \epsilon_1 \leq \delta$$

This is guaranteed by (1), for $\lambda = 1$.

Let now f_n be our iterative sequence of elements in $co(G)$, and assume that for $n \geq 2$,

$$||f - f_{n-1}||^2 \leq \delta/(n-1)$$

It is then possible to choose among different values of λ and ϵ_n so that:

$$(1-\lambda)^2||f_{n-1} - f||^2 + \lambda^2 b_f^2 \leq \frac{\delta}{n} - \epsilon_n \quad (2)$$

At step n, select g_n such that $||f-(1-\lambda)f_{n-1}-\lambda g_n||^2 \leq$

$$\inf_{g \in G} ||f - (1-\lambda)f_{n-1} - \lambda g||^2 + \epsilon_n \quad (3)$$

Hence, using (1), (3) and (2), we get: $||f - f_n||^2 \leq \frac{\delta}{n}$, and that completes the proof of Theorem 2.1.

2.3. DISCUSSION

The values of λ and ϵ_n in [5] and [7] are related to the parameter α, $\alpha = \delta/b_f^2 - 1$, in the following way:

$$[5] \;:\; \lambda = \frac{||f - f_{n-1}||^2}{b_f^2 + ||f - f_{n-1}||^2}; \qquad \epsilon_n = \frac{\alpha\delta}{n(n+\alpha)}$$

$$[7] \;:\; \lambda = \frac{1}{n}; \qquad \epsilon_n = \frac{\alpha b_f^2}{n^2}$$

It is easy to show that admissible values of λ which satisfy inequality (2) for positive values of ϵ_n fall in the following interval, centered at Barron's optimal value for λ:

$$\frac{||f - f_{n-1}||^2}{b_f^2 + ||f - f_{n-1}||^2} \pm$$

$$\frac{1}{b_f^2 + ||f - f_{n-1}||^2}\sqrt{||f - f_{n-1}||^4 - ||f - f_{n-1}||^2 + \frac{\delta}{n}}$$

To evaluate the possible choices for the bound ϵ_n we need to make use of the induction hypothesis; introducing it in inequality (2), values of λ should now satisfy

$$(1-\lambda)^2 \frac{\delta}{n-1} + \lambda^2 b_f^2 \leq \frac{\delta}{n} - \epsilon_n$$

Then, admissible values of λ for positive values of ϵ_n fall in the interval:

$$\frac{1+\alpha}{n+\alpha} \pm \frac{n-1}{n+\alpha}\sqrt{\frac{\alpha(1+\alpha)}{n(n-1)}}$$

The fact that $1/n$ falls within the limits of this interval (as can easily be checked) explains why the average of n elements [7] is always a solution to the problem.

3. OPTIMAL PARAMETERS

We now formulate the following questions:

1. What is the minimum bound for the global error using convex combinations of n elements from G?
2. What is the optimal choice of λ for a given bound, so that the tolerance allowed for ϵ_n is maximum?

Based on the assumptions made and in Lemma 2.2, let us formulate the problem again in a more general way: Our objective is to look for a constructive approximation so that the overall error using n elements from G satisfies the following inequality:

$$\|f - f_n\|^2 \leq \frac{\delta}{b(n)} \tag{4}$$

$b(n)$ being a function of the parameter n which indicates the order of the approximation (i.e. $b(n) = n$ both in [4] and [7]) and δ the parameter related to b_f^2 as defined before.

In what follows we will assume that the iterate f_n will be chosen as a convex combination of the previous iterate f_{n-1} and a point in G, g_n; this introduces a loss of generality, since other constructive approaches could be devised in order to re-optimize the coefficients of previous elements from G at each step. The facts that f_n is forced to be a convex combination of n elements from G, and our algorithm has to be constructive, mean that f_n is in the convex hull of $\{g_1, g_2, \ldots, g_n\}$ and f_{n-1} is in the convex hull of $\{g_1, g_2, \ldots, g_{n-1}\}$, but that does not imply that f_n must be a convex combination of f_{n-1} and g_n, as can be easily shown. We leave the more general problem for further investigation and concentrate here on the case where constructiveness of the algorithm is taken as in [4] and [7] to be equivalent to the constraint that, at each step, f_n is in the convex hull of $\{f_{n-1}, g_n\}$.

To answer the questions posed at the beginning of this section, let us now set up a framework where constructive results can be formulated.

Let $f_n = (1-\lambda)f_{n-1} + \lambda g_n$; we want to find λ, ϵ_n, and the function $b(n)$ so that:

$$\begin{aligned}
\|f - f_n\|^2 &\leq \inf_{0<\lambda<1} \inf_{g\in G} \|f - (1-\lambda)f_{n-1} + \lambda g\|^2 + \epsilon_n \\
&\leq \inf_{0<\lambda<1} (1-\lambda)^2 \|f - f_{n-1}\|^2 + \lambda^2 b_f^2 + \epsilon_n \\
&\leq \inf_{0<\lambda<1} (1-\lambda)^2 \frac{\delta}{b(n-1)} + \lambda^2 b_f^2 + \epsilon_n \quad (5) \\
&\leq \frac{\delta}{b(n)}
\end{aligned}$$

Since $\delta = (1+\alpha)b_f^2$, we can rewrite the last inequality in the following way:

$$\inf_{0<\lambda<1} (1-\lambda)^2 \frac{\delta}{b(n-1)} + \lambda^2\delta + \epsilon_n - \lambda^2\alpha b_f^2 \leq \frac{\delta}{b(n)}$$

This last expression represents the trade-off between the global error we are trying to achieve $\delta/b(n)$ and the error at each of the subproblems, ϵ_n.

We are going to prove the following: if we set $\epsilon_n = \lambda^2\alpha b_f^2$, then, for a given λ, the best rate of convergence of the approximation which can be achieved, measured in $b(n)$, is the one given in [7] and [5], and the optimal value of λ which minimizes ϵ_n for that best rate of convergence is precisely the value given in [7]. To see that, let's introduce the value of ϵ_n in (5); then:

$$(1-\lambda)^2 \frac{\delta}{b(n-1)} + \lambda^2\delta \leq \frac{\delta}{b(n)}$$

Hence,

$$(1-\lambda)^2 + b(n-1)\lambda^2 \leq \frac{b(n-1)}{b(n)}$$

and then:

$$P(\lambda) = \lambda^2(1 + b(n-1)) - 2\lambda + 1 - \frac{b(n-1)}{b(n)} \leq 0 \tag{6}$$

$P(\lambda)$ has to have a discriminant greater or equal than 0 for the inequality (6) to hold. So,

$$1 - (1 + b(n-1))\left(1 - \frac{b(n-1)}{b(n)}\right) \geq 0$$

and then, finally:

$$\begin{aligned}
b(n) &\geq (1 + b(n-1))\,(b(n) - b(n-1)) \Leftrightarrow \\
b(n-1) &\geq b(n-1)\,(b(n) - b(n-1)) \Leftrightarrow \\
b(n) &\leq 1 + b(n-1) \qquad (7)
\end{aligned}$$

Inequality (7) proves that, under the assumption that $\epsilon_n = \lambda^2\alpha b_f^2$, there is no better rate of convergence using these kind of convex constructive solutions that the one obtained in references [4] and [7], since the maximum rate is obtained when

$$b(n) = 1 + b(n-1) \Rightarrow b(n) = b(1) + n - 1 = n \tag{8}$$

Furthermore, for this rate of convergence there is only one zero of the function $P(\lambda)$, namely, $\lambda = (1/n)$ which is the optimal value and coincides with the one provided by Dingankar's algorithm.

We will next answer the questions posed at the beginning of this section, concerning the limits and bounds of the approximation.

4. BOUNDS FOR THE ERRORS

Looking back at expression (5), we will notice that, after using Lemma 2.2, we have at each step a quadratic problem in λ, which consists of minimizing

$$Q(\lambda_n) = (1-\lambda_n)^2 \frac{\delta}{b(n-1)} + \lambda_n^2 b_f^2$$

provided that the induction hypothesis (4) is satisfied for $k < n$. We have introduced the notation λ_n to stress the variation of this parameter along the iterative process.

Taking derivatives, we get

$$\lambda_n b_f^2 = (1-\lambda_n)\frac{\delta}{b(n-1)} \Rightarrow$$

$$\lambda_n = \frac{(1-\lambda_n)\delta}{b_f^2 b(n-1)} = \frac{1+\alpha}{1+\alpha+b(n-1)} \tag{9}$$

Hence, we get the following expression of the optimal error bound:

$$\begin{aligned} & \|f - f_n\|^2 \leq \\ \leq\ & (1-\lambda_n)^2 \left[\frac{\delta}{b(n-1)} + \frac{(1+\alpha)\delta}{b^2(n-1)}\right] + \epsilon_n \\ =\ & \left(\frac{\delta b^2(n-1)}{1+\alpha+b(n-1)}\right)\left(\frac{1+\alpha+b(n-1)}{b^2(n-1)}\right) + \epsilon_n \\ =\ & \frac{\delta}{1+\alpha+b(n-1)} + \epsilon_n = \frac{\delta}{b(n)} \end{aligned} \tag{10}$$

From (10) we can write the following expression for $b(n)$ and ϵ_n:

$$\frac{1}{b(n)} = \frac{1}{1+\alpha+b(n-1)} + \frac{\epsilon_n}{\delta} \tag{11}$$

and then

$$\epsilon_n = \frac{\delta\,[1+\alpha+b(n-1)-b(n)]}{b(n)(1+\alpha+b(n-1))} \tag{12}$$

From this last expression we conclude that there is a fundamental limitation in the rate of convergence that can be achieved under the hypothesis made so far, namely:

$$b(n) - b(n-1) \leq 1+\alpha = \frac{\delta}{b_f^2}$$

Assuming that we can solve the partial approximation problems at each step of the iteration, so $\epsilon_n = 0, n \geq 1$, then

$$b(n) = 1+\alpha+b(n-1) \quad \Rightarrow \quad b(n) = n(1+\alpha) \tag{13}$$

provided that we make $b(1) = 1+\alpha$, which means that we should find an element g_1 in G so that

$$\|f - f_1\|^2 \leq \frac{\delta}{1+\alpha}$$

which is guaranteed by Lemma 2.2. Hence, the best rate of convergence that can be obtained follows the law c/n, since

$$\frac{\delta}{n(1+\alpha)} = \frac{b_f^2}{n}$$

We have then reached the minimum value of the constant c, namely $c = b_f^2$.

Note that for this minimum to be reached we have

$$\lambda = \frac{1+\alpha}{(1+\alpha)n} = \frac{1}{n}$$

so the optimal convex combination would be the average of n elements from G, as in [7].

The remaining problem, namely: given the optimal value of λ find the maximum ϵ_n for a fixed convergence rate, thus making the quasi-optimization problem at each step easier to solve, was already explicitly solved in (12).

Again, to show how our results compare with [5] and [7], we will assume that our desired rate of convergence is given by $b(n) = n$.

The value $\lambda = (1+\alpha)/(n+\alpha)$ solves the optimization problem, and:

$$\epsilon_n = \frac{\alpha\delta}{n(n+\alpha)} \tag{14}$$

This is the best upper bound we can achieve for the partial error at each step of the iteration process. It coincides with Barron's bound, and is always greater than the bound found in [7]. To illustrate this fact, Figure 1 plots the bound ϵ_n for $n = 5$, for $b_f^2 = 1$ as a function of α. Optimal bound, solid line, Dingankar's [7] bound dotted line,

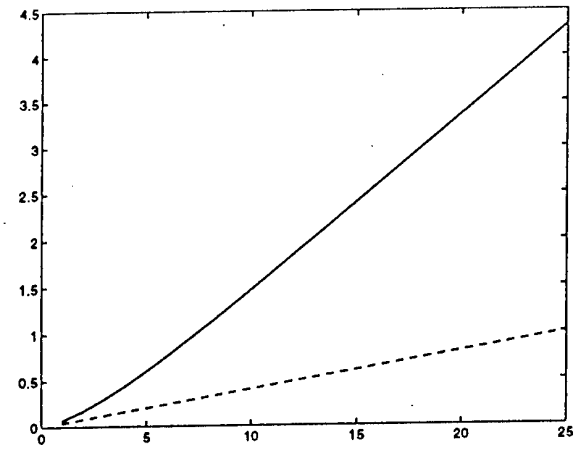


Figure 1: ϵ_n for $n = 5$

5. CONCLUSIONS

We have studied in this paper a theoretical framework where constructive algorithms based on convex combinations of elements from a subset of a Hilbert space can be formulated. We have derived the optimal values for the coefficients in the convex expansions to guarantee a desired convergence rate. We have also studied the trade-off between global and partial errors for that optimal value.

6. REFERENCES

[1] G. Cybenko, "Approximations by superpositions of a sigmoidal function", *Math. Contr. Signals, Syst.*, 1989.

[2] K. Hornik, M. Stinchcombe and H. White, "Multilayer feedforward networks are universal approximators", *Neural Networks*, 2, 1989.

[3] G. Pisier, "Remarks sur un resultat non publié de B. Maurey. *Sem. d'Analyse Fonctionelle*, 1(12) 1980-81. Ecole Polyt., Centre de Math., Palaiseau.

[4] L.K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training", *The Annals of Statistics* vol. 20 (1992)

[5] A.R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function", *IEEE Trans. on Inf. Theory* vol. 39. no.3 (1993)

[6] J.H. Friedman, "Multivariate adaptive regression splines" *The Annals of Statistics* vol. 19 (1991)

[7] A.T. Dingankar and I.W. Sandberg, "A note on Error Bounds for Approximations in Inner Product Spaces", *Circuits, Systems and Signal Processing* vol. 15. no.4 (1996)

[8] L. Breiman, "Hinging Hyperplanes for regression, classification and function approximation", *IEEE Trans. on Inf. Theory* vol. 39. no.3 (1993)

[9] Y. Zhao, "On projection pursuit learning", Ph.D. dissertation, Dept. Math. Art. Intell. Lab., M.I.T., 1992.

[10] F. Girosi and G. Anzellotti, "Convergence rates of approximation by translates", Art. Intell. Lab. Tech. Rep. 1288, M.I.T. 1992.

[11] E.W. Cheney, "Topics in Approximation Theory" *Department of Math*, UTA, (1993)

A HOPFIELD-BASED NEURAL NETWORK FOR THE SHORTEST PATH PROBLEM IN THE ATM COMPETITIVE STRATEGIES

Carlos Bousoño-Calzón †, *Aníbal R. Figueiras-Vidal* ‡

†ESI-Telecom, Universidad de Sevilla
C/Reina Mercedes s/n; 41012 Sevilla, Spain.
Ph:345.455.6872; Fax:345.455.6879; E-mail: cbousono@trajano.us.es
‡EPS-Telecom, Universidad Carlos III
C/Butarque, 15; 28911 Leganés-Madrid, Spain.
Ph:341.336.7226; Fax:341.336.7350; E-mail: anibal@gtts.ssr.upm.es

ABSTRACT

We have developed a Hopfield-based neural network for the shortest path problem and applied it to a competitive strategy for admission of calls and routing of permanent virtual circuits in ATM networks. The speed of this neural network is increased in an order of magnitude respect to previous approaches, and its performance in this competitive strategy is equivalent to the performance of Dijkstra's optimum algorithm.

1. INTRODUCTION

The design of the *Broadband ISDN* requires effective procedures of *call admission* and *routing* to provide a predefined *quality of service* without much understanding of the traffic characteristics. The *Competitive Strategies* provide a framework to address these problems [1].

These strategies can achieve provably good performance without the knowledge of the traffic patterns. Their structure can be resumed as follows:

- The formulation of a cost for every arc in the communication network. This cost is exponential in the current congestion of the arc, which leads to an approximation to the optimal multicommodity flow solution in polynomial time;
- The calculation of the *Shortest Path Problem* (SPP), as presented below, with these costs;
- If the minimum path for the SPP has a cost less than a threshold (to be defined in Sec. 3), the call is accepted and routed through such a path.

The SPP can be shortly formulated as follows [6]. Let G(N,L,C) be a directed graph, where N is a set of nodes, L is a set of ordered pairs of nodes in N (links), and C a cost for links. Undirected graphs can be considered if for every link (na,nb) in L, the reverse link (nb,na) also belongs to L. We define the cost of a path in G as the sum of the costs of the links in such a path. The SPP for a pair of nodes (n1,n2) is to find the path from n1 to n2 with the lowest cost.

Although there exist very efficient algorithms to solve exactly the SPP, like Dijkstra's or Fulkerson's methods [6], it is interesting to consider new approaches with a direct and fast hardware implementation as neural networks, which have already been applied to this problem [3].

The Hopfield model and its applications to optimization problems are finely described in texts as [7]. It consists of a set of nonlinear procesors (often a sigmoid), called *neurons*, connected each other with (usually) symmetric weights. These weights make up a Lyapunov function for the system; this function describes its dynamics towards a set of *attractor points*, which represents the set of solutions for the combinatorial problem.

Two different approaches based on the Hopfield net have been proposed in the literature for the SPP [3]: one of them formulates the problem with a quadratic cost function for the neural network; the other, coming from a remarkable paper by Ali [2], uses a linear cost function. Some advantages of the linear formulation are reported in [2], essentially the reduction of complexity (measured as the number of neurons in the neural net), an increase in the obtained performance, better scalability, and no assumptions about the length of the path restrict the optimality of the solution.

The proposed neural network has also a linear cost formulation inheriting the advantages mentioned before. A further reduction in complexity is accomplished by limiting the neurons in the neural net to those rep-

resenting links actually present in the communication network: then, for a given connectivity (average number of links for every node) in the communication network, the number of nodes in the neural net grows linearly with the size of the communication net, instead of the quadratic growth in Ali's neural net. The speed is also increased respect to Ali's net about an order of magnitude.

2. A HOPFIELD-BASED NEURAL NETWORK FOR THE SPP

The proposed neural network, following [2], is a matrix V of n x n neurons, where element V_{ij} represents the link from node i to node j; after convergence, it can take two logical values, '0' and '1'. In principle, we can assume that there exists a correspondence between neuron values 0 and 1 and logical values '0' and '1', respectively. After convergence, the selected path is defined as the set of links with their neurons in logical state '1'.

In order to separate the validity of solutions from the optimization of the linear cost, we follow [4, 8]: the neural network equations of motion are designed to provide valid solutions, while the linear cost optimization is achieved by setting the initialization point to an appropriate value.

The conditions to get valid solutions can be summarized as follows:

- All neurons representing non-existing links must converge to state '0';
- There has to be one and only one link starting at the source (S) node of a path, and one and only one link ending at the sink (T) node of such a path;
- For every node of a path different from source and sink, there must be one and only one link arriving at this node and one and only one link leaving it;
- No separated loops can exist.

In order to prevent the appearance of non-existing links, links arriving at S, or links leaving T, we define the *Inhibition Matrix* M: for i from 1 to n, $M[i,S] = M[T,i] = 0$; and if $L[i,j] = 0, M[i,j] = 0$.

The equations of motion for the Hopfied network are

$$\forall i,j: \quad dV_{ij} = A\ \partial_t V_{ij}\ M[i,j]\ dt \tag{1}$$

where the $\partial_t V_{ij}$ is different for neurons in Source row and Sink column than for the rest of the neurons in the network, being given by

$\forall i \neq S, j \neq T$:

$$\partial_t V_{ij} = -V_{ij}(1 - V_{ij})(EG1 + EG2 + B\,EG3 + D) \quad (3$$

$\forall j$:

$$\partial_t V_{Sj} = -V_{Sj}(1 - V_{Sj})(EG2 + B\,EG3 - C) \quad (3$$

$\forall i$:

$$\partial_t V_{iT} = -V_{iT}(1 - V_{iT})(EG1 + B\,EG3 - C) \quad (4$$

the rest of the terms being defined as follows

$$EG1 = \sum_p (V_{ip} - V_{pi}) \tag{5}$$

$$EG2 = \sum_p (V_{pj} - V_{jp}) \tag{6}$$

$$EG3 = \sum_{p \neq i} V_{pj} + \sum_{p \neq j} V_{ip} \tag{7}$$

Terms $EG1$ and $EG2$ try to impose that the number of incoming links to a node be equal to the number of outgoing links from that node; these terms are the same as in [2] for the neurons outside Source row or Sink column, but for these lines this constraint is changed, to take into account that it does not apply to Souce and Sink nodes. Term $EG3$ tries to avoid more than one '1' in a row or in a column; it prevents more than one outgoing or more than one incoming link in a node; then, it avoids loops in nodes belonging to the selected path. Note that in [2] the avoidance of loops were left to the linear minimization of the total cost, being possible if the costs for them are zero or very low. Parameter C forces one '1' in the souce row and another '1' in the sink column. Finally, parameter D preserves the network from spurious loops disconnected from the selected path.

We have selected the parameters A, B, C, and D to obtain valid solutions to the routing problem imposing the valid routes to be stable attractors of the Hopfield dynamical system, which is achieved by setting for every valid route:

- $\partial_t V_{ij} \geq 0$, for every neuron V_{ij} that has to be in state '0' to represent a valid route;
- $\partial_t V_{ij} \leq 0$, for every neuron V_{ij} that has to be in state '1' to represent a valid route.

These relations lead immediately to the constrains for parameters that follow:

- $0 \leq C \leq 1$
- $0 \leq B + D$

- $0 \leq D \leq 0$

In order to make the restrictions for parameter D consistent with its role of suppressing spurious loops ($D \neq 0$), it is necessary to redefine the logical states: a neuron V is in state '1' if $V \geq th1$, and it is in state '0' if $V \leq th0$; where $th0$ and $th1$ are thresholds which satisfy $0 \leq th0 \leq th1 \leq k_{min}$; k_{min} being given in Fig. 1 versus D for paths with a maximum number of hops, n, from 3 to 20. The value of these thresholds will also determine the speed of the network and the probability of retrieving valid solutions. For a further discussion about this topic see [5].

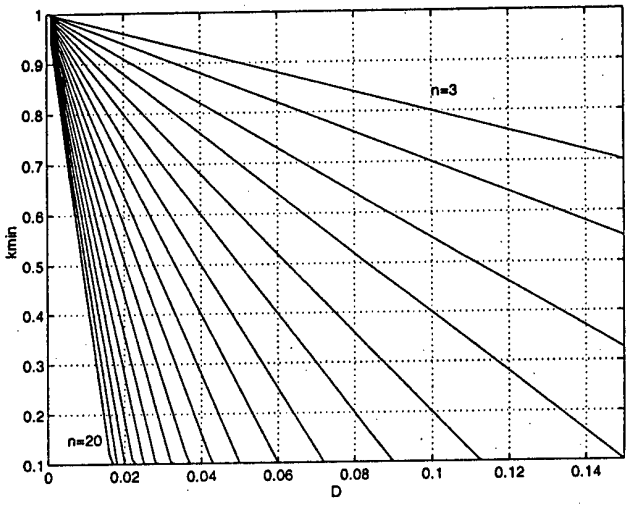


Figure 1: Minimum values to represent logical state '1'.

As the optimization problem is linear, the selection of the initialization point for the neural net can be used as a good heuristic to perform the optimization [8]. This approach allows to separate the problem of optimization from that of obtaining valid and fast solutions without aliasing (i.e., linear superposition of terms for optimization and constraints) in the Lyapunov function formulation. Assuming that costs, a_{ij}, lie in [0,1], the initialization point is selected to be:

$$V_{ij} = (1 - a_{ij}) M[i,j] \eta \qquad (8)$$

where the multiplication by the inhibition matrix makes that neurons representing non valid arcs converge trivially to 0; and η is a penalty for long paths.

The neural algorithm described so far allows a reduction in complexity and an increase of speed about an order of magnitude. The complexity of this net grows, for a given connectivity, as the number of links in the network and not as the square of it. However, the percentage of optimum paths obtained with this net decreases to about the 84% in the experiments.

In order to obtain better performance, a bank of nets is used: the initialization vectors ($V_{ij}^{(k)}$) for the bank of nets are given by

$$V_{ij}^{(k)} = (1 - a_{ij})\, M[i,j]\, \eta + \nu\, n_{ij}[k] \qquad (9)$$

where $\nu = 0.1$ (except one of them for which $\nu = 0$) and $n_{ij}[k]$ is a random number in [0,1].

The total number of neurons in the bank of nets is selected to be equal to the number of neurons in Ali's net, so that the complexity is comparable: the percentage of optimum solutions grows to 97%, the speed of the bank is essentially the speed of one neural net (an order of magnitude bigger than Ali's net), and the robustness obviously grows due to the parallelism.

3. THE COMPETITIVE STRATEGY

We take a simplified model of a competitive strategy for Permanent Virtual Circuit (PVC) routing from Plotkin [1]. The objective of this strategy is maximizing the benefit of the network, which is done by selecting the most profitable set among the incoming requests. As any other competitive strategy, a cost for every link is defined, which is exponential in the current load of the link; once a request is received, the minimum cost path is calculated for it. The call is admitted if the cost of the optimum route is below a threshold, which depends on the profit provided by this call, i.e., if the trade-off between the cost of network resouces and the profit provided by the call is positive for the network.

The communication network is represented by a graph G(V,E,u), where V is the set of nodes, E the set of (directed) edges, and u a function from E to the set of positive real numbers which represents the capacity of the edges.

The traffic in the network is represented by a sequence of requests β_i, where each request β_i is described by $\beta_i = (s_i, t_i, r_i, \rho_i)$: nodes s_i and t_i are the source and the destination for the request i, r_i is the bandwith required by the request, and ρ_i is the profit got if the request is accepted. As we consider only *permanent* circuits, the holding time is infinite; so, no timing parameters of the original model are needed except the order index i.

The *relative load on the edge e*, just before considering the jth request, is defined by

$$\lambda_e(j) = \sum_{e \in P_i, i<j} \frac{r_i}{u(e)} \qquad (10)$$

where P_i is the path for the ith request; the *load* on the network, just before considering the jth request, results

$$\lambda(j) = \sum_{e \in E} \lambda_e(j) \tag{11}$$

Let $Hmax$ be the maximum number of hops allowed for a path in the communication network, and let $\mu = 2\,Hmax+1$; to establish a feasible competitive method, we need to impose some constraints:

- $1 \leq \frac{\rho_i}{r_i} = Hmax$
- $r_i \leq \frac{min_{e \in E} u(e)}{log \mu}$

Then, the competitive strategy for throughput maximization can be formulated as follows:

1. For the ith arriving call to be admitted, it must be checked if there exists a path P from s_i to t_i satisfying the condition

$$\sum_{e \in P} (\mu^{\lambda_e(i)} - 1) \leq \frac{\rho(i)}{r_i} \tag{12}$$

2. If such a path exists, accept the call and route this call on path P satisfying the above condition.

4. SIMULATIONS

Results of simulating the competitive strategy of Section 3 with Dijkstra's optimum algorithm [6] and with Hopfield's algorithm of Section 2 for the SPP computation are presented here: we will refer to both algorithms as OPT and HOP, respectively. The objective is to compare, after saturation of the network(the simulations have been stopped after a consecutive rejection of 100 requests by both algorithms), the obtained profit by both algorithms versus the required resources. Some conclusions are then drawn.

The simulation characteristics are: the communication network has 13 nodes, and a connectivity of 2.62; the capacity of links is normalized to 1; $Hmax$ has been set to 10; parameter r_i is given a random value in [0,0.2], while the source and the destination nodes have been randomly chosen among the nodes in the network.

For the Hopfield algorithm: $A = 1$, $B = 0.9$, $C = 0.7$, $D = 0.05$; $dt = 0.1$; $th0 = 0.1$, $th1 = 0.55$. Note from Fig. 1 that the values selected for $th1$ and D force the maximum number of in a path to be $n = 7$, although this number was never reached in the experiments due to its associated high cost. Since the competitive strategy costs were not constrained to [0,1], they have been mapped to the costs a_{ij}, considered in Ecs. (8) and (9), as

$$a_{ij} = \frac{1}{max - min} cost[i,j] - \frac{min}{max - min} \tag{13}$$

where $cost[i,j]$ is the cost given to the link L[i,j] and max and min are the maximum and minimum of those costs, respectively.

We select the number of nets in the bank to have roughly the same number of neurons (170) than Ali's net (169), so the complexity is comparable. The number of optimum solutions obtained with the net is about 97%. It would be interesting to study the behaviour with the number of nets in the bank to obtain a fixed performance in optimization, to see how the total complexity (measured as the number of neurons in the bank) grows. Nonetheless, it is worth to note that the evolution of the neurons in different nets are not related each other, so it could be expected a better behaviour increasing the number of nets in the bank than increasing the number of neurons in Ali's net.

Fig. 2 makes a global comparison between these algorithms, showing the relative difference of profit between HOP and OPT versus the resources used to allocate the calls (load). These results are averages over 10 realizations. The average number of iterations for our Hopfield net to reach convergence is 240 against the 3000-5000 iterations needed in [2].

Let us discuss briefly these results. The inhibitory matrix reduces the number of active neurons in the Hopfield net to the number of links in the communication network, so that, if the connectivity of the network is fixed, the complexity of the Hopfield net grows as $O(n)$, where n is the number of nodes. Obviously, the use of a bank of nets obviously multiplies the number of neurons by the number of nets in such a bank, but further research is needed to determine how this last number grows with n for a fixed performance. Nontheless, the use of a bank increases the optimization performance without affecting the speed of the whole system, and it provides additional robustness. We repeat that the number of iterations of the neural network to get convergence is an average of 240 iterations versus 3000-5000 iterations in [2]: about 10% of such a number in the previous work of Ali [2].

5. CONCLUSIONS AND FURTHER RESEARCH

We have developed a Hopfield net for the SPP that, compared to existing neural nets for the SPP, provides good scalability, a reduced complexity, and higher speed; but worse quality. If a bank of nets with a complexity comparable to Ali's neural net is consid-

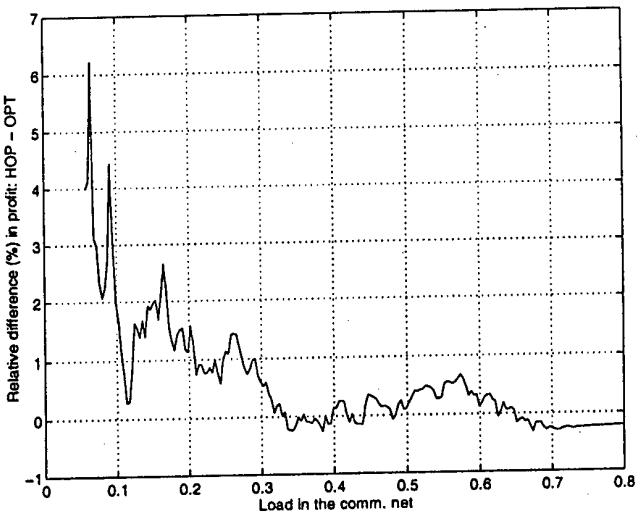


Figure 2: Relative performance of the neural net versus Dijkstra's algorithm.

ered, the obtained quality is equivalent in the competitive strategies to Dijkstra's optimum algorithm, and the speed remains to be about an order of magnitud better than Ali's.

It would be interesting a further understanding of the dynamics of the net as well as about its ability to provide optimal solutions, this issue being closely related to the capacity of the Hopfield net. The development of distributed Hopfield algorithms for routing is also of interest, as the current research in networking seems to suggest.

6. REFERENCES

[1] S. Plotkin, "Competitive Routing of Virtual Circuits in ATM Networks", IEEE J. Select. Areas Comm., vol. 13, no. 6, pp. 1128-1136, Aug. 1995.

[2] H. K. Ali, F. Kamoun, "Neural Networks for Shortest Path Computations and Routing in Computer Networks", IEEE Trans. on Neural Networks, vol. 4, no. 6, pp. 941-953, Nov. 1993.

[3] T. Fritsh, M. Mittler, P. Tran-Gia, "Artificial Neural Net Applications in Telecommunication Systems", Research Report, Institute of Computer Science, University of Wuerzburg, 1992.

[4] C. Bousoño-Calzón, M. Manning, "The Hopfield Neural Network Applied to the Quadratic Assignment Problem", Neural Computing and Applications Journal, vol. 3, pp. 64-72, March 1995.

[5] C. Bousoño-Calzón, *Optimización Combinatoria Basada en el Esquema Neuronal de Hopfield.* Ph.D. Thesis, Univ. Politécnica de Madrid, 1996.

[6] A. V. Aho, J. E. Hopcroft, J. D. Ullman, *The Design and Analysis of Computer Algorithms*; Reading, MA: Addison-Wesley, 1974.

[7] J. Hertz, A. Krogh, R. G. Palmer, *Introduction to the Theory of Neural Computation*; Redwood City, CA: Addison-Wesley, 1991.

[8] W. J. Wolfe et al., "Inhibitory Grids and The Assignment Problem", IEEE Trans. on Neural Networks, vol. 4, no. 2, pp. 316-331, March 1993.

CANONICAL PIECEWISE LINEAR NETWORK FOR NONLINEAR FILTERING AND ITS APPLICATION TO BLIND EQUALIZATION

Tülay Adalı and Xiao Liu

Information Technology Laboratory
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County,
Baltimore, MD 21228-5398, U.S.A
Tel: (410) 455-3521, Fax: (410) 455-3969
adali@engr.umbc.edu xliu@engr.umbc.edu

ABSTRACT

The canonical piecewise linear structure is used for nonlinear filtering and its nonlinear approximation capacity is shown by utilizing piecewise linear partions. It is also shown that CPL network can approximate a given nonlinear continuous function with any degree of accuracy. This result is extended to show that CPL networks can be used to equalize a nonlinear channel. We show that if the distribution of the output of equalizer is the same as the distribution of the sequence at the input of the nonlinear channel, then the global system is identity (except for a sign factor) under some regularity conditions. Thus, distribution learning [1] by maximizing an appropriate objective function achieves nonlinear channel blind equalization.

1. INTRODUCTION

The traditional theory of signal processing is established upon the assumption of linearity. The linear filter with or without feedback whose output is a linear combination of the input signal is both easy to implement and to analyze, and has been widely used in a variety of applications. However, most practical systems are better approximated by nonlinear models. Also, the growing demand for signal processing in more demanding environments to achieve very high data rates has driven the need to improve on existing methods using nonlinear filters. Even simple nonlinear models can successfully capture types of system behavior which would not be possible to describe with linear models. However, nonlinear models also offer great challenges because of their complex structure and dynamics. The statistical problems of model identifications are similarly more intricate and the introduction of nonlinearity into the filter's operation will lead to an increase in the implementation complexity.

Several approaches have been proposed for the estimation and characterization of the nonlinear model. Polynomial or Volterra filters [13] assume that the nonlinear function can be represented as an expansion of polynomial terms, and are probably the best known method of nonlinear filtering. It turns out that although sufficiently high order polynomials can yield a small asymptotic probability of error, they will also, in general, converge very slowly. Neural network structures, such as multilayer perceptron and the radial basis functions, have also been introduced as nonlinear filters (e.g. [6]), but they require a large amount of training time and large network sizes. Recurrent neural network (RNN) equalizer [10] can accurately model the inverse of a nonlinear communication channel with smaller network size, but because of its complex nonlinear structure, its dynamics are very hard to explain. Also, for blind equalization, it is quite difficult to incorporate statistics information into the network structure because of the highly nonlinear structure of a general RNN.

Piecewise linear models constitute a comprise between the complexity of the nonlinear approximation and the theoretical abundance of the linear domain. Piecewise linear models have been proven very useful in control engineering [14], nonlinear circuit analysis, and other nonlinear problems [3]. Here, we study the approximation and dynamical properties of a special kind of piecewise linear function, canonical piecewise linear (CPL) network [3], present its application to nonlinear filtering, particularly for blind equalization of nonlinear channels. CPL network offers the following benefits: (1) it makes use of standard linear adaptive filtering techniques to perform training tasks and allows for efficient selection of the partition boundaries; (2) it offers

savings in computation time and implementation cost, especially when required to model strong nonlinearities; (3) because of its piecewise linear nature, it allows for easy incorporation of known statistical information into the network structure.

In this paper, we first prove the nonlinear approximation capacity of CPL network by utilizing piecewise linear partions and show that CPL network can approximate a given nonlinear continuous function with any degree of accuracy, and explain dynamics of learning on the CPL network. We then extend this result to show that CPL networks can be used to equalize a nonlinear channel, and present a proof of this ability of the CPL equalizer for a general nonlinear channel. The theoretical results reported for this property have always assumed a linear distorting channel characteristics [2],[4]. We show that if the distribution of the output random variable of equalizer is the same as the the distribution of the sequence at the input of the nonlinear channel, then the global system is identity (except for a sign factor) under some regularity conditions. Thus, distribution learning [1] by maximizing an appropriate objective function achieves nonlinear channel blind equalization.

2. REPRESENTATION AND CAPACITY OF CPL NETWORK

The CPL network is defined as [3]:

Definition 1 (*Canonical Piecewise-Linear Function*): A piecewise linear function f: $D \rightarrow Q$ with a compact subset $D \subset R^N$ and compact subset $Q \subset R^M$ is called a canonical piecewise linear (CPL) function if it can be expressed by a *global* representation

$$f(\mathbf{x}) = \mathbf{a} + \mathbf{B}\mathbf{x} + \sum_{i=1}^{\tau} \mathbf{c}_i \, |\langle \alpha_i, \mathbf{x} \rangle + \beta_i| \qquad (1)$$

where $\mathbf{B} \in R^{M \times N}, \mathbf{a}, \mathbf{c}_i, \alpha_i \in R^N$ and $\beta_i \in R$.

CPL representation only requires a minimal amount of memory space for storing the parameters of multidimensional piecewise linear functions. Since the domains of the functions describing such models are partitioned into polyhedral regions where the functions are linear throughout, all the nonlinearities are localized in the region boundaries. This makes them much more amenable for analysis than virtually any other type of nonlinear functions. Moreover, the class of CPL function is closed since the composition and inverse (if it exists) of CPL functions is again a canonical piecewise linear function [9]. Thus, CPL function provides us with minimum number of boundaries to partition the training patterns, and in each partitioned region, we can use a linear model to approximate the given mapping. While the representation capability of (1) has been studied in [3], this network was proposed and used without proof of its approximation capability. In [11], CPL network is considered as a special case of the multilayer perceptron model, hence claiming its approximation ability, however no attempt is made to construct a proof to demonstrate the claim and the connection with the multilayer perceptron model.

Before discussing the capacity of CPL network, we introduce following definitions and lemmas:

Definition 2 (*Nondegenerate Partition*) partition

$$\langle \alpha_i, \mathbf{x} \rangle + \beta_i = 0 \quad i = 1, 2, \cdots, q$$

is said to be *nondegenerate* if for every set of linearly dependent vectors $\alpha_{i_1}, \alpha_{i_2}, \cdots, \alpha_{i_m}$, $m \leq q$, the rank of the matrix $[\alpha_{i_1}, \alpha_{i_2}, \cdots, \alpha_{i_m}]$ is strictly less than the rank of $(N+1) \times m$ matrix:

$$\begin{bmatrix} \alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_m} \\ \beta_{i_1}, \beta_{i_2}, \ldots, \beta_{i_m} \end{bmatrix}$$

Definition 3 *Consistent Variation* [3]: A function f: $D \rightarrow Q$ with a compact subset $D \subset R^N$ and compact subset $Q \subset R^M$ is said to possess the *consistent variation* property if and only if

1) f has a linear partition,

2) $\mathbf{J}_{R_{i1}+} - \mathbf{J}_{R_{i1}-} = \mathbf{J}_{R_{i2}+} - \mathbf{J}_{R_{i2}-} = \ldots = \mathbf{J}_{R_{in_i}+} - \mathbf{J}_{R_{in_i}-} = \cdots = \mathbf{c}_i \alpha_i^T, i = 1, 2, \ldots, q$ where $\mathbf{J}_{R_{ij}+}$ and $\mathbf{J}_{R_{ij}-}$ denote the Jacobian matrices of the regions $R_{ij}{}^+$ and $R_{ij}{}^-$ respectively, which are separated by the boundary

$$\langle \alpha_i, \mathbf{x} \rangle + \beta_i = 0$$

Here, $j = 1, 2, \ldots, q$, and q pairs of regions are separated by this boundary such that $\langle \alpha_i, \mathbf{x} \rangle + \beta_i \geq 0$ for $\mathbf{x} \in R_{ij}{}^+$ and $\langle \alpha_i, \mathbf{x} \rangle + \beta_i \leq 0$ for $\mathbf{x} \in R_{ij}{}^-$.

Lemma 1 [3]: If the domain space of a continuous piecewise linear function f is partitioned by a set of nondegenerate linear partition boundaries, then f has the consistent variation property.

Lemma 2 (*Necessary and Sufficient Condition*) [3]: A piecewise linear function f has a canonical piecewise linear representation if and only if it possesses the consistent variation property.

We prove the following theorem based on the nondegeneration partitions property:

Theorem 1: Let domain D be a compact space of N dimensions and $\mathcal{F}$ be a set of canonical piecewise linear functions on D. Then, for any continuous function $\bar{f}$ on D, there exists a function $f \in \mathcal{F}$ such that $|f(\mathbf{x}) - \bar{f}(\mathbf{x})| < \epsilon$ for all $\mathbf{x} \in D$

Proof. Since $\bar{f}$ is continuous on D, then for every $\xi \in D$, there is a sequence $\{I_{n\xi}\}$ of closed intervals containing ξ, whose lengths converge to zero, such that

$$||\bar{f}(\mathbf{x}) - \bar{f}(\xi)|| < \frac{\epsilon}{2} \tag{2}$$

Then, the set of intervals

$$I = [I_{n\xi} | \xi \in D, n = 1, 2, \cdots]$$

covers D in the *Vitali* sense [7]. Hence, by *Vitali Covering Theorem* [7], there are a finite number of disjoint intervals $I_{n_1\xi_1}, I_{n_2\xi_2}, \cdots, I_{n_k\xi_k}$ in I covering all of D, and, in each interval $I_{n_i\xi_i}$ condition in (2) always holds for $i = 1, 2, \cdots, k$. On the other hand, because of the continuity of linear function, there exist Jacobian matrices $\mathbf{J}_{n_i\xi_i}$ and vectors $\mathbf{w}_{n_i\xi_i}$ such that

$$||\mathbf{J}_{n_i\xi_i}\mathbf{x} + \mathbf{w}_{n_i\xi_i} - \bar{f}(\xi_i)|| < \frac{\epsilon}{2} \quad \mathbf{x} \in D, \quad i = 1, 2, \cdots, k \tag{3}$$

and they have the same value at the common boundary. Define $f(\mathbf{x})$ as follows:

$$f(\mathbf{x}) = \mathbf{J}_{n_i\xi_i}\mathbf{x} + \mathbf{w}_{n_i\xi_i} \quad \mathbf{x} \in I_{n_i\xi_i}$$

then, from (2) and (3), we have

$$||f(\mathbf{x}) - \bar{f}(\mathbf{x})|| < \epsilon \quad \mathbf{x} \in D$$

Since $I_{n_i\xi_i}, i = 1, 2, \cdots, k$ are closed intervals in D, they can be obtained by partitioned D with a set of nondegenerate partitions. Thus, by Theorem 2.1, the piecewise linear function $f(\mathbf{x})$ possesses a CPL representation.

Hence, any nonlinear channel can be represented as a CPL function, and furthermore, if we use a CPL network as an equalizer, then, the global system is still a CPL function.

3. BLIND EQUALIZATION BY CPL NETWORK

Blind equalizers are a special class of equalizers that determine their parameters based on the statistics of the channel input and the measured output when training sequences are not accessible. Since almost all of blind equalizers such as [5] are developed for linear channels, the use of these algorithms will suffer from a severe performance degradation for unknown nonlinear channels. A RNN-based blind equalizer is proposed in [10], however, the equalizer does not guarantee the consistency of the cost function utilized. In this work, we establish the consistency of the estimation scheme by using a CPL equalizer for nonlinear channels as follows:

Let the global system $\mathcal{T} = \mathcal{F}(\mathcal{S})$ (a cascade of a nonlinear channel $\mathcal{S}$ and the CPL equalizer $\mathcal{F}$) be the CPL network (1) and $\{x(n)\}$ be an *i.i.d.* random variable with distribution ν. $x(n) \in \Omega$. Assume that CPL network (1) divides the input space into m disjoint regions, $R_1, R_2, \cdots, R_m$, and in each region R_i, (1) is equivalent to the following linear model:

$$M_i: \qquad \tilde{x}(n) = \sum w_{ij} x_j(n) \tag{4}$$

The basic assumptions on $\mathcal{T}$ and ν are the following:

(i) The distribution ν is symmetric with finite variance.

(ii) For each model M_i, there exists at least a subset $I \subset \Omega$ and a region R_i, $\tilde{I} = I \times I \cdots \times I \subset R_j$ such that the mapping

$$\tilde{x}(n) = \sum w_{ij} x_j(n)$$

is a $\tilde{I}$ onto I mapping.

Then, we have the following conclusion:

Theorem 2: Consider the global system $\mathcal{T} = \mathcal{F}(\mathcal{S})$. We assume that $\{x(n)\}$ is an *i.i.d.* process with distribution ν and assumptions (i) and (ii) are satisfied. If the distribution of $\{\tilde{x}(n)\}$ is still ν, then, the global system $\mathcal{T}$ is identity except for a possible delay and a sign factor.

Proof. Since $\{x(n)\}$ and $\{\tilde{x}(n)\}$ have the same distribution and ν is symmetric, then, $E\{x(n)\} = \{\tilde{x}(n)\} = 0$ and

$$\int_{I+I'} x^2 d\nu_x = \int\int \cdots \int_{\tilde{I}+\tilde{I}'} \tilde{x}^2 d\nu_{x_1} d\nu_{x_2} \cdots d\nu_{x_k}$$

where $I' = -I$, $\tilde{I}' = (-I) \times (-I) \cdots \times (-I)$. Then,

$$\int_{I+I'} x^2 d\nu_x = \sum_{j=1}^{k} \int\int \cdots \int_{\tilde{I}+\tilde{I}'} {w_{ij}}^2 {x_j}^2 d\nu_{x_1} d\nu_{x_2} \cdots d\nu_{x_k}$$

i.e

$$\int_{I+I'} x^2 d\nu_x = \rho^{k-1} \sum {w_{ij}}^2 \int_{I+I'} x^2 d\nu_x$$

here, $\rho = \int_I d\nu_x$. We can easily obtain that $\rho^{k-1} \sum {w_{ij}}^2 = 1$. Let f be the characterization function of $x(n)$ on I, we have [8]

$$f(\tau) = \prod f(w_{ij}\tau) \tag{5}$$

Let $g = |f|$ and G be the distribution function corresponding to g. From (5),

$$g(\tau) = \prod g(w_{ij}\tau)$$

Setting $\psi(\tau) = -\ln g(\tau)/\tau^2$, then,

$$\psi(\tau) = \sum {w_{ij}}^2 \psi(w_{ij}\tau)$$

which can be rewritten as

$$\sum \rho^{k-1} {w_{ij}}^2 [\psi(\tau) - \psi(w_{ij}\tau)/\rho^{k-1}] = 0 \tag{6}$$

It follows from (6) that, for any τ, there exists at least one w_{ij}, such that $\psi(\tau) \leq \psi(w_{ij}\tau)/\rho^{k-1}$. Since ν has finite variance, then $\psi(0)$ exists and we get

$$\psi(0) \leq \psi(0)/\rho^{k-1} \tag{7}$$

From (6), it also follows that for every τ, there exists a q such that $\psi(\tau) \geq \psi(w_{iq}\tau)/\rho^{k-1}$. therefore, we have

$$\psi(0) \geq \psi(0)/\rho^{k-1} \tag{8}$$

We know that (7), (8) hold if and only if $\rho^{k-1} = 1$, i.e $k = 1$. This means that the model M_i has only one non-zero coefficient w_{il} and $w_{il}^2 = 1$. Thus proves the theorem.

Hence, in order to obtain the solution for the blind equalization problem, we have to adjust the tap values of the CPL equalizer in such a way that the instantaneous distribution of the output $\tilde{x}(n)$ of the equalizer converges to the input distribution ν. Several cost functions such as moment error objective function, generalized Godard/Sato objective function and Shannon entropy can be used for distribution learning. In the next section, We present a blind equalization algorithm based on the moment error objective function and partial likelihood function. We present simulation results that show that the CPL blind equalizer outperforms the constant modulus algorithm [5],[15] by orders of magnitude when equalizing nonlinear channels.

4. EXAMPLE

Assume that the only available information is channel observations: $y_n, y_{n-1}, \cdots, y_0$, and the statistics of channel input $E[{x_n}^j], j = 1, \cdots, 4$. Let $p_w(x_n|\mathcal{F}_n)$ be the estimated probability mass function (pmf) of the input sequence x_n, where $\mathcal{F}_n$ is the σ-field generated by events $[y_n, y_{n-1}, \cdots, y_1, y_0]$. By [1], a distribution learning can be achieved by maximizing partial log-likelihood function , i.e., $\max_w \sum_{i=1}^{i=n} \ln p_w(x_i|\mathcal{F}_i)$. The true channel input x_i is not known, but it can be shown that the maximization of partial likelihood is equivalent to the maximization of quasi-partial log-likelihood function $\sum_{i=1}^{i=n} \ln p_w(\bar{x}_i|\mathcal{F}_i)$, $\bar{x}_i \in \mathcal{S}$ with respect to the ω and $\bar{x}_i \in \mathcal{S}$. Thus, By using this conclusion, the blind equalization algorithm is given as follows:

- Start with an initial estimate of ω
- Maximize quasi-partial log-likelihood function with respect to $\bar{x}_i$
- Maximize quasi-partial log-likelihood function with respect to ω based on the updated $\bar{x}_i$
- Repeat these steps until the algorithm converges

For the binary communication channel, $\mathcal{S} = \{-1, 1\}$, we choose

$$p_w(\bar{x}_n|\mathcal{F}_n) = \begin{cases} \frac{e^{-J(x_n=1)}}{e^{-J(x_n=1)}+e^{-J(x_n=-1)}} & \text{if } \bar{x} = 1 \\ 1 - \frac{e^{-J(x_n=1)}}{e^{-J(x_n=1)}+e^{-J(x_n=-1)}} & \text{if } \bar{x} = -1 \end{cases}$$

where

$$J(\bar{x}_n) = \sum_{j=1}^{4} \pi_j (E[\hat{x}_n^j] - E[x(n)^j])^2$$

$$E[\hat{x}_n^j] = \frac{1}{n}((n-1)E[\hat{x}_n^j] + \hat{x}_n^j), \quad \hat{x}_n^j = (\bar{x}_n^j + \hat{x}(n))/2$$

$\hat{x}(n)$ is the output of CPL equalizer, and π_j are positive constants. Our blind algorithm and CMA algorithm [5], [15] are tested for equalization of the nonlinear channel

$$G(z) = 1 + 0.7z^{-1} + 0.15(1 + 0.7z^{-1})^2$$

$$+0.1(1 + 0.7z^{-1})^3 + 0.05(1 + 0.7z^{-1})^4$$

Even for this relatively simple communication channel, the CMA based equalizer exhibits a very poor performance. In Figure 1, we plot the BER curves for both equalizers. The results show that CPL based blind equalizer outperforms the CMA equalizer by orders of magnitude. If we choose the test channel as $G(z) = 1 + 0.7z^{-1}$, Figure 2 shows that CPL blind equalizer and CMA algorithm have comparable performance.

5. CONCLUSIONS

canonical piecewise linear structure is introduced as blind equalizer in this paper. The mapping ability CPL network is studied. A methodology to study ntically distribution and blind equalization is pre- ted for a global CPL system. A blind algorithm is ived based on the moment error objective function l partial likelihood function. The simulation results nonstrate that CPL based equalizer outperforms the IA equalizer by orders of magnitude when equalizing ılinear channel and they perform similarly in linear ınnel equalization.

6. REFERENCES

T. Adalı, X. Liu, and M. K. Sönmez, "Conditional distribution learning with neural networks and its application to channel equalization," to appear *IEEE Trans. Signal Processing.*

A. Benveniste, M. Goursat, and G. Ruget, "Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications," *IEEE Trans. Automatic Contr.*, 1980, vol. AC-25, pp. 385-399.

L. O. Chua and A. Deng, "Canonical piecewise-Linear networks," *IEEE Trans. Circuits Syst.*, vol. 35, No. 1, pp. 101–111 , Jan. 1988.

D. Donoho, "On minimum entropy deconvolution," *Applied Time Series Analysis II*, ed. D. Finley, Academic Press, New York, 1981, pp. 556-608.

D. N. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. on Communications*, 1980, vol. COMM-28, pp. 1867-1875.

G. J. Gibson, S. Siu, and C. F. N. Cowan, "The application of nonlinear structures to the reconstruction of binary signals," in *IEEE Trans. Signal Processing*, vol. 39, no. 8, pp. 1877-1884, Aug. 1991.

C. Goffman, "Real Functions," Prindle, Weber & Schmidt, Incorporated. 1953.

A. M. Kagan, Y. V. Linnik, and C. R. Rao, "Characterization Problems in Mathematical Statistics," John Wiley & Sons, Inc., 1973.

C. Kahlert and L. O. Chua, " The complete canonical piecewise-linear representation-part I: the geometry of the domain space," *IEEE Trans. Circuits Syst.*, vol. 39, No. 3, pp. 222–236, Jan. 1992.

G. Kechriotis, E. Zervas, and E. S. Manolakos, "Using recurrent neural networks for adaptive communication channel equalization," *IEEE Trans. on Neural Networks*, vol. 5, no. 2, pp. 267-278, March 1994.

J. Lin and R. Unbehauen, "Canonical piecewise-linear networks," *IEEE Trans. on Neural Networks*, vol. 6, No. 1, January 1995

[12] X. Liu and T. Adalı, "Channel equalization using partial likelihood estimation and recurrent canonical piecewise-Linear neural network," to appear in *Proc. 1996 European Conf. on Signal Processing*, Trieste, Italy, Sep. 1996.

[13] V. J. Mathews, "Adaptive polynomial filters," *IEEE Trans. Signal Processing Mag.*, pp. 10-26, July 1991.

[14] N. B. O. L. Pettit and P. E. Wellstead, " Analyzing piecewise linear dynamic systems," *IEEE Control Systems magazine*, vol. 8, pp. 43-50, 1995.

[15] J.R. Treichler and B.G. Agee, "A new approach to multipath correction of constant modulus signals," in *IEEE. Acoust., Speech and Signal Process.*, vol. 31, no. 2, pp. 459-471, 1983.

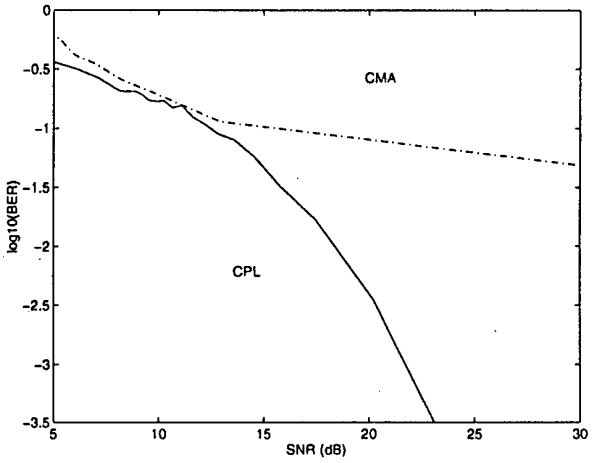


Figure 1: 2-PAM Nonlinear Blind Equalization

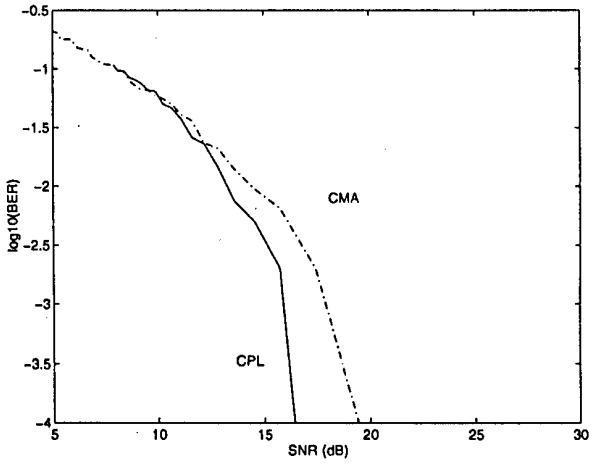


Figure 2: 2-PAM linear Blind Equalization

DISTRIBUTION LEARNING BY PARTIAL LIKELIHOOD ESTIMATION AND DYNAMICS OF RELATIVE ENTROPY MINIMIZATION

Tülay Adalı[1] *and Xiao Liu*[1] *and M. Kemal Sönmez*[2]

[1]Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County, Baltimore, MD 21228-5398
[2]Institute for Systems Research, University of Maryland, College Park, MD 20742

ABSTRACT

We introduce a unified statistical framework for real-time signal processing with neural networks by using a recent extension of maximum likelihood (ML) estimation, partial likelihood (PL) estimation theory, which allows for (i) dependent observations and (ii) sequential processing. For a general neural network conditional distribution model and for the general case of dependent observations, we establish a fundamental information-theoretic relationship for PL estimation, show its equivalence to relative entropy minimization. We study the dynamics of relative entropy minimization (maximum partial likelihood estimation) within the *well-formed* cost functions framework, show that these are well-formed cost functions, hence their gradient descent minimization is guaranteed to converge to a solution if one exists. The formulation is applied to adaptive channel equalization and simulation results are presented to show the ability of the least relative entropy equalizer to realize complex decision boundaries and to recover during training from convergence at the wrong extreme in cases where the mean square error based MLP equalizer can not.

1. INTRODUCTION

Statistical parameter estimation theory has as its fundamental support maximum likelihood (ML) estimation which provides estimators with nice large sample optimality properties and invariant with respect to functions of the parameters. However, ML theory is traditionally developed for independent observations, and a majority of signal processing applications require processing of dependent observations. In this paper, we introduce a conditional distribution learning framework for real-time signal processing with neural networks based on partial likelihood (PL) theory [4], [10]. Obtained as a partial factorization of the full likelihood, PL also possesses nice large sample properties of ML, and more importantly, it can easily be characterized for dependent data and sequential processing. Hence, it overcomes the difficulties with other extensions of ML for dependent data, such as conditional likelihood which, for easy specification, requires that the auxiliary information be known for the whole period (i.e. including future observations) [7]. Some of the other problems with other factorizations of likelihood for dependent data are initial state specification requirements (e.g. when using Markovian representations for the data) and the problems when dealing with missing data. Therefore, PL provides us with a particularly suitable formulation for real-time signal processing which most of the time requires on-line processing of dependent observations.

We introduce a general neural network conditional probability model, and for this model, establish a key information-theoretic connection, namely the equivalence of maximum PL estimation and accumulated relative entropy (ARE) minimization. Hence, distribution learning using relative entropy between the true and estimated probability mass functions can be achieved by maximum PL estimation which does not require that the true conditionals be known (which in general are not available). This result can be regarded as the extension of the ML and minimum ARE equivalence for independent and identically distributed (i.i.d.) data [9] to the general case of dependent observations. While providing the theoretical foundation for statistical analysis of maximum PL estimation, this connection can also be used to derive a new class of real-time signal processing algorithms based on information-theoretic alternating projections [3].

We then consider a perceptron probability model for binary distribution learning and its application to adaptive channel equalization. For the MLP model, we derive the least relative entropy (LRE) algorithm by gradient optimization and show that it possesses nice dynamical properties which can be beneficial to the channel equalization problem. Particularly, it is

shown that LRE can always recover from convergence at the wrong extreme whereas the mean-squared-error (MSE) based gradient descent learning on MLP can not. This property of the algorithm is discussed within the *well-formed* cost functions framework of Wittner and Denker, [8] stating that gradient descent learning on such cost functions is always guaranteed to find a solution if one exists. In a gradient descent dynamics framework, it has been shown that MSE cost function is not a *well-formed* cost function [8], therefore finding a solution can not always be guaranteed. With MLPs, it is also often the case that the MSE based learning algorithms can not recover from convergence at a wrong extreme.

2. DISTRIBUTION LEARNING BY PARTIAL LIKELIHOOD ESTIMATION

Consider the discrete valued sequence $\{x_k\}$ taking values from the alphabet $\mathcal{S} = \{a_0, a_1, ..., a_M\}$. Define the σ-field $\mathcal{F}_{n-1} = \sigma\{1, x_{n-1}, \cdots, x_1, x_0; y_n, \cdots, y_1, y_0\}$ where the inclusion of 1 is only a mathematical convenience. We parametrize the conditional probability mass function (pmf) $p_\theta(x_n|\mathcal{F}_{n-1})$ by a neural network as follows:

$$p_\theta(x_n|\mathcal{F}_{n-1}) = f(x_n, g(\mathbf{y}_n^N, \theta)). \qquad (1)$$

Here, θ is the vector of network weights, $\theta \in \Theta$ where Θ is a compact parameter set, and $\mathbf{y}_n^N = [y_n, y_{n-1}, \cdots, y_{n-N+1}]$. The term $g(\mathbf{y}_n^N, \theta)$ is the output of the neural network, and $f(\cdot)$ and $g(\cdot)$ are continuous and differentiable functions. Since $\mathcal{F}_n$ includes the entire history the network can assume a *recurrent* structure as well. The task is then to estimate the conditional pmf $p_\theta(x_n|\mathcal{F}_{n-1})$, or the conditional probabilities

$$P_\theta(x_n = a_i|\mathcal{F}_{n-1}) \quad \forall a_i \in \mathcal{S}.$$

In (1), $f(\cdot)$ is included to account for the functional representation of the pmf using the M conditional probabilities. Also, an additional constraint on (1) is that, $f(\cdot)$ has to be chosen such that $\sum_{a_j \in \mathcal{S}} P_\theta(x_n = a_j|\mathcal{F}_{n-1}) = 1$. Neural network learning (extracting the information represented by the data through adaptation of the weights θ) can now be viewed as a distribution learning problem, i.e. estimation of the parameters of the conditional pmf, such that the PL function given by

$$\mathcal{L}^p(\mathbf{x}_n; \theta) \equiv \mathcal{L}_n^p(\theta) = \prod_{i=1}^{n} p_\theta(x_i|\mathcal{F}_{i-1}). \qquad (2)$$

is maximized.

3. EQUIVALENCE TO RELATIVE ENTROPY MINIMIZATION

The relative entropy (RE), or the Kullback-Leibler distance [6], is a fundamental information theoretic measure of how accurate the estimated conditional pmf $p_\theta(x_n|\mathcal{F}_{n-1})$ is an approximation to the true conditional pmf $p_{\theta_0}(x_n|\mathcal{F}_{n-1})$ (assuming it is realized by θ_0 for (1)). We define *accumulated* relative entropy (ARE) at time n as $\mathcal{I}_n(\theta) = \sum_{k=1}^{n} i_k(\theta)$ where the relative entropy distance: $D_k(p_{\theta_0}\|p_\theta) \equiv i_k(\theta) = E\{r_k(\theta)|\mathcal{F}_{k-1}\}$, and $\mathcal{J}_n(\theta) = \sum_{k=1}^{n} j_k(\theta)$ where $j_k(\theta) = Var\{r_k(\theta)|\mathcal{F}_{k-1}\}$ with $r_k(\theta) = \ln \frac{P_{\theta_0}(x_k=a_j|\mathcal{F}_{k-1})}{P_\theta(x_k=a_j|\mathcal{F}_{k-1})}$.

Based on the theory of PL [10], we establish the relationship between MPL estimation and ARE minimization by the following theorem:

Theorem: If there exist a constant $\delta > 0$ and continuous functions $f(\cdot)$ and $g(\cdot)$, such that, for each $\theta \neq \theta_0$, as $n \to \infty$,

$$P(\mathcal{I}_n(\theta)/n > \delta) \longrightarrow 1 \qquad (3)$$

and

$$\mathcal{J}_n(\theta)/n^2 \longrightarrow 0 \text{ in probability} \qquad (4)$$

then

$$\arg\min_\theta \mathcal{I}_n(\theta) - \arg\max_\theta \tilde{\mathcal{L}}_n^p(\theta) \longrightarrow 0 \qquad (5)$$

almost surely on $\Omega = \{\mathcal{I}_n(\theta) \uparrow \infty, \sum_{i=1}^{n} j_i(\theta)/\mathcal{I}_i^2(\theta) < \infty\}$ where $\tilde{\mathcal{L}}_n^p(\theta) \equiv \ln \mathcal{L}_n^p(\theta)$.

Note that the first condition of the theorem, (3), represents the rate by which the Kullback-Leibler information accumulates with n, and guarantees that for each $\theta \neq \theta_0$, $\mathcal{I}_n(\theta) \to \infty$ as $n \to \infty$, i.e. the information continues to accumulate. The second condition, (4), on the other hand implies asymptotical stability of variance. Thus, the maximum PL estimate $\hat{\theta}$ also minimizes ARE distance between the true and estimated conditional distributions asymptotically providing an estimate of the true parameter θ_0. We emphasize the fact that the result holds for the general case of dependent observations and hence provides a generalization of the ML and ARE equivalence which is shown for independent observations, [9]. Proof of this theorem is given in [2]. In [2], we also establish consistency and asymptotic normality of partial likelihood estimates for the conditional probability model of (1).

4. PERCEPTRON CONDITIONAL PROBABILITY MODEL

If we consider the special case where the sequence x_n takes values from the binary alphabet $\mathcal{S} = \{0, 1\}$, the

problem reduces to estimation of the conditional probability $P(x_n = 1|\mathcal{F}_{n-1})$. The PL function is then characterized as:

$$\mathcal{L}_n^p(\theta) = \prod_{i=1}^{n} P_\theta(x_i = 1|\mathcal{F}_{i-1})^{x_i} \left(1 - P_\theta(x_i = 1|\mathcal{F}_{i-1})\right)^{1-x_i}. \tag{6}$$

Consider the following single hidden layer MLP structure as the conditional pmf model:

$$P_\theta(x_n = 1|\mathcal{F}_{n-1}) = g\left(\sum_{i=1}^{q} h\left(\mathbf{y}_n{}^T \mathbf{w}^i\right) v^i\right) \tag{7}$$

where $\mathbf{w}^i \in \mathbf{R}^{N\times 1}$ is the weight vector between the inputlayer and the hidden node i, $(i = 1, \cdots, q$, where q is the number of hidden nodes), $\mathbf{y}_n \in \mathbf{R}^{N\times 1}$ is the observation vector, and v^i is the weight between the hidden node i and the output node. We represent the entire set of weights by $\theta = [\mathbf{W}, \mathbf{v}] \in \mathbf{R}^{q\times N+1}$ where $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \cdots, \mathbf{w}^q]^T \in \mathbf{R}^{q\times N}$ and $\mathbf{v} = [v^1, v^2, \cdots, v^q]^T \in \mathbf{R}^{q\times 1}$. The hidden node activation function $h(\cdot)$ is chosen to ensure network approximation capabilities [9], e.g. it can be chosen as the familiar logistic or the radial basis function. However, for learning parameters by gradient descent minimization, note that $g(\cdot)$ has to be chosen such that $g'(\cdot) > 0$.

If we choose both $g(\cdot)$ and $h(\cdot)$ as sigmoidal function, i.e.

$$g(\mathbf{s}_n^T \mathbf{v}) = \frac{1}{1 + \exp(-\mathbf{s}_n^T \mathbf{v})} \tag{8}$$

where

$$s_n^i = \frac{1}{1 + \exp\left(-\mathbf{y}_n^T \mathbf{w}^i\right)}$$

for $i = 1, \cdots, q$, $\mathbf{s}_n = [s_n^1, s_n^2, \cdots, s_n^q]^T$, gradient descent minimization of the negative log PL cost function results in the following updates:

$$v_{n+1}^i = v_n^i + \mu_1 s_n^i e_n \tag{9}$$

$$\mathbf{w}_{n+1}^i = \mathbf{w}_n^i - \mu_2 \mathbf{y}_n g(s_n^i)\left(1 - g(s_n^i)\right) v_n^i e_n \tag{10}$$

for $i = 0, \cdots, q$, where $s_n^i = g\left(\mathbf{y}_n^T \mathbf{w}_n^i\right)$ and $e_n = x_n - g(\mathbf{s}_n^T \mathbf{v}_n = x_n - P_\theta(x_n = 1|\mathcal{F}_{n-1})$ where μ_1 and μ_2 are the step sizes at the output and hidden layers respectively.

The binary equalization problem can be rephrased as follows in order to comply with the development in [8]. For the remainder of the section, assume that the nonlinearity is the hyperbolic tangent, an odd function, without loss of generality. Therefore, $p_\theta(x_n = 1|\mathcal{F}_n) \in (-1, 1)$. (Note that the transformation to the probability measure is immediate by the application of transformation: $\frac{1}{2}[(\cdot) + 1]$).

Divide the training set $\mathcal{Y} = \{\mathbf{y}_n \in \mathbf{R}^{N\times 1}, n = 1, \ldots, M\}$ into two disjoint subsets specified by the desired output:

$$\mathcal{Y} = \underbrace{\{\mathbf{y}_n | p_\theta(x_n = 1|\mathcal{F}_n) \geq 0\}}_{B_1} \cup \underbrace{\{\mathbf{y}_n | p_\theta(x_n = 1|\mathcal{F}_n) < 0\}}_{B_2}. \tag{11}$$

Now, define $B \equiv B_1 \cup \{-\mathbf{y}_n | \mathbf{y}_n \in B_2\}$ so that the solution set can be defined as $\Theta \equiv \{\theta | p_\theta(x_n = 1|\mathcal{F}_n) > 0, \forall \mathbf{y}_n \in B\}$.

Next, we state the definition of a well-formed cost function in the sense of Wittner and Denker [8]. Consider cost functions of the form

$$J(\theta) = \sum_{i=1}^{n} \nu(\mathbf{y}_i^T \theta). \tag{12}$$

Definition: The cost function $J(\cdot)$ is *well-formed* if $\nu(\cdot)$ is differentiable and satisfies the following:
(i) For all s, $-\nu'(s) \geq 0$ ($\nu(\cdot)$ does not push in the wrong direction).
(ii) There exists some $\epsilon > 0$ such that $-\nu'(s) \geq \epsilon$ for all $s \leq 0$ ($\nu(\cdot)$ keeps pushing if there is a misclassification).
(iii) $\nu(\cdot)$ is bounded below.

Proposition 1: If the cost function is well-formed, then gradient descent is guaranteed to enter Θ, provided Θ is not empty.
Proof: See [8].

Proposition 2: Negative log PL cost function, $-\bar{\mathcal{L}}_n = -\sum_{i=1}^{n}\left[\frac{1+x_i}{2} \ln\left(\frac{1+p_\theta(x_i=1|\mathcal{F}_i)}{2}\right) + \frac{1-x_i}{2} \ln\left(\frac{1-p_\theta(x_i=1|\mathcal{F}_i)}{2}\right)\right]$ is well-formed.
Proof:
With the hyperbolic tangent as the nonlinearity and for the target x, ν becomes

$$\nu(s) = -\frac{1+x}{2} \ln\left(\frac{1+\tanh(s)}{2}\right) - \frac{1-x}{2} \ln\left(\frac{1-\tanh(s)}{2}\right)$$

with $-\nu'(s) = x - \tanh(s)$. In the rephrased version of the binary equalization problem for the development in this section, the target $x = 1$, therefore $-\nu'(s) = 1 - \tanh(s)$ and
(i) $-\nu'(s) = 1 - \tanh(s) \geq 0$,
(ii) $-\nu'(s) = 1 - \tanh(s) \geq 1$ for $s \leq 0$,
(iii) $\nu(s) \geq \ln 2$.
Therefore, gradient descent on the negative log PL cost function is guaranteed to find a solution provided that one exists. As is well known, there is no such guarantee with the MSE cost function when used on MLP's, even on those without any hidden units.

Some further aspects of gradient descent learning on the ARE cost function are considered in [1]. In par-

ticular, the dynamics is studied by considering its parameter updates [1], and it is shown that for LRE, the backpropagated output error is a non-vanishing control signal and hence the algorithm can always recover from convergence at the wrong extreme while the MSE based MLP may not.

The ability of the algorithm to track large variations during training can be quite beneficial for channel equalization. For example, in LEOS (Low Earth Orbit Satellite) communication systems, these abrupt changes occur quite frequently. Due to the Doppler shift, combined with multi-path reflections, the channel characteristics undergo an abrupt change as the channel is switched from one satellite (usually receding with a negative carrier shift) to the next successive satellite (usually approaching with a positive carrier shift). Another typical case is in land mobile communications where multiple cells are transmitting the same information (usually with a small frequency offset) to cover an entire area. In this case the channel variation occurs when the mobile unit switches reception from one antenna to another one having a stronger signal at that particular point. In the next section, we present simulation results to demonstrate these dynamics for LRE and MSE minimizations in practical channel equalization schemes.

5. APPLICATION TO ADAPTIVE CHANNEL EQUALIZATION

In this section, we present application of partial likelihood estimation with neural networks to adaptive channel equalization. We consider a simple binary pulse amplitude modulation (PAM) data transmission system. The supervised adaptive channel equalization problem is posed such that the probability that the transmitted signal x_n takes the value 1 from the binary alphabet is to be determined from a training sequence given the finite past of the received signal: $\mathbf{y}_n = [y_n, y_{n-1}, \cdots, y_{n-N+1}]$. The equalizer structure is shown in Fig. 1.

We study the performance of the LRE algorithm given in (9)-(10) as follows: First, we present test results to demonstrate the capability of the structure to realize complex decision boundaries and of the algorithm to learn parameters to achieve these boundaries. This is done for minimum and non-minimum phase channels at different SNR levels. We then present simulation results to demonstrate the ability of the algorithm to track abrupt changes during training, a property discussed in [1]. The performance of LRE is compared with that of the steepest descent learning (backpropagation) based on the MSE criterion for the same structure, i.e. the perceptron model since this is the structure we have considered in this paper.

For the first simulation study, we consider two simple multipath channels: $H(z) = 1+0.5z^{-1}$ and $H(z) = 0.5 + z^{-1}$, i.e. a minimum phase and a nonminimum phase channel respectively. Figures 2(a) and 2(b) show the decision regions for the first, and figures 3(a) and 3(b) show the regions for the second channel, for approximately 21 dB and 11 dB SNRs respectively. The MLP structure used in these figures is 2-7-1. As observed in both cases, LRE successfully learns the coefficients for achieving the given partition. These results compare favorably with those presented in [5] for the MLP equalizer based on the MSE criterion.

Next, we consider a nonlinear channel example and compare the learning characteristics of LRE with MSE based backpropagation. We model the nonlinear channel as a multipath channel ($H(z) = 1+0.5z^{-6}+0.25z^{-16}$) followed by a nonlinearity $0.5(\cdot)^3$, and the PAM communication system has 8 bits per sample with Nyquist pulse shaping. Note that since 8 bit pulse shaping is used, the multipath structure corresponds to fractional previous symbol interference, and full interference of second previous symbol. We implement the LRE algorithm for binary alphabet given in (9) and (10), and the gradient descent minimization of the MSE on the same MLP structure for equalization of the given channel. Both algorithms have a 3-8-1 MLP structure.

To show the recovery property of LRE discussed in the previous section, we introduce an abrupt change (an exact sign change) in the channel characteristics after 150 iterations, effectively causing the current parameter estimates to be at the wrong extreme. In Fig. 4(a), we show the transient characteristics of both algorithms with the abrupt change at 150 iterations at a signal to noise ratio (SNR) of 19 dB. As observed in the figure, LRE can recover from convergence at the wrong extreme very effectively. Starting from the very first iteration after the change it can follow the changes by adapting both its hidden and output layer weights in a few iterations. As we can observe in Fig. 4(a), MSE based MLP produces many wrong decisions before it can adapt to this new operating condition. Note that both algorithms have not fully converged at 150 iterations, and if the sudden change causing misclassifications occurs later MSE based MLP might not be able to recover. This is shown in Fig. 4(b), by introducing the sudden change at iteration 1000. Again LRE can very rapidly adapt to the new operating condition, rapidly recovering from convergence at the wrong extreme.

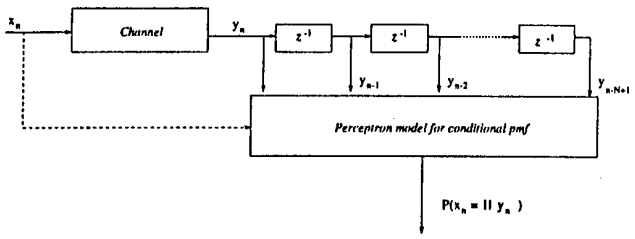
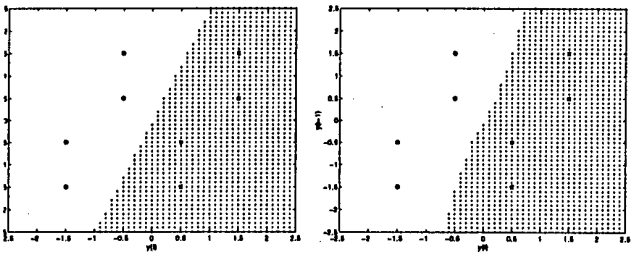
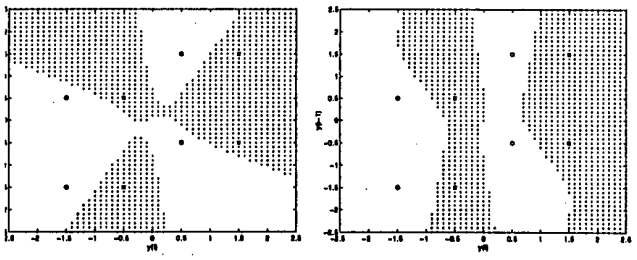


Figure 1: Conditional distribution learning for the binary communications channel



(a) (b)

Figure 2: Decision regions formed by the LRE equalizer for $H(z) = 1+0.5z^{-1}$ for (a) 21 dB and (b) 11 dB SNR "*" represents $x_n = 1$ and "o" $x_n = -1$).



(a) (b)

Figure 3: Decision regions formed by the LRE equalizer for $H(z) = 0.5+z^{-1}$ for (a) 21 dB and (b) 11 dB SNR "*" represents $x_n = 1$ and "o" $x_n = -1$).

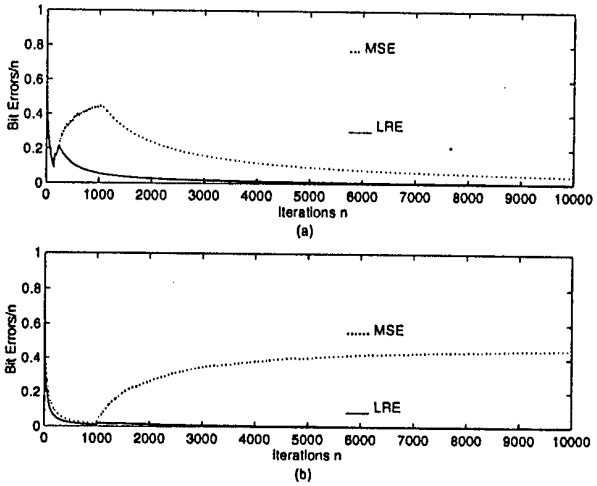


Figure 4: Recovery characteristics for MSE and LRE MLP equalizers with an abrupt change at (a) 150 (b) 1000 iterations (SNR =19 dB)

6. REFERENCES

[1] T. Adalı, M. K. Sönmez, and K. Patel, "On the dynamics of the LRE algorithm," in *Proc. ICASSP*, (Detroit, MI), May 1995, pp. 929-932.

[2] T. Adalı, X. Liu, and M. K. Sönmez, "Conditional distribution learning by neural networks and its application to channel equalization," to appear *IEEE Trans. Signal Processing.*

[3] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedure," *Statistics and decisions, Supplementary issue, No. 1,* (E. Dedewicz *et. al.*, eds.), pp. 205-237, Munich, Oldenburg Verlag, 1984.

[4] D.R. Cox, "Partial likelihood," *Biometrika*, vol. 62, pp. 69-72, 1975.

[5] G. J. Gibson, S. Siu, and C. F. N. Cowan, "The application of nonlinear structures to the reconstruction of binary signals," in *IEEE Trans. Signal Processing*, vol. 39, no. 8, pp. 1877-1884, Aug. 1991.

[6] L. Kullback, and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics 22*, pp. 79-86, 1951.

[7] E. Slud and B. Kedem, "Partial likelihood analysis of logistic regression and autoregression," Statistica Sinica, vol. 4, no. 1, pp. 89-106, Jan. 1994.

[8] B. S. Wittner and J. S. Denker, "Strategies for teaching layered networks classification tasks," *Neural Info. Proc. Systems* (Denver, CO), 1988, pp. 850-859.

[9] H. White, *"Estimation, inference, and specification analysis,"* New York, NY, USA : Cambridge University Press, 1994.

[10] W. H. Wong,"Theory of partial likelihood," *Ann. Statist.*, 14, pp. 88-123, 1986.

Image Processing

USE OF NON-LINEAR PRINCIPAL COMPONENT ANALYSIS AND VECTOR QUANTIZATION FOR IMAGE CODING

Dimitrios Tzovaras and Michael G. Strintzis

Information Processing Laboratory
Electrical and Computer Engineering Department
Aristotle University of Thessaloniki, Thessaloniki 540 06, Greece
Tel. : +3031.996359, Fax : +3031.996398, e-mail : tzovaras@dion.ee.auth.gr

ABSTRACT

In the present paper the non-linear principal component analysis method is combined with vector quantization for the coding of images. The proposed coder is fully implemented using neural networks (NN). The NLPCA is realized using the backpropagation NN, while vector quantization is performed using the LVQ NN. The effects of quantization in the quality of the reconstructed image are then compensated by using codebook vector optimization. Experimental results are presented for the coding of a sequence of images.

I. INTRODUCTION

The use of the Karhunen-Loeve Transform (KLT) for image coding is well known to be an optimal scheme for data compression based on the exploitation of correlation between neighboring pixels or groups of pixels [1]. Its superior performance has made it a benchmark against which other methods such as the Discrete Cosine Transform (DCT) and the Walsh Transform are compared. Despite its optimality properties however, it has not found widespread application because of the difficulties associated with the needed computation of the eigenvectors of the image covariance matrix. The use of DCT as an approximation to the KLT is a well established, easily implementable practical alternative. Other well researched alternatives include the use of neural networks to implement the image Principal Component Analysis (PCA) thus computing its KLT [2]; and the use of neural networks to realize a direct coder/ decoder autoassociative mechanism based on examples [7].

It can be shown ([1], see also [8]) that under certain conditions, notably Gaussian statistics, the optimum data compression method is linear and is therefore under the minimum mean-square criterion, identical to the PCA (KLT) method. However, this is not the case when the data statistics are, more realistically, a mixture of Gaussian distributions. A nonlinear Principal Component Analysis (NLPCA) has been functionally defined in [6] by a class of autoassociative Neural Networks. The resulting data compression scheme has been shown to outperform linear PCA and to be relatively easy to implement.

The present paper utilizes the NLPCA of [6] to effect image coding, aiming at higher efficiency coding than is possible with the linear PCA method [2]. It also combines this with a proposed data compressor based on vector quantization. According to this technique a codebook is created based on the frequency of occurrence of vectors in a series of images. This vector quantization is realized by the Counterpropagation (Learning Vector Quantizer (LVQ)) Neural Network. Furthermore, an analysis is presented of the error due to coefficient quantization in the bottleneck layer. Finally, a technique is proposed for post-processing of the LVQ codebook vectors in order to minimize the quantization effects to the output of the decoding network.

II. NON LINEAR PRINCIPAL COMPONENT ANALYSIS (NLPCA)

Let $\mathbf{Y}$ represent a $n \times m$ data matrix (n is the number of the observed data vectors and m is the dimension of those vectors). The goal of NLPCA is the replacement of each data vector (row of $\mathbf{Y}$) $\underline{Y}$ with the corresponding row $\underline{T}$, of a matrix $\mathbf{T}$, with dimension $n \times f$, $f < m$:

$$\underline{T} = \underline{G}(\underline{Y}) \tag{1}$$

where $\underline{G}$ is a non-linear vector function composed of f non-linear functions :

$$\underline{G} = \{G_1, G_2, \ldots, G_f\}; \quad T_i = G_i(\underline{Y}) \tag{2}$$

This work was supported in part by the ACTS PANORAMA project 092 and the Greek Secretariat for Research and Technology projects NIKA and IHIS.

and by analogy with the linear PCA terminology, T_1 is referred to as the prime nonlinear factor of $\underline{T}$, and T_i as the i-th nonlinear factor. Reconstruction of the data vector is done by using a second nonlinear function $\underline{H}$:

$$\underline{H} = \{H_1, H_2, \ldots, H_m\} \quad (3)$$

and the reconstructed data vector $\underline{Y'}$ has components

$$Y'_j = H_j(\underline{T}), \quad j = 1, \ldots, m. \quad (4)$$

A measure of the loss of information is the error $\underline{E} = \underline{Y} - \underline{Y'}$. The functions $\underline{G}$ and $\underline{H}$ are selected so as to minimize the Euclidean norm of $\underline{E}$.

III. NEURAL NETWORKS REALIZATION OF NLPCA

The neural network implementing the NLPCA in [6] consists of joined realizations of the coding and the decoding functions. The realization of the coding function $\underline{G}$ has the rows of the data matrix $\mathbf{Y}$ as inputs and thus has m input neurons (see Figure 1). The hidden "mapping" layer consists of M_1 neurons, where M1 must be higher than f. The output of the neural network is the corresponding row of matrix $\mathbf{T}$ and is composed of f neurons. It represents the projection of the data vector on the f-dimensional space. The output neurons can be either linear or sigmoidal.

In the implementation of the decoding function $\underline{H}$, input is one row of matrix $\mathbf{T}$ (see Figure 2). Thus the input layer consists of f linear neurons. The hidden "demapping" layer consists of K_2 sigmoidal neurons, where $K_2 > f$. The output of this network is the reconstructed data vector where each one of the m neurons represents one component of the data vector.

The combined network contains three hidden layers, the mapping layer involved in modeling in $\underline{G}$, the middle layer whose outputs represent the features $\underline{T}$, and the demapping layer involved in modeling $\underline{H}$ (see Figure 3). The second hidden layer of the combined network is the "bottleneck" layer and its size determines the data compression to be achieved. The input and output layers of the combined network represent $\underline{Y}$ and $\underline{Y'}$, respectively. $\underline{Y}$ is both the input to $\underline{G}$ and the desired output from $\underline{H}$; thus, the combined network must be trained to produce the identity mapping, $Y - Y$. Supervised training can thus be applied to the combined network. Training to learn the identity mapping has been called self-supervised backpropagation or autoassociation [7].

IV. VECTOR QUANTIZATION

Additional data compression was achieved by using the Hecht-Nielsen Counter Propagation Neural Network (CPN) on the coder output $\underline{T}$ to produce its quantized version $\underline{T}_q$. The CPN [9] is a feedforward, unsupervised learning neural network formed as a combination of Kohonen's Learning Vector Quantizer with a Grossberg outstar. In our case it serves for the further quantization of the coder output to indexed clusters; the CPN output is the index number of the cluster. Thus, only the index (an integer) need be transmitted, effecting further data compression of the (real number) coder outputs. Upon reception of the index, the decoder reconstructs the image from a codebook formed on the basis of the frequency of occurrence of vectors in a series of images.

The LVQ network guarantees the minimization of the quantization error norm $||\underline{T} - \underline{T}_q||$. However, for data compression purposes the minimization of the input-output error $\underline{Y} - \underline{Y'}_q$, is needed ($\underline{Y'}_q = \underline{H}(\underline{T}_q)$). Thus in our approach, the statistics of the vector quantization error $\underline{T} - \underline{T}_q$ are then used to update the weights of the NLPCA decoder in such a way that the input-output reconstruction error is minimized.

V. CODEBOOK VECTOR OPTIMIZATION

A basic drawback of the proposed coding approach, is that the NLPCA networks and the LVQ networks are trained independently. This is due to the fact that if the two networks were trained simultaneously (i.e. at each iteration of the training of the NLPCA network the output of the third level was led as an input to the LVQ training network, and the output of the LVQ network was led as an input to the fourth level of the NLPCA network, as shown in Figure 3) the stability of the steepest descent training procedure would be no longer guaranteed.

Due to the non-linearity of the proposed coder a small error (due to quantization) may lead to large errors in the decoding phase of the network. Motivated from this fact we propose a post processing phase after the training of the NLPCA and LVQ networks, that modifies appropriately the codebook vectors in order for the input-output mean square error to be minimized. According to the proposed technique, if b_i is the output of the ith neuron of the first layer of the decoding network the following error must be minimized

$$\min \sum_{j=1}^{m} (y_j - y'_j)^2 \quad (5)$$

where y is the input vector and y' is the output vector.

The equations describing the output vector y' in terms of the quantized coefficient vectors (output of the LVQ network) are :

$$b_i = f(\sum_{h=1}^{f}(v_{hi}T_h) + \Theta_i) \quad (6)$$

$$y'_j = f(\sum_{j=1}^{M_2}(w_{ij}b_i) + \Gamma_i \quad (7)$$

where

$$f(x) = \frac{1}{1+e^{-x}}$$

where v are the weights connecting the first with the second layer of the decoding network and w are the weights connecting the second with the third layer of the decoding network. Minimization of (5) in terms of the T_h, $h = 1, \ldots, f$ implies

$$\frac{\partial \sum_{j=1}^{m}(y_j - y'_j)^2}{\partial T_k} = 0 \text{ for } k = 1, \ldots, f$$

thus,

$$\sum_{j=1}^{m}(y_j - y'_j)\frac{\partial y'_j}{\partial T_k} = 0$$

Note that

$$\frac{\partial y'_j}{\partial T_k} = y'_j(1-y'_j)\sum_{i=1}^{M_2} w_{ij}b_i(1-b_i)v_{kj}$$

Thus, the following system of k non-linear equations must be solved in terms of T_k, for $k = 1, \ldots, f$

$$\sum_{j=1}^{m}(y_j - y'_j)y'_j(1-y'_j)\sum_{i=1}^{M_2} w_{ij}b_i(1-b_i)v_{kj} = 0$$

for $k = 1, \ldots, f$. Note that the T_k's are implicitly contained in the above equations, as seen by equations (6) and (7) for b_i and y'_j.

The steepest descent method was used for the solution of the above system of non-linear equations. The LVQ codebook vectors were used as initial estimates for a_k. The post-processed codebook vectors are then used for the coding of the coefficients of the coding network of NLPCA.

VI. EXPERIMENTAL RESULTS

The proposed still image coding method was applied for the coding of the 9 first frames of the image sequence "Trevor White" of size 256 × 256. The images were divided into blocks of dimension 4 × 4, creating vectors of dimension 16. The neural network implementing the NLPCA was chosen with $m = 16$, i.e. 16 neurons in the first and last level, $M_1 = M_2 = 35$ neurons in the second and fourth layer and $f = 8$ neurons in the middle level.

We have examined two approaches for the training of the NLPCA network : a) only the first image was used as a training set, b) the first, the fifth and the ninth image were used as a training set. The LVQ network was trained with the coefficients corresponding to the training set of the NLPCA network. Codebooks of 1024, 512 and 256 were tested corresponding to bit rates of respectively.

The PSNR of the reconstructed image is shown in Tables 1 and 2. The original and decoded frames 1, 5, and 7 of the "Trevor White" sequence are shown in Figures 4-9. The results show that the method works satisfactory when the NLPCA network is trained with the set of the three images. When only the first image (I-frame) is used for the training of the NLPCA network only the first 4 images are coded efficiently, since the motion is very small and the correlation with the first image is high.

The use of the method proposed in Section V for quantization error compensation during codebook initialization has seen to improve considerably the results (more than 1 dB improvement compared to Tables 1 and 2), especially for the codebooks of 256 and 512 vectors, where the quantization effects are not negligible.

The NLPCA method was applied to the compression of "Lenna" both with and without additive noise. Blocks of dimension 4x4 were used, with various sizes of the mapping and the bottleneck layer and with linear bottleneck layer units. The results were compared to the results of PCA coding (KLT). It was seen that with Gaussian noise, the results were comparable with those of the linear PCA scheme; in cases of non-Gaussian noise addition, however, the NLPCA method performed considerably better than the PCA method.

VII. REFERENCES

[1] A. N. Netravali and B. G. Haskell, *Digital Pictures. Representation and Compression*, Plenum Press, New York and London, 1988.

[2] T. D. Sanger, "Optimal Unsupervised Learning in a Single Layer Linear Feedforward Neural Network", *Neural Networks*, Vol. 2, pp. 459-463, 1989.

[3] E. Oja, "A Simplified Neural Model as a Principal Component Analyser", *J. Math. Biology*, Vol. 15, pp. 267-273, 1982.

[4] P. Baldi and K. Hornik, "Neural Networks and Principal Component Analysis : Learning from Examples without Local Minima," *Neural Networks*, Vol. 2, pp. 53-62, 1989.

Frame	0.625	0.5625	0.5
	bits/pixel	*bits/pixel*	*bits/pixel*
1	35.00	31.10	29.10
2	31.60	29.91	28.50
3	31.21	29.77	28.35
4	30.03	29.01	27.95
5	29.68	28.64	27.77
6	29.30	28.58	27.70
7	29.22	28.30	27.61
8	28.85	28.11	27.40
9	28.56	28.01	27.30

Table 1: PSNR for the coding of the first 9 frames of the "Trevor White" sequence using the NLPCA network trained using only the first frame.

Frame	0.625	0.5625	0.5
	bits/pixel	*bits/pixel*	*bits/pixel*
1	32.25	29.95	28.85
2	30.80	29.17	28.50
3	31.02	29.21	28.48
4	30.68	29.01	28.25
5	30.81	29.34	28.70
6	30.77	29.22	28.82
7	31.25	29.67	29.05
8	31.26	29.70	29.13
9	32.31	30.20	29.35

Table 2: PSNR for the coding of the first 9 frames of the "Trevor White" sequence using the NLPCA network trained using the first, fifth and ninth frames.

[5] F. M. Silva and L. B. Almeida, "A Distributed Solution for Data Orthonormalization", *Proc. ICANN*, Helsinki, 1991.

[6] M. A. Kramer, "Nonlinear Principal Component Analysis Using Autoassociative Neural Networks", *AIChE Journal*, vol. 37, num. 2, pp. 233-243, 1991.

[7] G. W. Cottrell, P. W. Munro and D. Zipser, "Image Compression by Error Backpropagation : A Demonstration of Extensional Programming", *Advances in Cognitive Science*, Vol.2, Norwood N.J.

[8] H. Bourlard and Y. Kamp, "Auto-Association by Multilayer Perceptrons and Singular Value Decomposition", *Biol. Cybernetics*, Vol. 59, 1988, pp. 291-294.

[9] P. K. Simpson, *Artificial Neural Systems*, Pergamon Press, New York 1990.

[10] G. Dundar and K. Rose, "The Effects of Quantization on Multilayer Neural Networks", *IEEE Trans. om Neural Networks*, vol. 6, no. 6, Nov. 1995.

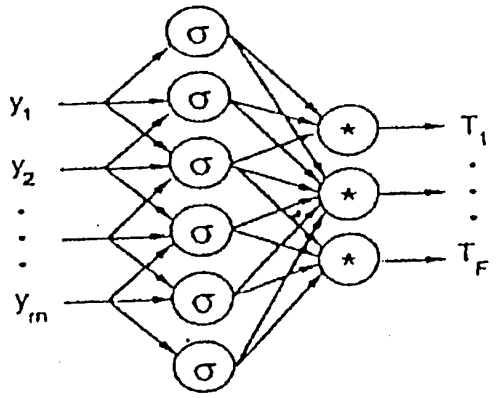


Figure 1: The encoding phase of the NLPCA network.

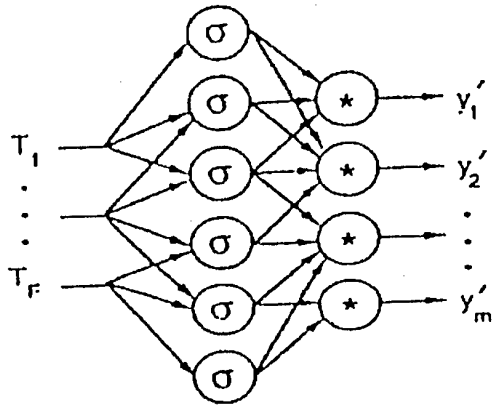


Figure 2: The decoding phase of the NLPCA network.

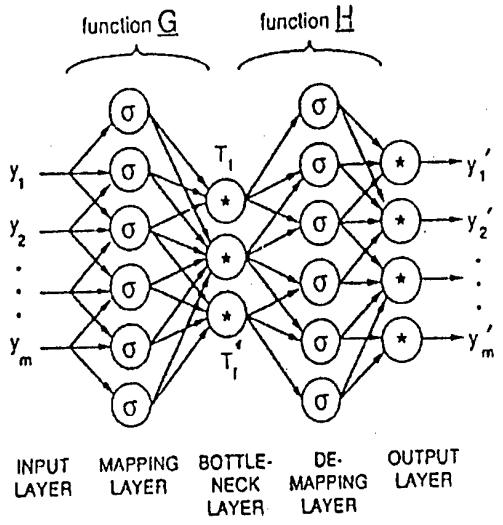


Figure 3: The encoding and decoding phases of the NLPCA network.

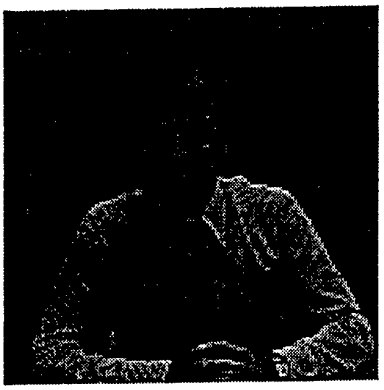


Figure 4: Original frame 1 of the "Trevor White" sequence.

Figure 5: Reconstructed frame 1 of the "Trevor White" sequence using NLPCA trained only using this frame.

Figure 6: Original frame 5 of the "Trevor White" sequence.

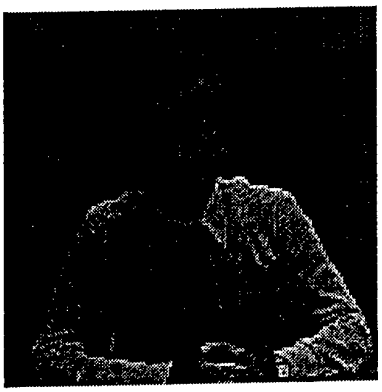


Figure 7: Reconstructed frame 5 of the "Trevor White" sequence coded at 0.625 *bits/pixel* using NLPCA trained only using the first frame.

Figure 8: Original frame 7 of the "Trevor White" sequence.

Figure 9: Reconstructed frame 7 of the "Trevor White" sequence coded at 0.625 *bits/pixel* using NLPCA trained with the first, fipth and ninth images.

CODING OF MULTICHANNEL IMAGES USING OPTIMAL VECTOR HIERARCHICAL DECOMPOSITION

Dimitrios Tzovaras and Michael G. Strintzis

Information Processing Laboratory
Electrical and Computer Engineering Department
Aristotle University of Thessaloniki, Thessaloniki 540 06, Greece
Tel. : +3031.996359, Fax : +3031.996398, e-mail : tzovaras@dion.ee.auth.gr

ABSTRACT

In the present paper we determine two families of analysis and synthesis vector filters which achieve optimal construction of multiresolution vector sequences by minimizing the variance of the error signals between successive pyramid levels. A measure of the entropy reduction achieved by the pyramid is in this way maximized. The effect of this is to ensure that the lower-resolution image produced by the primary subband bears maximum resemblance to the input image. Furthermore, it is assumed that additive transmission noise corrupts the downsampled signal prior to the synthesis stage. It is seen that under noiseless or lossless transmission conditions, the two above families of optimal analysis and synthesis filters coincide. The results are evaluated experimentally for the vector coding of color images.

I. INTRODUCTION

Subband analysis/synthesis techniques have been extensively studied for image and video coding applications [1]. According to the subband coding technique the image is decomposed into several sub-images in terms of different frequency bands by a filter bank and to code the sub-images instead of the original image.

Pyramidal image coding has also been studied [2, 3] and optimal construction of the pyramid sequence was sought by minimizing for each level of the pyramid the variance of the error image. In this way, a measure of the entropy reduction achieved by the pyramid is maximized. If the pyramid is to be used for the scalable or progressive coding of the sequence, this construction also ensures the production of a same-size copy of the signal or image which at a lower resolution bears as much resemblance to the original as possible. In a typical scalable coding application this copy may be transmitted via a slower communication channel, while the original is perfectly reconstructed from the entire pyramid. In an alternative scheme, a perfect reconstruction filter bank may be used for the transmission of the full-resolution signal or image, of which one or more bands are retained for the construction of one or more copies of the original at lower resolutions. Again, the filters are chosen so as to ensure that at the lower resolutions, the copies bear as much resemblance the original image as is possible. This is achieved by minimization of the error occuring when only one band is retained of the perfect reconstruction filter bank.

Along with scalar processing, vector processing has attracted particular interest in the signal and image processing community recently [4]. Vector transform coding techniques have recently been used for image coding applications [5] to remove the inter-vector correlation.

In the present paper the results of [2, 3, 6] are first generalized and the problem of the optimal design of a vector pyramid or a vector subband coding scheme is addressed. Furthermore, the results are generalized to the case where transmission noise corrupts the downsampled signal prior to the synthesis stage.

In the examined scheme the analysis or the synthesis part is considered fixed (i.e. the analysis or the synthesis filters are fixed) and the statistics of the quantization part involved for transform coefficient coding are considered known. The problem then is to define the optimal synthesis (analysis) vector filters that minimize the distortion due to the quantization of the subband or pyramid vector transform coefficients. Thus specific knowledge about the power spectrum of the original signal and the quantization noise, can be incorporated to design optimally the vector filter bank so as to minimize the quantization distortion.

This work was supported in part by the ACTS PANORAMA project 092 and the Greek Secretariat for Research and Technology projects NIKA and IHIS.

II. ORTHOGONAL TWO-CHANNEL VECTOR FILTER BANKS

Recently in [7] the vector filter banks were introduced and the perfect reconstruction orthogonal analysis / synthesis vector filters $\mathbf{H}(\omega)$ and $\mathbf{G}(\omega)$ were defined that satisfy the following properties

$$\mathbf{H}(0) = I_N, \text{ or } \sum_k H_k = I_N$$

where

$$\mathbf{H}(z) = \sum_k H_k z^{-1}$$

and

$$\mathbf{H}(z)\mathbf{H}^T(z) + \mathbf{H}(-z)\mathbf{H}^T(-z) = I_N$$
$$\mathbf{G}(z)\mathbf{H}^T(z) + \mathbf{G}(-z)\mathbf{H}^T(-z) = I_N$$

and

$$\mathbf{G}(z)\mathbf{G}^T(z) + \mathbf{G}(-z)\mathbf{G}^T(-z) = I_N$$

and also that the H_k is symmetric. In the same work, it was proven that an $M \times M$ matrix $\mathbf{F}(z)$ is paraunitary and FIR if and only if

$$\mathbf{F}(z) = e^{im_0\omega}\mathbf{U}_\rho(z)\ldots\mathbf{U}_1(z)\mathbf{F}$$

where m_0 is an integer, ρ is a nonnegative integer, $\mathbf{F}$ is an $M \times M$ constant unitary matrix and

$$\mathbf{U}_l(\omega) = I_M + (e^{i\omega} - 1)\mathbf{v}_l\mathbf{v}_l^T$$

where $\mathbf{v}_l$ is a unit-norm constant $M \times 1$ vector, for $l = 1, \ldots, \rho$.

In the present work we deal with deriving the optimal biorthogonal perfect reconstruction vector filters.

III. OPTIMAL VECTOR PYRAMIDAL AND SUBBAND DECOMPOSITIONS

A multiresolution data representation consists of a sequence of linear transformations of the data with successively reduced resolution. If the vector sequence $\mathbf{x}[m]$ represents the original data, the construction of the multiresolution sequence begins with the computation of the predicted value $\mathbf{u}[m]$ of each $\mathbf{x}[m]$ as a local weighted average :

$$\mathbf{u}[m] = \sum_{i=-N}^{N} \mathbf{h}[i]\mathbf{x}[m-i] = \mathbf{h} \star \mathbf{x} \qquad (1)$$

where the asterisk $\star$ denotes convolution. A well-known specific multiresolution representation is based on the construction of the "Gaussian Pyramid" in which (1)

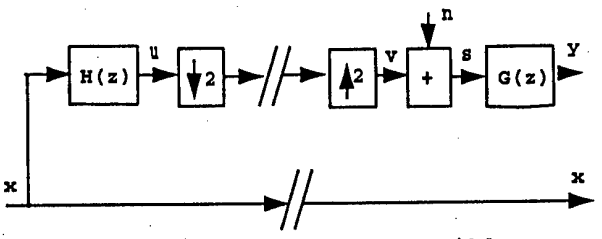


Figure 2 : Hierarchical Vector Pyramidal encoder.

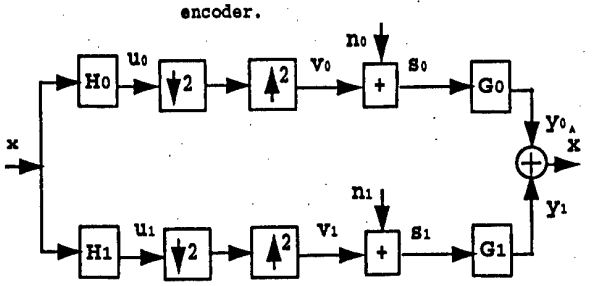


Figure 2 : Hierarchical Vector Subband encoder.

is followed by a decimating filter. Interpolation is then used to revert to the original image size:

$$\mathbf{v}[m] = \begin{cases} \mathbf{u}[2m_1] & \text{if } m = 2m_1 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

If the pyramid is used in signal or image transmission, the downsampled signal is subject to corruption by noise. The quantizers and transmission noise sources are physically located before the upsamplers. For the sake of notational simplicity, however, we shall equivalently place them after the upsamplers, assuming simply that the zeros interpolated by the upsamplers are always quantized to zero. If additive noise $\mathbf{n}[m]$ is assumed,

$$\mathbf{y}[m] = \sum_{i=-N'}^{N'} \mathbf{g}[i](\mathbf{v}[m-i] + \mathbf{n}[m-i]) \qquad (3)$$

The error image is:

$$\mathbf{e}[m] = \mathbf{x}[m] - \mathbf{y}[m] \qquad (4)$$

and the total error variance :

$$E = E\{\mathbf{e}^T[m]\mathbf{e}[m]\} \qquad (5)$$

The process is repeated for the reduction in resolution of the sequence $\mathbf{x}'[m] = \mathbf{u}[2m]$. The corresponding sequence of error images is known as the "Laplacian" pyramid.

An optimal construction of the pyramid sequence may be sought by minimizing for each level of the pyramid the variance of the error image (5). In this way, a

measure of the entropy reduction achieved by the pyramid is maximized.

The output of the prefilter followed by decimation and interpolation by zeros is given by

$$\mathbf{v}[m] = w[m] \sum_i \mathbf{h}[i]\mathbf{x}[m-i] = w[m]\mathbf{u}[m] \quad (6)$$

where

$$w[m] = \frac{(1+(-1)^m)}{2} \quad (7)$$

The sequence u[m] is not wide-sense stationary; however the time averages

$$\mathbf{R_{ux}}[p] = \lim \frac{1}{2K+1} \sum_{m=-K}^{K} \mathbf{u}[m]\mathbf{x}^T[m-p],$$

$$\mathbf{R_v}[p] = \lim \frac{1}{2K+1} \sum_{m=-K}^{K} \mathbf{v}[m]\mathbf{v}^T[m-p] \quad (8)$$

are seen to exist, under nonrestrictive conditions on $\mathbf{x}[.]$. With the above definition, it can be proven that,

$$\mathbf{R_{ux}}[m] = \frac{1}{2}\sum_i \mathbf{h}[i]\mathbf{R_x}[m-i] = \frac{1}{2}\mathbf{h} * \mathbf{R_x}$$

Also,

$$\mathbf{R_v}[p] = \frac{1}{2}c[p](\mathbf{h}[p] * \mathbf{R_x}[p] * \mathbf{h}[-p]) \quad (9)$$

The corresponding Z-transforms are therefore related by

$$\mathbf{\Phi_{vx}}(z) = \frac{1}{2}\mathbf{H}(z)\mathbf{\Phi_x}(z) \quad (10)$$

$$\mathbf{\Phi_v}(z) = \frac{1}{4}(\mathbf{A}(z) + \mathbf{A}(-z)) = \frac{1}{2}\mathbf{P}(z) \quad (11)$$

where

$$\mathbf{A}(z) = \mathbf{H}(z)\mathbf{\Phi_x}(z)\mathbf{H}^T(z^{-1}) \,. \quad (12)$$

Then using the above arguments, it is easily seen that the sequences $\mathbf{s}[m]$, possess auto- and cross-correlations and spectra defined as in (8). Note also that since $\mathbf{s}[m] = 0$ if $m =$ odd, their spectra satisfy

$$\mathbf{R_s}[2n+1] = 0, \qquad \mathbf{\Phi_s}(z) = \mathbf{\Phi_s}(-z) \quad (13)$$

Likewise, the output $\mathbf{y}$ of the interpolating filter is seen as before to possess cross and autocorrelation functions defined as in (8). Their Z-transforms are related by

$$\mathbf{\Phi_{yx}}(z) = \mathbf{G}(z)\mathbf{\Phi_{sx}}(z) \quad (14)$$

$$\mathbf{\Phi_y}(z) = \mathbf{G}(z)\mathbf{\Phi_s}(z)\mathbf{G}^T(z^{-1}) \quad (15)$$

From (5), the error variance is found by

$$2\pi j E = tr[E\{\mathbf{e}[m]\mathbf{e}^T[m]\}] = tr[\oint \mathbf{\Phi_e}(z)z^{-1}dz] \quad (16)$$

where $tr[\mathbf{F}]$ is the trace of the matrix $\mathbf{F}$ and $\mathbf{\Phi_e}(z)$ is the power spectrum of the error $\mathbf{e}[m]$. Clearly $\mathbf{\Phi_e}(z) = \mathbf{\Phi_x}(z) - \mathbf{\Phi_{xv}}(z) - \mathbf{\Phi_{yx}}(z) + \mathbf{\Phi_y}(z)$ hence from (14-16) :

$$2\pi j E = tr[\oint (\mathbf{\Phi_x}(z) - 2\mathbf{G}(z)\mathbf{\Phi_{sx}}(z))\, z^{-1}dz] +$$

$$tr[\oint \mathbf{G}(z)\mathbf{\Phi_s}(z)\mathbf{G}^T(z^{-1})z^{-1}dz] \quad (17)$$

Thus, the design of either the pyramidal or the subband decomposition scheme should aim at the minimization of the error variance (17).

IV. OPTIMAL FIR AND IIR VECTOR FILTERS

With arbitrary given $\mathbf{h}[i]$, the optimum FIR filter $\mathbf{g}[i]$ in (3) will minimize the error variance (5) if the well known orthogonality condition holds :

$$E\{\left(\mathbf{x}[m] - \sum_{i=-N}^{N} \mathbf{g}[i]\mathbf{s}[m-i]\right)\mathbf{s}^T[m-l]\} = 0,$$

for $l = -N, \ldots, N$. This implies

$$\mathbf{R_{xs}}[l] = \sum_{i=0}^{N} \mathbf{g}[i]\mathbf{R_s}[l-i], \quad , l = -N, \ldots, N$$

Given (11) this may be separated into two sets of equations for the identification respectively of the even- and odd- indexed coefficient matrices $\mathbf{g}[i]$:

$$\mathbf{R_{xs}}[2l_1] = \sum_{i1} \mathbf{g}[2i_1]\mathbf{R_s}[2l_1 - 2i_1]$$

$$\mathbf{R_{xs}}[2l_2+1] = \sum_{i2} \mathbf{g}[2i_2+1]\mathbf{R_s}[2l_2 - 2i_2]$$

which define fully $\mathbf{g}[i], i = -N, \ldots, N$. The optimal IIR filters are found by direct minimization of (17). We shall consider the noiseless case first.

IV.1. NOISELESS CASE

In this case $\mathbf{s} = \mathbf{v}$, and hence

$$\mathbf{\Phi_{sx}}(z) = \mathbf{\Phi_{vx}}(z) = \frac{1}{2}\mathbf{H}(z)\mathbf{\Phi_x}(z) \quad (18)$$

$$\mathbf{\Phi_s}(z) = \mathbf{\Phi_v}(z) = \frac{1}{2}\mathbf{P}(z) \quad (19)$$

with $\mathbf{P}$ given by (19). The error variance is found to be

$$2\pi jE = tr \oint \mathbf{\Phi_X}(z) - \mathbf{H}(z)\mathbf{\Phi_X}(z)\mathbf{G}(z) + tr \oint \frac{1}{2}\mathbf{H}(z)\mathbf{\Phi_X}(z)\mathbf{H}^T(z^{-1})\mathbf{Q}(z)z^{-1}dz \quad (20)$$

where

$$\mathbf{Q}(z) = \frac{1}{2}\left(\mathbf{G}^T(z^{-1})\mathbf{G}(z) + \mathbf{G}^T(-z^{-1})\mathbf{G}(-z)\right) . \quad (21)$$

The optimal pyramidal and subband decompositions will be obtained by minimization of the above expression (20) for the error variance. Assuming first the analysis filter $\mathbf{H}(z)$ fixed and given, the optimum corresponding synthesis filter minimizing (20) is given by

$$\mathbf{G}(z) = \mathbf{\Phi_X}(z)\mathbf{H}^T(z^{-1})\mathbf{P}^{-1}(z) . \quad (22)$$

Conversely, if the synthesis filter $\mathbf{G}(z)$ is fixed, the optimum analysis filter can be found. This is achieved when

$$\mathbf{H}(z) = \mathbf{Q}^{-1}(z)\mathbf{G}^T(z^{-1}) \quad (23)$$

where $\mathbf{Q}(z)$ is given by (21). The globally optimum filter pair $(\mathbf{H}(z), \mathbf{G}(z))$ will be found by either (22) or (23) and the minimization of the resulting expression in (20). The minima found either way are easily seen to be identical.

To develop the optimum filter bank for vector signal coding, an extra constraint is that the filters $\mathbf{H}(z)$ and $\mathbf{G}(z)$ selected be such that a perfect reconstruction filter bank can be built with those filters respectively its analysis and synthesis filters for its primary band. The analysis and synthesis filters in the remaining bands are chosen to satisfy the perfect reconstruction principle. In the noiseless case, the filters defined by (14), (15) always satisfy the perfect reconstruction condition. Thus in this case, the low-pass band of the filter has analysis and synthesis filters identical to those in the optimal pyramidal analysis.

IV.2. NOISY CASE

With considerable insight into the structure of the optimal pyramids and filter banks gained from the consideration of the noiseless case, the noisy case may now be considered. Again, if the analysis filter is fixed, the optimum synthesis filter in the pyramidal configuration will be found by minimizing (17). It can be shown that this will be given by

$$\mathbf{G}(z) = \mathbf{\Phi}^T_{\mathbf{SX}}(z^{-1})\mathbf{\Phi}^{-1}_{\mathbf{S}}(z) . \quad (24)$$

This is a completely general expression for the optimal synthesis filter, which can be further analyzed under some additional simplifying assumptions. For example, the additive noise may be assumed to be uncorrelated with the input :

$$\mathbf{\Phi_{VN}}(z) = 0 \quad (25)$$

This assumption is reasonable in the instance of transmission noise and is justified for a large class of practical quantizers which includes fine and dithered quantizers [1]. In this case

$$\mathbf{\Phi_{SX}}(z) = \mathbf{\Phi_{VX}}(z) = \frac{1}{2}\mathbf{H}(z)\mathbf{\Phi_X}(z)$$

$$\mathbf{\Phi_S}(z) = \mathbf{\Phi_V}(z) + \mathbf{\Phi_R}(z) = \frac{1}{2}\mathbf{P}(z) + \mathbf{\Phi_R}(z) \quad (26)$$

where $\mathbf{\Phi_R}(z)$ is the noise power spectral density. Note also that :

$$\mathbf{\Phi_R}(z) = \mathbf{\Phi_R}(-z) \quad (27)$$

From (24), the optimal $\mathbf{G}(z)$ is

$$\mathbf{G}(z) = \mathbf{\Phi}^T_{\mathbf{X}}(z^{-1})\mathbf{H}^T(z^{-1})[\mathbf{P}(z) + 2\mathbf{\Phi_R}(z)]^{-1} \quad (28)$$

It can be seen, that in this case, unlike the noiseless case, the optimal pyramidal synthesis filters fail to satisfy the necessary condition for perfect reconstruction and therefore do not offer a solution to the problem of optimal filter bank construction whether in the general form (24) or in the special case (28). However, in the latter case, the analysis filters may be chosen so as to form a perfect reconstruction bank with the optimal synthesis filters.

V. EXPERIMENTAL RESULTS

The proposed vector subband coding method was tested for the coding of multichannel images. The results are evaluated in the coding of the color RGB image "Peppers" of size 256 × 256. In all the cases examined, uniform quantization of the transform coefficients was applied.

For the definition of the first family of analysis / synthesis filters the conversion from YUV to RGB format given by

$$\mathbf{x}_{rgb} = \mathbf{A}x_{yuv}$$

was used where $\mathbf{A}$ is defined by

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1.402 \\ 1 & -0.34414 & -0.71414 \\ 1 & 1.772 & 0 \end{bmatrix} . \quad (29)$$

Thus, one choice for the synthesis vector filter is

$$\mathbf{G}(z) = \mathbf{A}\Lambda(z)$$

where

$$\mathbf{\Lambda}(z) = \begin{bmatrix} \Lambda_1(z) & 0 & 0 \\ 0 & \Lambda_2(z) & 0 \\ 0 & 0 & \Lambda_3(z) \end{bmatrix} \qquad (30)$$

and $\Lambda_i(z)$, for $i = 1, \ldots, 3$ are FIR low-pass filters. In this case the analysis filter would be given by

$$\mathbf{H}(z) = \mathbf{D}^{-1}(z)\mathbf{\Lambda}^T(z^{-1})\mathbf{A}$$

where

$$\mathbf{D}(z) = \mathbf{\Lambda}^T(z)\mathbf{A}^T\mathbf{A}\mathbf{\Lambda}(z) + \mathbf{\Lambda}^T(z^{-1})\mathbf{A}^T\mathbf{A}\mathbf{\Lambda}(z^{-1}) \ .$$

The above filter matrices $\mathbf{H}(z)$ and $\mathbf{G}(z)$ are chosen as the analysis / synthesis filters of the primary band of the vector subband coding scheme. The analysis and synthesis filters in the remaining bands are chosen to satisfy the perfect reconstruction principle. The above filters were tested for the coding of the color image "Peppers". In Figure 1 the proposed technique is compared in terms of PSNR versus bitrate, with the scalar subband coding with the same type of filters [6].

The conversion between the standard Red Green Blue (RGB) format to YUV format may also be used for the definition of the second family of analysis/synthesis filters. The conversion, written in a matrix form, is

$$\mathbf{x}_{yuv} = \mathbf{B}x_{rgb}$$

where

$$\mathbf{B} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.1687 & -0.3313 & 0.5 \\ 0.5 & -0.4187 & -0.0813 \end{bmatrix} . \qquad (31)$$

Thus a good choice for the analysis filters will be

$$\mathbf{H}(z) = \mathbf{B}\mathbf{\Lambda}(z)\mathbf{\Phi}_{\mathbf{X}}(z) \ .$$

The power spectra of the input may be approximated by making the assumption that the input signal can be modeled with a three-channel AR model of the form

$$\hat{\mathbf{x}}[i] = \sum_{m=1}^{p_1} \mathbf{C}[m]\mathbf{x}[i-m] + \mathbf{w}[i]$$

in which the model coefficients $\mathbf{C}[m]$ are 1×3 vectors, and image pixels $\mathbf{x}[m]$ are vectors of length 3. The three-channel Yule-Walker equations are

$$\sum_m \mathbf{C}[m]\mathbf{R}_{xx}[i-m] = \begin{cases} \mathbf{P}_w & \text{for } m = 0 \\ 0 & \text{otherwise} \end{cases}$$

The autocorrelation matrices are given by $\mathbf{R}_{xx}[k,l] = E\{\mathbf{x}[i+k]\mathbf{x}^T[i]\}$. The solution of (V) gives the coefficients $\mathbf{C}[m]$ and the correlation matrix $\mathbf{P}_w$.

The fitness of the multichannel AR models to color image modeling is measured in terms of the prediction mean square error

$$MSE = \frac{1}{M}\sum_i \|\mathbf{x}[i] - \sum_m \mathbf{C}[m]\mathbf{x}[i-m]\|^2$$

The power spectral of the input is then given by

$$\mathbf{\Phi}_{\mathbf{X}}(z) = \frac{\mathbf{\Phi}_{\mathbf{W}}(z)}{1 - \sum_{m_1}\sum_{m_2} \mathbf{C}(m_1)\mathbf{C}^T(m_2)z^{m_1-m_2}}$$

Simulations were performed also using the second family of vector filters and the results were comparable with the ones obtained with the first choice of filters.

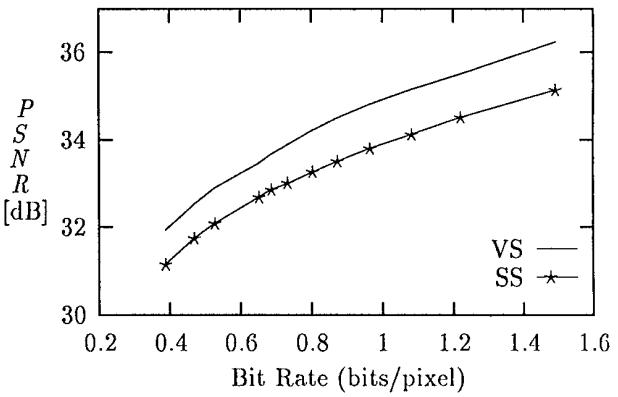


Figure 1: Bit rate versus PSNR performance of the vector subband coding technique (VS), compared to the scalar subband coding (SS) of each vector component.

VI. REFERENCES

[1] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, 1993.

[2] M. G. Strintzis, "Optimal Filters for the Generation of Multiresolution Sequences," *Signal Processing*, vol. 39, No. 2, pp. 55-68, June 1994.

[3] M. G. Strintzis, "Optimal Biorthogonal Wavelet Bases for Signal Representation", *IEEE Trans. Signal Processing*, vol. 44, No. 6, June 1996.

[4] W. Li and Y.-Q. Zhang, "A Study of Vector Transform Coding of Subband-Decomposed Images," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 4, pp. 383–391, Aug. 1994.

[5] J. Wus and W. Li, "Vector Subband Coding of High Resolution Images," in *Picture Coding Symposium (PCS' 94)*, (Sacramento), pp. 123–125, Sep. 1994.

[6] S. N. Efstratiadis, D. Tzovaras and M. G. Strintzis, "Hierarchical Partition Priority Wavelet Image Compression", *IEEE Trans. Image Processing*, vol. 5, No. 7, July 1996.

[7] X. -G. Xia and B. W. Suter, "Vector-Valued Wavelets and Vector Filter Banks", *IEEE Trans. on Signal Processing*, Vol. 44, No. 3, Mar. 1996.

SOURCE CODING OF STEREO IMAGE PAIRS

Halûk Aydınoğlu and Monson H. Hayes

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250 USA
Phone: (404) 894-2910 Fax: (404) 894-8363

ABSTRACT

This paper focuses on the source coding of stereo image pairs. In particular, the conditional coder (CONCOD) and its properties are discussed. The problems attributed to disparity compensation (DC), a widely used conditional coder, are introduced. Finally, a new conditional coding strategy, the Subspace Projection Technique (SPT), is proposed. The SPT is a transform-domain approach with a space-varying transformation matrix and may be interpreted as a spatial-transform domain representation of the stereo data.

1. INTRODUCTION

The human brain can process the subtle differences between the images that are presented to the left and right eyes to perceive a three-dimensional outside world. This ability is called stereo vision. A stereoscopic system may be used to artificially stimulate the stereo vision ability. A stereo pair, a pair of images of the same scene acquired from two different perspectives, is presented to the observer so that the right image is seen by the right eye and the left image is seen by the left eye. The human observer then perceives the scene in depth by processing the relative displacement, i.e., the disparity, of the objects between the two images of a stereo pair. There exists an inherent redundancy between the images of a stereo pair that can be exploited for efficient transmission and storage of stereo images.

2. CONDITIONAL CODER

A stereo pair (X, Y) can be modeled as a vector-valued outcome of two correlated discrete random processes. Since stereo image communication is generally considered an optional extension to the basic monocular system, only a small percentage of the total bit rate is allocated for the second image (or equivalently for the source Y) [1, 2]. Moreover, at least one of the images should be decoded separately. One coding strategy that satisfies both conditions is the conditional coder (CONCOD) and it may be described as "code one image and then code the second image given the coded first image". The CONCOD structure is suboptimal in the sense that

$$R_X(D_X)+R_{Y|X}(D_Y) \le R_{X,Y}(D_X,D_Y) \le R_X(D_X)+R_{Y|\hat{X}}(D_Y), \quad (1)$$

where $R_X(D_X)$ is the rate-distortion function for X, $R_{X,Y}(D_X,D_Y)$ is the joint rate-distortion function for X and Y (the optimal coder), $R_{Y|X}(D_Y)$ is the conditional rate-distortion function for Y given the original X, and $R_{Y|\hat{X}}(D_Y)$ is the conditional rate-distortion function for Y given the encoded X (the conditional coder) [3]. An interesting property for very low bit rate coding of the second image follows if we rewrite Eq.(1) with a looser lower bound:

$$R_X(D_X) \le R_{X,Y}(D_X, D_Y) \le R_X(D_X) + R_{Y|\hat{X}}(D_Y). \quad (2)$$

For the extreme case, in which we allocate zero bits for the second image (the distortion D_Y for the second image is equal to its maximum value $D_{Y_{max}}$, i.e., $R_{Y|\hat{X}}(D_{Y_{max}}) = 0$), the CONCOD structure is optimal since the lower and upper bounds of Eq. (2) are identical. As a result of the continuity of the rate-distortion functions, for any $\epsilon > 0$ there exists a δ such that if $|D_{Y_{max}} - D_{Y_\delta}| < \delta$ then

$$|R_X(D_X) + R_{Y|\hat{X}}(D_{Y_\delta}) - (R_X(D_X) + R_{Y|\hat{X}}(D_{Y_{max}}))| = R_{Y|\hat{X}}(D_{Y_\delta}) < \epsilon. \quad (3)$$

However, since

$$\begin{aligned} R_{X,Y}(D_X, D_{Y_{max}}) &= R_X(D_X) + R_{Y|\hat{X}}(D_{Y_{max}}), \quad (4) \\ R_{X,Y}(D_X, D_{Y_\delta}) &> R_{X,Y}(D_X, D_{Y_{max}}), \end{aligned}$$

we can rewrite Eq. (3) in the following form:

$$|R_X(D_X) + R_{Y|\hat{X}}(D_{Y_\delta}) - R_{X,Y}(D_X, D_{Y_\delta})| < \epsilon, \quad (5)$$

which implies that the CONCOD structure is performing arbitrarily close to the optimal solution given that $R_{Y|\hat{X}}(D_Y)$ is small.

This work was supported in part by the Joint Services Electronic Program, Grant No. DAAH-04-93-G-0027.

Based on this observation, conditional coding techniques that minimize the bit rate for the second image by exploiting the stereo redundancy while preserving a required quality have been investigated [1, 4]. Disparity information, for example, is used to displace pixels in a reference image to form a predicted frame. For some applications the prediction error is transmitted to improve the quality of the coded images. This particular scheme is referred to as the disparity compensated (DC) prediction and is a special case of the CONCOD structure. Note that the operational rate-distortion curve of any DC scheme is bounded from below by the theoretical rate-distortion curve of the CONCOD structure. In general, an intensity-based block matching algorithm is employed to obtain the disparity field [5]. For very low bit rate applications, an important portion of the bit budget is spent for the disparity vectors and DC-based algorithms suffer from several problems including the failure of the constant intensity axiom, false predictions for the occlusion regions, and blocking artifacts [4, 6]. Rate-distortion and perceptual performance of the stereo coding schemes may be improved if the characteristics of the human visual system and physical properties of the stereo imaging systems are exploited. In particular, the following properties are worth mentioning:

- The intensity of light that is reflected from an object and recorded by the camera depends on the position of a camera relative to the object [7]. Moreover, different camera characteristics may yield systematic luminance differences.
- In the transform domain, low frequency coefficients are more correlated than the high frequency coefficients [3]. Moreover, low frequency coefficients are perceptually important [3].
- Exploiting and alleviating the interblock redundancy of the transform domain coefficients diminish the visibility of the blocking artifacts [8].

The next section describes a coding technique that considers these observations.

3. SUBSPACE PROJECTION TECHNIQUE

We recently proposed a new approach to stereo image coding that performs disparity compensation in a spatial-transform domain framework [4]. The novelty of the proposed approach is that an incomplete transform basis for transform domain compensation and a multiplicative gain factor for spatial domain compensation are used. The proposed algorithm, which is called the Subspace Projection Technique (SPT), may be interpreted as a data dependent block transformation.

The SPT obtains an estimate for each $m \times m$ block (or equivalently m^2-dimensional vector in the Euclidean space R^{m^2}) $\mathbf{b}_r$ of the right image by post processing the prediction $\hat{\mathbf{b}}_r$ that is obtained through disparity compensation (DC). The SPT algorithm does not specify the disparity compensation technique. One can obtain disparity vectors using fixed block size DC [5], variable block size DC (VDC) [1], windowed DC (WDC) [2], or hierarchal block matching techniques [9].

As with many data compression algorithms, our motivation is to find an efficient representation for each vector $\mathbf{b}_r$ so that we can transmit or store the given data with fewer bits. The idea of the SPT is to achieve this objective by creating a suitable transformation, $\mathbf{T}$, from the Euclidean space R^{m^2} to a proper subspace $\mathcal{S}$. We consider the use of a block-varying transformation in order to exploit the stereo redundancy explicitly. Therefore, the vector $\hat{\mathbf{b}}_r$ is included in the span set of the subspace $\mathcal{S}$ as well as a set of fixed orthogonal polynomial vectors $\mathbf{V} = \{\mathbf{v}_i\}_1^{N-1}$. The choice of polynomial vectors is motivated by the computational savings they introduce and by the smooth intensity variations found in most natural images. Moreover, the polynomial vectors yield a good approximation to the incomplete discrete cosine transform [10].

The transform domain representation for the vector $\mathbf{b}_r$ is given by:

$$\mathbf{T}(\mathbf{b}_r) = \gamma \cdot \mathbf{b}_0 + \sum_{i=1}^{N-1} \alpha_i \cdot \mathbf{v}_i, \tag{6}$$

where $\mathbf{b}_0$ is the orthogonal component of the vector $\hat{\mathbf{b}}_r$ to the fixed vectors and γ and α_i's are the projection coefficients that can be obtained by:

$$\alpha_i = \frac{< \mathbf{b}_r, \mathbf{v}_i >}{< \mathbf{v}_i, \mathbf{v}_i >} \quad i = 1, 2, \cdots, N-1 \tag{7}$$

$$\gamma = \frac{< \mathbf{b}_r, \mathbf{b}_0 >}{< \mathbf{b}_0, \mathbf{b}_0 >}. \tag{8}$$

The coefficient γ is a multiplicative gain factor that adapts to the local changes in cross-correlation statistics of the stereo data. An average gain of 1-2dB is achieved over the estimate that assumes $\gamma = 1$. A typical histogram of the coefficient γ is presented in Fig. 1. Although, the peak magnitude of the histogram is image dependent, typical values range between $0.8 - 1.1$. The peak magnitude is equal to 0.9 for this example.

The fixed subspace estimate $\sum_{i=1}^{N-1} \alpha_i \cdot \mathbf{v}_i$ is the low frequency component of the block $\mathbf{b}_r$. In general, either one (zero order approximation), three (first order approximation), or six (second order approximation) fixed vectors are used. We employ DC in the

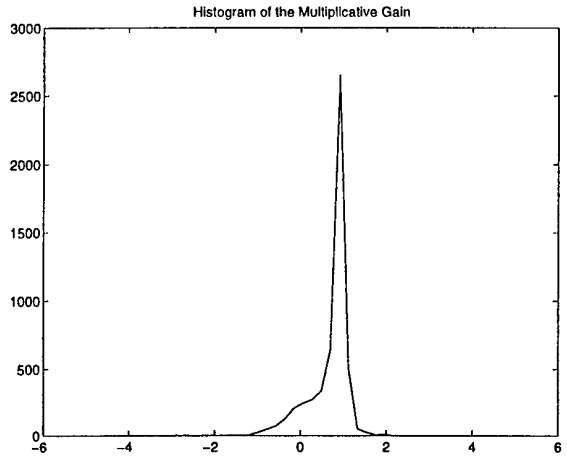


Figure 1: Histogram for the multiplicative gain factor.

transform domain for the fixed subspace coefficients. This is equivalent to correcting for systematic (local and global) luminance differences. The disparity compensated prediction for those are given by:

$$\hat{\alpha}_i = \frac{< \hat{\mathbf{b}}_r, \mathbf{v}_i >}{< \mathbf{v}_i, \mathbf{v}_i >} \qquad i = 1, 2, \cdot, N-1. \tag{9}$$

We transmit the prediction error, $\epsilon_i = \alpha_i - \hat{\alpha}_i$.

There are two implementations for the SPT. The first is to use a large projection block size (e.g., 16-*by*-16) to obtain an initial estimate and to code and transmit the residual given by $\mathbf{b}_r - \mathbf{T}(\mathbf{b}_r)$. The second is to use a smaller projection block size (e.g., 4-*by*-4) and quantize the coefficients ϵ_i's and γ using a subband coder so that the spatial correlation between the coefficients of the neighboring blocks (inter-block redundancy) is exploited and the visibility of the blocking artifacts diminishes [8]. A locally optimal bit allocation scheme is employed to determine the quantizer for each coefficient [2].

4. EXPERIMENTAL RESULTS

This section presents a comparison of several stereo image coding algorithms. In the experiments, we code the left image independently and the right image based on the coded left image.

Three different stereo pairs are used for the experiments. The first pair was obtained from the "flower garden" sequence (frame numbers 0 and 2). These images are 352-by-240. The second is the "Lab" pair (512-by-480), which was obtained by shifting a video camera and taking two pictures of a stationary scene from two horizontally shifted locations. Finally, the third one is the "Room" pair (256-by-256), which was obtained using the same procedure as the "Lab" pair.

method	f.garden	lab	room
DC	22.97	28.84	23.47
WDC	23.24	29.28	23.80
VDC	23.53	29.91	24.16
SPT	23.70	30.87	26.83
V-SPT	24.11	31.51	27.05
W-SPT	23.95	31.13	27.13

Table 1: Rate-distortion performance of the stereo image coding algorithms. The PSNRs of the encoded images (in dB) are presented at a fixed bit rate (0.06bpp for the lab and room pairs and 0.05bpp for the flower garden pair).

First, the performance of several stereo image coding algorithms at very low bit rates (around 0.05 bpp) are compared. These include fixed block size (8-by-8) disparity compensation (DC), windowed DC (WDC), variable block size DC (VDC), subspace projection technique (SPT) using fixed block (16-by-16) size DC estimates, SPT using variable block size DC estimates (V-SPT), SPT using WDC estimates (W-SPT). An average gain of 0.75dB is achieved for the given pairs by using VDC. Occlusion regions are better estimated by this method. Similar gains are obtained by V-SPT. WDC improves the PSNR by 0.32dB on the average over DC for the test data. Moreover, blocking artifacts are removed by the windowing operation. All implementations of SPT are found to be superior to DC-based techniques in the rate-distortion sense. The quality of the images that are coded by the SPT can be improved by either decreasing the projection block size or increasing the number of fixed vectors if a small increase in the bit rate or computational complexity is allowed. For example, for the lab image, one can obtain 31.2 dB (at the same bit rate) if six fixed vectors are used instead of three fixed vectors. Same PSNR can be obtained with three fixed vectors if the projection block size, m, is chosen to be 4. In Fig 2, we present the effect of increasing the number of fixed vectors. At low bit rates both implementation have similar performance [1]. Therefore, the one with the lower computational complexity is preferred.

Second, the performance of the following coding al-

[1] In theory, increasing the number of fixed vectors improves the rate-distortion performance for all bit rates. However, for this particular example a locally optimal bit allocation scheme was used. As a result, the implementation with three fixed vectors performs slightly better than the implementation with six fixed vectors at very low bit rates.

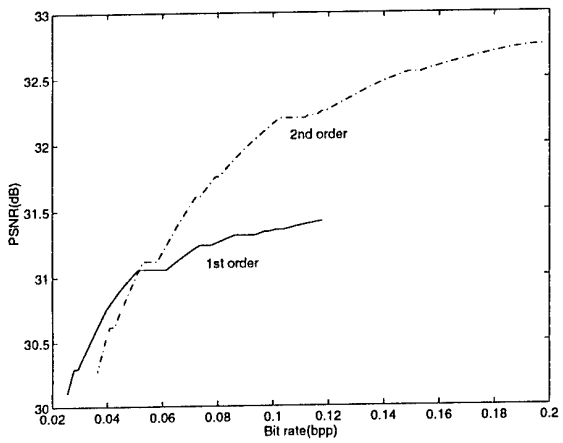


Figure 2: Effects of the number of fixed vectors on the performance of the SPT.

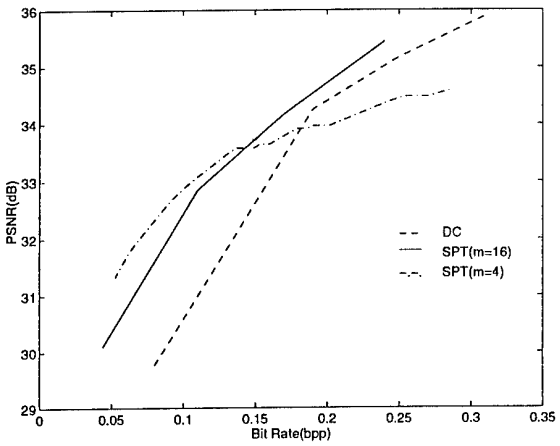


Figure 3: A comparison of the rate distortion performance of the stereo coding algorithms.

method	f.garden	lab	room
DC-RC	24.80	32.91	29.19
SPT-RC	25.84	33.40	30.24
VDC	24.00	30.10	24.72
WDC-RC	25.43	33.15	29.65
WSPT-RC	26.07	34.00	30.55

Table 2: Rate-distortion performance of the stereo image coding algorithms. The PSNRs of the encoded images (in dB) are presented at a fixed bit rate (0.13bpp for the lab pair, 0.16bpp for the flower garden pair, and 0.15bpp for the room pair).

gorithms at low bit rates (around 0.15 bpp) are investigated: Disparity compensated ($n = 16$) residual coding (DC-RC), VDC, windowed DC-RC (WDC-RC), SPT compensated residual coding ($n = 16, m = 16$) (SPT-RC), W-SPT-RC. It is easier to code the residual for windowed techniques. In fact, WDC improves the coding results both perceptually and in the rate-distortion sense. It is better practice to increase the bit rate for the residual than to increase the bit rate for the disparity field. However, slightly increasing the bit rate for the disparity field using the VDC method and then coding the residual field is a feasible solution. Different implementations of SPT are superior to the DC-based techniques. Finally, the operational rate distortion curves for the DC and the SPT algorithms are compared for the "Lab" stereo pair. The SPT is implemented in two different ways as described in the previous section. For the first implementation of the SPT a 4-by-4 projection block size is used, no residual is coded, the projection coefficients are quantized using a subband coder, and a locally optimal bit allocation scheme is employed as described in [2]. For the second implementation a 16-by-16 block size is used for both disparity estimation and subspace projection. The residual is coded using a subband coder. Finally, DC is implemented using 16-by-16 blocks for estimation. The residual due to the disparity compensation is coded with a subband coder. The rate-distortion curves are presented in Figure 3. For low bit rates, the first implementation of the SPT is 3 dB better than the disparity compensation and 2 dB better than the second implementation of the SPT. Another interpretation of this result is that we can achieve a bit rate reduction of 60% over DC. For bit rates higher than 0.15 bpp, the second implementation of the SPT is preferable. Note that the maximum PSNR achieved by the first SPT implementation is bounded from above since even without the quantization, the algorithm cannot achieve exact reconstruction. The original and coded images are presented in Fig 4-5, respectively.

5. CONCLUSION

This paper summarizes our recent work on stereo image coding. We propose a new coding technique based on subspace projection. The novelty of the approach is that the transformation matrix of the projection operation adaptively changes to exploit the inherent stereo redundancy and non-stationary cross-correlation characteristics between the images of a stereo pair. In addition, we used a combined transform-subband coding scheme that is very efficient for coding transform domain coefficients. The subspace projection technique

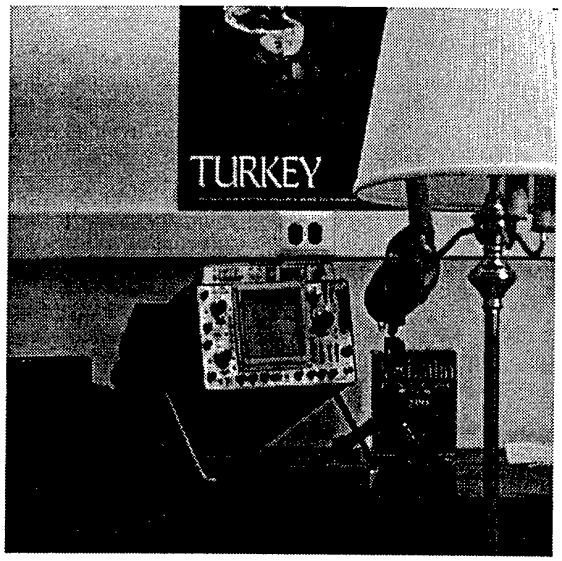


Figure 4: Original right image.

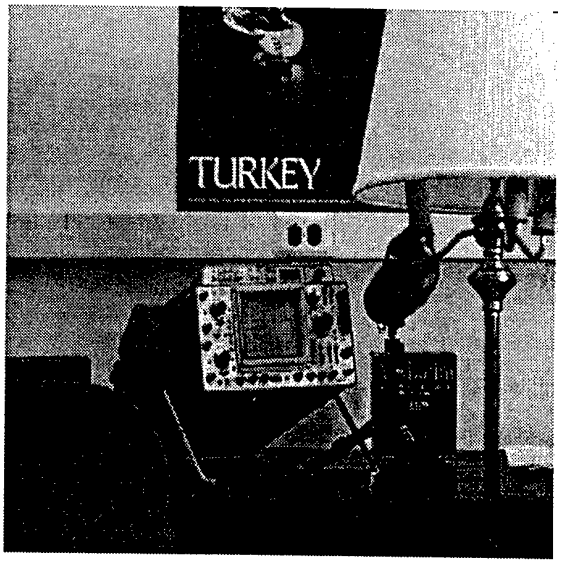


Figure 5: Coded right image at 0.11bpp with a PSNR of 32.85dB using the SPT algorithm and residual coding.

is appealing since its performance at very low bit rates was found to be superior to the standard stereo coding algorithms. The proposed coder is flexible and can operate over a broad range of bit rates and does not require training.

6. REFERENCES

[1] S. Sethuraman, A. Jordan, and M. Siegel, "A Multiresolutional Region Based Hierarchical Segmentation Scheme for Stereoscopic Image Compression," in *Digital Video Compression: Algorithms and Technologies*, vol. 2419, SPIE, 1995.

[2] H. Aydınoğlu and M. H. Hayes, "Performance Analysis of Stereo Coding Algorithms," in *ICASSP*, vol. IV, May 1996.

[3] M. G. Perkins, "Data Compression of Stereopairs," *IEEE Trans. on Communications*, vol. 40, pp. 684–696, April 1992.

[4] H. Aydınoğlu and M. H. Hayes, "Stereo Image Coding: A Subspace Approach," *submitted to Trans. on Image Processing*, 1996.

[5] H. Yamaguchi, Y.Tatehira, K. Akiyama, and Y. Kobayashi, "Stereoscopic Images Disparity for Predictive Coding," in *ICASSP*, pp. 1976–1979, IEEE, 1989.

[6] H. Aydınoğlu and M. H. Hayes, "Compression of Multi-View Images," in *International Conference on Image Processing*, vol. II, pp. 385–389, November 1994.

[7] S. D. Cochran and G. Medioni, "3-D Surface Description from Binocular Stereo," *IEEE Trans on PAMI*, vol. 14, pp. 981–994, October 1992.

[8] K. Lim, K. Chun, and J. Ra, "Improvements on Image Transform Coding by Reducing Interblock Correlation," *IEEE Trans. on Image Proc.*, vol. 4, pp. 1146–1150, August 1995.

[9] S. Sethuraman, A. Jordan, and M. Siegel, "Multiresolution Based Hierarchical Disparity Estimation for Stereo Image Compression," in *Symposium on Application of Subbands and Wavelets*, 1994.

[10] H. Aydınoğlu and M. H. Hayes, "Image Coding with Polynomial Transforms," in *submitted to Asilomar*, 1996.

ADAPTIVE BLOCK-BASED MOTION ESTIMATION IN VIDEO CODING

Marco Accame, Francesco G.B. De Natale, and Daniel D. Giusto

Dept. of Electrical and Electronic Engineering, University of Cagliari, Italy
ph: 39 (70) 675-5866 - fax: 39 (70) 675-5900 - email: giusto@diee.unica.it

ABSTRACT

A block based motion estimation (BBME) strategy that uses blocks of adaptive size for encoding video sequences. A novel joint strategy for the vector propagation and the adaptive correction has been implemented, that allows both an efficient level-to-level update of the motion field and an effective recovery from wrongly estimated vectors. The developed algorithm yields a spatial quantization of the motion field, which is locally variable and it allows to segment regions with uniform motion in big blocks. The method produces a reduction both in bitrate and in computational load by keeping the reconstruction quality nearly constant, when compared to the classical BBME.

1. INTRODUCTION

In video coding, motion compensation techniques predict future frames of the video sequence using a previously computed estimation of the objects' motion. In H.261 recommendations [1], as well in MPEG [2,3] and in some other video coding techniques, the motion field is computed blockwise, by a spatial quantization of the dense optical flow into a few displacement vectors associated to squared blocks. The estimation of each vector is usually performed by comparing the block of the future frame with every displaced block of the current frame that falls within a pre-defined search window (full search). The lowest distortion measure gives the searched displacement.

In this paper, a new method that performs motion estimation for video coding is presented. The method allows to achieve a motion field that is spatially quantized with blocks of variable sizes, keeping on their inside constant motion activities, so that a large number of motion vectors are joined in big blocks. The estimation is performed by processing the motion at several level of resolution and by choosing, level by level on a local basis, the dimension of the block that fits the motion field. The initial estimation is achieved by a full-search BBME at the coarsest level of a multiresolution pyramid, whereas on the next levels a novel adaptive strategy performs only the necessary correction to the field propagated from the previous level. The innovative strategy propagates and fixes the motion field across the levels as follows. Starting from the coarse spatial quantization of the motion field with large blocks achieved in the first level, it decides which regions require a finer spatial quantization with smaller blocks. For such regions, the strategy initializes new displacement vectors in the next resolution level and decides for each vector the required amount of correction. It is so possible to fit the correction for each vector, and to perform actions ranging from a robust error recovery to only a small correction of previously estimated vectors; the computational load and the number of noisy vectors are both reduced. For the regions where a denser spatial quantization is not needed, the blocks do not generate new blocks and they are just propagated increasing their size level by level, whereas the related displacement is refined in its value. This joint strategy of propagation and correction acts locally and yields a not-uniform spatial quantization of the motion field. With respect to the traditional BBME with fixed size blocks, these fewer vectors are enough for achieving nearly the same reconstruction quality, as they are placed according to the motion activities in the scene. Moreover, the computational load is reduced; at first because the number of vectors is reduced too, and then because for each one of them the number of computations is carefully tuned on the basis of the local field.

2. HIERARCHICAL BBME

Classical BBME has the serious drawback that it has a huge computational load; considering to divide an image with N and M square blocks of size B in the two spatial directions, and a search window S of $\pm W$, the number of elementary operations (EOPs) required for full search is:

$$\Delta_f = M\,N\,B^2\,(2W+1)^2$$

When a CIF-formatted luminance picture is used, with B=16 and W=16, the number of operations grows to about Δ_f=1.1x10^8. For this reason, the full search BBME cannot be easily managed for real time encoding of video sequences and fast search methods have been studied for increasing the speed of the estimation task. In recent years, several works have been presented that are based on a hierarchical approach for BBME [4,5]. Hierarchical methods are well suited for motion estimation, as they give robust motion fields and they can lower by even two orders of magnitude the computational load of the full search BBME. Let us suppose to use a L-levels multiresolution representation (e.g., the Gaussian pyramid introduced in [6]). It includes the original image as the (0)-th level and its L-1 approximations at a dyadic sequence of resolutions. At the $(L-1)$-th level, the BBME is carried out with blocks of dimension B by using a full search in a window that is reduced by a factor 2^{L-1} respect to the case of fixed

resolution (i.e., the search is done within $\pm W2^{-L+1}$). At the coarsest level, the number of blocks in each spatial direction is reduced too by the same factor, resulting $M2^{-L+1}$ and $N2^{-L+1}$ respectively. Thus, only Δ_{L-1} elementary operations are required at the $(L-1)$-th level for the coarsest BBME and Δ_l for the corrections in each of the remaining levels $(l=L-2,\cdots,0)$. But, as the main displacement has been processed in the coarsest level, only a reduced offset $v_l << W2^{-l}$ is used for Δ_l for it needs to process just scale errors, and the resulting number of elementary computations Δ_H is greatly lower if compared to Δ_f (i.e., the value of the equivalent full search done at the full resolution). For $v_l=v$ invariant with the

$$\begin{aligned}\Delta_{L-1} &= M\,N\,B^2\,2^{-2L+2}\,(2\,W2^{-L+1}+1)^2\\ \Delta_l &= M\,N\,B^2\,2^{-2l}\,(2\,v+1)^2\\ \Delta_H &= \Delta_{L-1} + \Sigma_l\,\Delta_l = \\ &= \Delta_{L-1} + (4/3)\,M\,N\,B^2\,(2\,v+1)^2\,(1-2^{-2L+2})\end{aligned}$$

In the case of a hierarchical BBME for a CIF sequence, using a $L=3$ multiresolution pyramid and a correction at each level of $v=2$, the number of computations is about $\Delta_H = 3.7\text{x}10^6$ for a gain of 30, whereas the gain becomes 65 using $v=1$, and 92 using $L=4$ and $v=1$. However, the great gain of the hierarchical BBME produce a result in the final estimation that is similar to the full search only if the first estimation is coherent to the true motion at that resolution, even if it is coarsely spatially quantized. Indeed, a block of size B at the $(L-1)$-th level is equivalent to a block of size $2^{-L+1}B$ of the (0)-th resolution level. If the block contains several objects with different motions on its inside, the first estimation surely gives a vector that minimizes the error at the $(L-1)$-th level, but that requires large corrections on the next ones. In this situation, a small value of v cannot recover from the false estimates done in the earlier levels. On the other hand, a greater value for v allows a better correction but a lower computational gain. One solution is to improve the propagation of the motion field toward the next level of the pyramid. A vector in the (l)-th level generates four new vectors in the $(l-1)$-th level, each one associated to a block of dimension B as its father on the previous level. Instead of replicating the value of the father vector (with a factor of 2 due to the resolution increase), in [7] the motion field is propagated by searching for the best initial vector among the neighboring vectors at the previous level. In this way, a bad propagation coming from the father vector can be recovered if the son block has motion activities similar to the ones of any spatially adjacent block. In other words, the high spatial correlation of the motion field is exploited for error recovery.

3. PROPAGATION OF THE MOTION FIELD IN HIERARCHICAL BBME

Hierarchical BBME methods allow to process motion activities at different scales, and they are well suited for reducing the computational load. However, it is necessary that the initial coarse estimation is properly propagated and corrected in the finer levels. Moreover, the introduction of variable size blocks in BBME helps to reduce both the computational load and the number of the vectors of the motion field that need to be coded. Thus, a hierarchical BBME scheme that aims at reducing both the number of vectors and the computational load, without reconstruction degradation, must have the following properties. At first, it must be able to recover from earlier incoherent estimations. That can be achieved by a proper choice of the vector field that is to be refined at each level, and by setting adaptively the correction window on the basis of the required correction. Then, the improvement of the spatial resolution of the motion field should stop when smaller blocks do not take a substantial improvement on the reconstruction. These properties better the estimation and the coding performances. Indeed, the propagation stop allows to have less vectors to code, whereas a good joint propagation and correction strategy allows to achieve smooth motion fields with fewer noisy vectors, so that the entropy coding of the field may give better results.

The used multiresolution representation is the L-level Gaussian pyramid [6], already introduced in Section II. Here, the first BBME is performed at the $(L-1)$-th level with B-sized blocks and the resulting motion field is evaluated block by block for its propagation toward the next level and for its correction. The propagation and the correction of the motion field is performed until the finest resolution is reached at the (0)-th level. Let $b_{B,l}(i,j)$ be the generic B-sized block at the (l)-th level; there are $M2^{-l}$ and $N2^{-l}$ blocks in the two spatial directions, thus $i\in\{1,\cdots,N2^{-l}\}$ and $j\in\{1,\cdots,M2^{-l}\}$. In case the compensation for the block $b_{B,l}(i,j)$ is not satisfactory enough, the block generates four new blocks at the next level where their displacements are corrected on the basis of their necessities. These four B-sized blocks encloses the same area of the scene as $b_{B,l}(i,j)$ does, because they lie on the $(l-1)$-th level where the scale is doubled, and they allow to improve the estimation by means of a finer spatial quantization.

Otherwise, if the compensation of $b_{B,l}(i,j)$ is satisfactory, the displacement is projected directly on the (0)-th level, where it is applied to a block of dimension $B2^{+l}$ and the estimation for the area it encloses is stopped.

3.1. Adaptive Size Block Generation

Each block $b_{B,l}(i,j)$ is made up of four sub-blocks of size $B/2$ at the same level: $b_{B/2,l}(2i,2j)$, $b_{B/2,l}(2i+1,2j)$, $b_{B/2,ll}(2i,2j+1)$, and $b_{B/2,ll}(2i+1,2j+1)$. The MAD of the compensation is evaluated for each of these by using $d_{B,l}(i,j)$, i.e., the same displacement vector of the block $b_{B,l}(i,j)$. If at least one of them has a MAD that results to be higher than a fixed threshold T_L, the block $b_{B,l}(i,j)$ is split and generates four new blocks at the $(l-1)$-th level that are: $b_{B,l-1}(2i,2j)$, $b_{B,l-1}(2i+1,2j)$, $b_{B,l-1}(2i,2j+1)$, and $b_{B,l-1}(2i+1,2j+1)$. The new blocks are given four adequate initial displacement vectors that are the starting position for the correction of the motion field in this area. The split and update process is iterated across levels (see Fig.1), until the compensation error given by stopping the spatial segmentation of the motion field with the block $b_{B,l}(i,j)$ is

satisfactory. This condition should be checked at the *(0)-th* for every block, by displacing the block $b_{2^lB,0}(i,j)$ of dimension 2^lB, that is the projection on the finest resolution of $b_{B,l}(i,j)$. However, for reducing useless computations, the *(0)-th* level compensation for verifying a propagation stop is performed only for a block that reveals good reconstruction at the *(l)-th* level. This block should contain a constant motion for every pixel and the reconstruction error should be uniformly distributed in its inside. The MADs computed inside its sub-blocks are used, for they do not give further computations as higher order moments would need. In conclusion, the MADs of the sub-blocks are compared to the threshold, and if all of them are lower than T_L, the block is projected on the finest resolution level because there is an hypothesis of propagation stop that is to be checked.

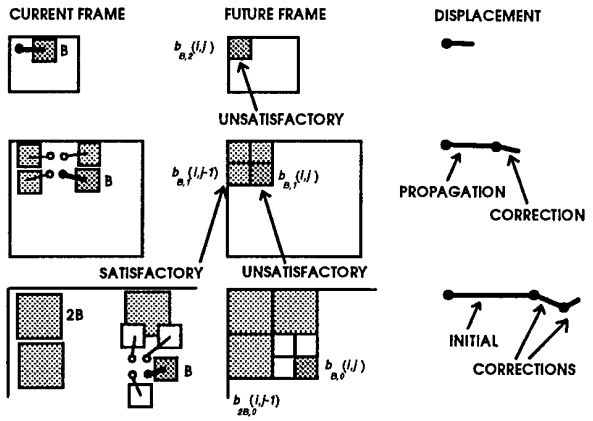


Figure 1. An example of the adaptive size BBME.

To perform this control, the vector $d_{B,0}(i,j)$ is scaled by a factor 2^l, and the compensation is evaluated with the best vector $d_{2^lB,0}(i,j)$ that lies within $\pm 2^{l-1}$, that is a resolution window that allows to reach the pixel accuracy at the *(0)-th* level (see Fig.2). Within the resolution window, the MAD results to be a monotonic function of the displacements; thus, the global minimum can be reached by using a fast search method for BBME. In our work we have used a modified version of the algorithm proposed in [8], that allows to find the best displacement computing a reduced number of EOPs. If the compensation is satisfactory, the block $b_{B,l}(i,j)$ is not further split and just one vector is used for its displacement field instead of 2^{2l} vectors; this fact gives a considerable gain in coding efficiency, since $4^l + \cdots + 4^l = (4/3)\ [4^l\text{-}1]$ vectors are avoided by stopping to propagate a block at the *(l)-th* level. The mechanism allows to create blocks of different dimensions, which adaptively cover the future frame, obtaining a higher vectors' concentration where a more dense displacement field is needed. However, since the motion field is not spatially quantized with a regular block disposition, it is necessary to code the position of the blocks inside the future frame. The propagation of the blocks within the Gaussian pyramid is equivalent to a quad-tree growing process, whose root is the father block at the *(L-1)-th* level. This helps to code very compactly the side information needed to place the blocks; it is just necessary to code with one bit each split decision and, based on such information, it is possible to reconstruct the quad-tree and place the blocks on the future frame with the related vectors so that the decoder can perform the compensation.

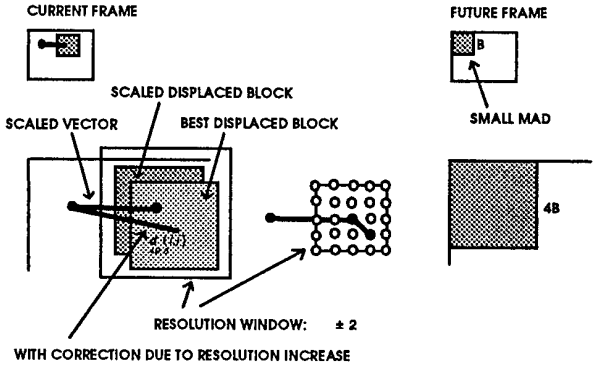


Figure 2. Example of scaling from the *2-nd* to the *(0)-th* level with relative resolution window for testing the compensation quality.

3.2. Vector Propagation Strategy

The propagation of a vector from the generic *(l)-th* level is different whether the related compensation is satisfactory or not. In the first case the vector is at first expanded on the *(0)-th* level, and then the best $d_{2^lB,0}(i,j)$ within the correction window of Fig.2 is taken. In the other case, four new vectors are propagated to the *(l-1)-th* level as follows: (i) before beginning the correction phase, each vector among the set of its neighbors at the *(l)-th* level is scaled to the *(l-1)-th* level, and it is used for compensating the associated block; the vector with the lowest MAD is chosen as the starting point; (ii) during the correction phase at the *(l-1)-th* level, the vector previously initialized is again compared to its neighbors on the same *(l-1)-th* level, that have been already corrected. The one giving the lowest MAD is taken. The first step allows to exploit the neighborhood of the motion field computed on the *(l)-th* level for recovering from a wrongly estimated motion vector. If a *B*-sized block of the *(l)-th* level contains some objects with different motions, the estimated vector that minimizes the MAD cannot be the real motion. In this case, if a sub-block contains just one of these objects and an adjacent block is completely covered by the same object (so that its displacement estimation is correct), then, the sub-block could start from the true displacement for its motion correction in the *(l-1)-th* level. The second step, allows to start with more accurate initial vectors if the neighbor blocks at the same level have a similar motion. Indeed, it is necessary to perform the correction just for the first block that covers an area with constant motion (e.g., an object moving against a still background) and then the same correction is taken by its neighbor blocks that cover the same area.

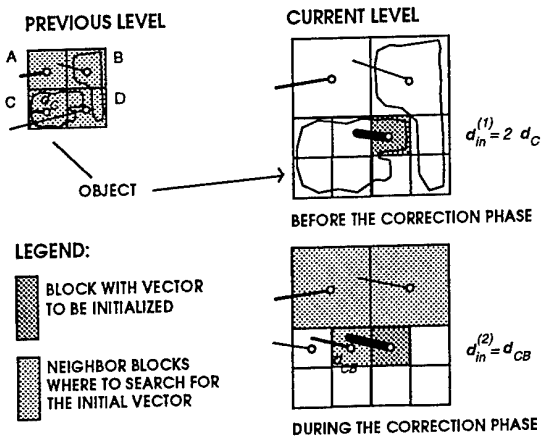


Figure 3. At first it is chosen the initial vector from the block C in the previous level as the best among those of blocks {A, B, C, D}. Then, it is chosen among the already refined neighbor vectors.

3.3. Adaptive Size Setting for the Search Window

With an initialization strategy as the one seen in the previous Sub-Section, the initial vectors are as much as possible near (or even equal) to the target vector. Thus, to reduce both the computational load and the possibility to fall in a false displacement, it is necessary to set the correction window so that its size is proportional to the amount of required correction. For instance, if the MAD of the initial vector is large (i.e., the compensation is not satisfactory enough) a small correction may be not enough and even a new estimation may be required; but if only with the initialization phase the compensation is satisfactory, it is not necessary to search any more. In order to modulate between these two extreme situations, the width of the search window has been set on the basis of the compensation MAD of the initial vector. In particular the highest value of the MADs of the four sub-blocks is considered. If the initial MAD has a value lower than the threshold T_L, then, the hypothesis for a propagation stop is fulfilled, a further correction is not needed anymore and the search window is set to zero. Otherwise, if it is bigger than an upper threshold T_H, a simple correction is not enough and a new estimation is performed: the initial vector is set to zero and the window size is set to the maximum displacement allowed at that level, i.e., $\pm W2^{-l}$. In all other cases the correction window size is set proportional to the initial MAD value. By applying this strategy, it is possible to recover from bad estimations or to perform only small corrections on already satisfactory estimations, by using a search-window width which is set adaptively.

4. EXPERIMENTAL RESULTS

The proposed adaptive BBME allows to reach the same reconstruction quality of the full search BBME, with a reduced number of vectors. In Fig. 5, we show the results for the standard CIF sequence *Salesman*. The full search BBME has been performed with block size B=16 within a search window of W=±16. Moreover, for the hierarchical BBME, the number of the levels of the Gaussian pyramid have been set to L=3, the minimum block dimension is B=16 and the initial search windows at the coarsest level to ±4. The hierarchical method is able to exploit the same displacements of the full search and it allow blocks of dimensions B_0=16, B_1=32 and B_2=64. For T_L and T_H two settings have been chosen, so that different performances are obtained. The first allows to achieve a reconstruction quality as similar as possible to the one of the full search, but with a reduction in the number of the used vectors. The latter setting reaches a higher decrease in the number of used vectors, but with a reconstruction quality that results to be a little lower to the one of the full search. The diagrams show the comparative values for the PSNR of the achieved compensations, and the entropy coding of the motion field achieved using adaptive arithmetic coding [9]. PSNR values are very similar for each considered compensation, whereas the bitrate for the adaptive method is lower even by a factor two respect to the full search method. Finally, the compensations between the 2nd and the 4th frame of *Salesman* are shown. It is reported the compensated 4th frame, its DFD with an offset of 128 and quantized with 1 bpp, and the relative motion field. The full search is compared to the adaptive method with two different settings of the thresholds. The reconstruction quality is nearly the same for the three compensations, but the number of vectors and the bitrate is greatly reduced for the adaptive BBMEs. Moreover, the method gives a gain in computational load that reaches even a factor 100, with L=3 levels of the multiresolution representation, that can be compared with the factor 65 achieved in Section.

REFERENCES

[1] CCITT SG XV, "Recommendation H.261," 1990.

[2] ISO-IEC JTC1 CD 11172, "Coding of Moving Pictures and Associated Audio for Digital Storage Media up to 1.5 Mbps," 1991.

[3] ISO-IEC/DIS 13818-2, "Generic Coding of Moving Pictures and Associated Audio Information," 1994.

[4] M.Bierling, "Displacement estimation by hierarchical block matching," *Proc. SPIE-VCIP*, 942-951, 1988.

[5] H.Dufaux, F.Moscheni, "Motion Estimation Techniques for Digital TV: A Review and a New Contribution," *Proc. IEEE* 83(6), 858-876, 1995.

[6] P.J.Burt, E.H.Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans.* COM-31(4), 532-540, 1983.

[7] P.Anandan, "A computational framework and an algorithm for the measurement of visual motion," *Int. J. Comp. Vision* 2, 283-310, 1989.

[8] Ghanbari, "The Cross-Search Algorithm for Motion Estimation", *IEEE Trans.* COM-38(7), 950-953, 1990.

[9] R.M.Witten, I.H.Neal, J.G.Cleary, "Arithmetic Coding for Data Compression," *Commun. ACM* 30(6), 520-540, 1987.

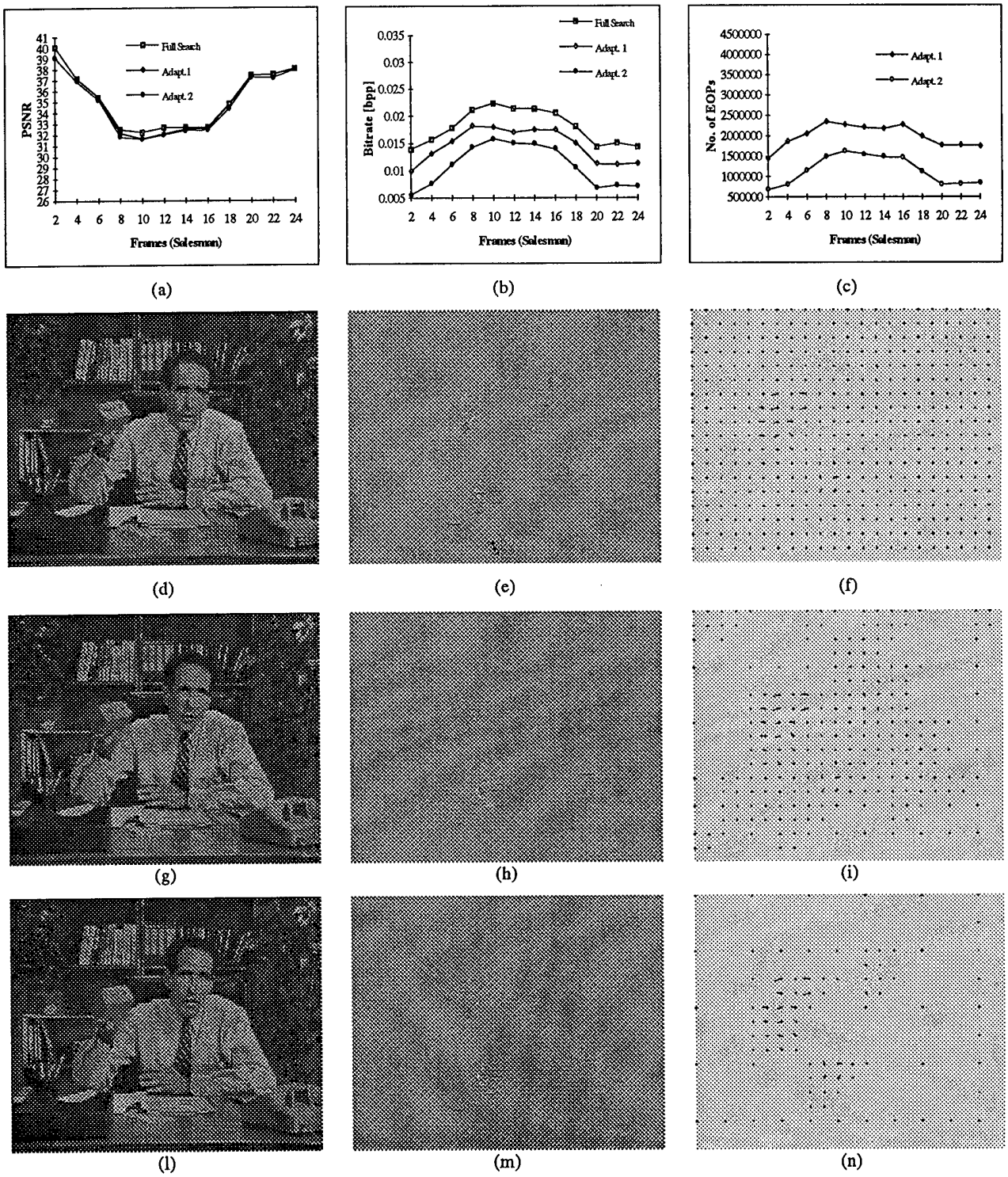


Figure 5. In (a, b, c), the diagrams show the PSNR of the achieved compensations, the bitrate for coding the motion field and the computational load (for full search there are 99,847,168 EOPs). The reconstructed frames, the DFD images and the needle diagrams are reported for the compensation between the 2nd and the 4th frame of *Salesman*. The full search BBME (d, e, f) gives a PSNR=37.19 with 396 vectors coded with 0.0157 bpp. The proposed method allows to reach the same PSNR=37.19 with 225 vectors coded with 0.0123 bpp (g, h, i). By a different tuning of the thresholds, with only 87 vectors and 0.0073 bpp it is possible to achieve a PSNR=36.93 (l, m, n).

EXPLOITING CHARACTERISTICS OF A LARGE NUMBER OF MPEG VIDEO SOURCES FOR STATISTICALLY MULTIPLEXING VIDEO FOR TV BROADCAST APPLICATIONS

Luís Miguel Lopes Teixeira and Teresa Andrade
INESC Porto, Largo Mompilher 22, 4000 PORTO, PORTUGAL
UCP, Rua Diogo Botelho 1327, 4150 Porto, PORTUGAL
email: lmt@inescn.pt

ABSTRACT

In this paper, an efficient algorithm for multiplexing video sources, using a dynamic bandwidth allocation scheme, is presented. Video sources are grouped into classes regarding different combined levels of spatial detail and amount of movement. Simulations are being performed and quality improvements have been obtained for sequences with higher local activity/motion.

1. INTRODUCTION

Digital technology and techniques are progressively being introduced in TV broadcast applications. Among digital techniques video compression emerges by allowing the use of a significantly smaller bandwidth for the transmission of a full-motion video digital signal, while still providing a high-quality service. It is thus foreseen its use in the TV broadcast industry, in an economic and efficient way.

MPEG [1][2] specifies the compressed audio and video bitstream formats and how to maintain synchronisation between the two. With the MPEG2 video standard an almost constant-quality signal can be produced by allowing the bit rate of the compressed bitstream to vary. Therefore the video coder will produce more bits when encoding a more complex picture, in order to maintain constant the overall quality.

Television channels will be MPEG encoded and transmitted. Considering the fact that they are statistically independent, it is very unlikely that encoders working in parallel on different video sources, will be dealing simultaneously with difficult scenes to encode. Or, in other words and applying this concept to TV broadcasting, it is very unlikely that all TV channels will be presenting at the same time scenes with the same kind of complexity and motion. The most probable situation is to have distinct TV channels showing different kinds of programmes. Some will be presenting sporting events with more or less motion, while others will be transmitting news programmes, documentaries, movies, talk-shows, etc. Thus by allowing individual encoders to generate variable bit rates, transmission bandwidth is dynamically allocated to each channel according to their individual needs [3][4]. The entity responsible for multiplexing the different TV channels in one single transmission link, will automatically assign more bandwidth to a channel when it detects a great amount of complexity in the scene being encoded. This will allow a more efficient use of the total

available transmission bandwidth, while maintaining nearly constant quality across all sources. This scheme can be applied to satellite or terrestrial broadcasting as well as ATM transmission.

The paper is organised as follows: section II presents an algorithm to classify the video sources and to dynamically allocate the total bandwidth for each channel regarding its needs, section III refers to the associated system and network problems and in section IV some simulation results are presented. Finally, section V, discusses the results and future work.

2. DYNAMIC BANDWIDTH ALLOCATION ALGORITHM

The dynamic bandwidth allocation algorithm consists in 4 steps and is applied during a picture period.

In the first step, the reference bandwidth (BWref) of each video source is determined based on the total available transmission bandwidth, the picture coding complexity and type, GOP structure of each video source and the current state of the total virtual buffer.

$$BW_{ref\_I} = \frac{R \times X_i}{X_i + \frac{N_P \times X_p}{K_p} + \frac{N_b \times X_i}{K_b}} \quad (1)$$

$$BW_{ref\_P} = \frac{R \times \frac{X_p}{K_p}}{X_i + \frac{N_P \times X_p}{K_p} + \frac{N_b \times X_i}{K_b}} \quad (2)$$

$$BW_{ref\_B} = \frac{R \times \frac{X_b}{K_b}}{X_i + \frac{N_P \times X_p}{K_p} + \frac{N_b \times X_i}{K_b}} \quad (3)$$

where

$$X_i = \frac{160 \times bit\_rate}{115},\ X_p = \frac{60 \times bit\_rate}{115} \quad (4)$$

$$X_b = \frac{42 \times bit\_rate}{115},\ R = bit\_rate \times \frac{N_{GOP}}{picture\_rate} \quad (5)$$

BW_{ref_I}, BW_{ref_P} and BW_{ref_B} are bandwidth for I, P and B pictures for a video source within a GOP. $K_P = 1.0$ and $K_b = 1.4$ are constants dependent on the quantisation matrices [5] and N_p and N_b are the numbers of P pictures and B pictures remaing in the current GOP in the encoding order. R is the transmission bandwidth allocate to the channel during one GOP and N_{Gop} is the total number of pictures in the GOP. bit_rate is the ratio between the total available transmission bandwidth and the number of sources. X variables are the "global complexity measure" of the different pictures types. They are updated by calculating the product of the number of bits generated by encoding a picture and the average quantization parameter (computed with the actual quantization values used during the encoding of all macroblocks, including the skipped ones) for each of the different pictures types.

In the second step, the estimated bandwidth is determined for the optimal distribution of the total available transmission bandwidth according to picture coding type and video source complexity. A measure

of the picture complexity is obtained by computing the average value of spatial local activity. For the macroblock j, spatial local activity is measured from the four luminance frame-organised sub-blocks and the four luminance field-organised sub-blocks, using the original pixel values:

$$act_j = 1 + \min_{sblk=1,8}(\text{var\_}sblk) \quad (6)$$

where

$$\text{var\_}sblk = \frac{1}{64}\sum_{k=1}^{64}(P_k - \overline{P})^2 \quad (7)$$

$$\overline{P} = \frac{1}{64}\sum_{k=1}^{64} P_k \quad (8)$$

and P_k are the pixel values in the original 8×8 block.

The complexity of the scene being coded can be estimated from the complexity of the different frame types in the GOP: the spatial complexity of the I frame and the motion, which determines the complexity of the P and B frames.

Class	Content Complexity	Video test material
A	low spatial detail & low amount of movement	mad, akiyo, hall monitor, container ship, sean
B	medium spatial detail & low amount of movement or vice versa	foreman, news, silent, coast guard
C	high spatial detail & medium amount of movement or vice versa	bus, table tennis, stefan, m&c

Table 1) Classification regarding content complexity

The sequences used in the simulation are divided according to their content complexity (table 1) measured through the mean and the variance values of the local activity.

In the third step, the available bandwidth is allocated to each video source by considering the estimated bandwidth.

$$BW_i = BW_{Est_i} \times \frac{\sum_{i=1}^{n} BW_{ref_i}}{\sum_{i=1}^{n} BW_{Est_i}} \quad (9)$$

where n is the number of video sources.

Finally, in the last step, a reference value for the quantization parameter is determined for the picture. The reference value of the quantisation parameter is then modulated according to the spatial activity in the macroblock to obtain the value of the quantization parameter, mquant, that is used to quantize each macroblock.

3. NETWORK AND SYSTEMS ASPECTS

The implementation of the proposed bandwidth allocation scheme for multiplexing several MPEG2 video sources in broadcast applications, presents some challenges and raises a number of problems which are presently being addressed.

One of such aspects concerns the fact that individual encoder sources and multiplex buffer are physically apart. This implies that a protocol must be defined so that communication links may exchange valuable information between them. The protocol should take in account several requirements such as the quality of service imposed by the system (minimum delays and losses, procedures to add/drop new TV channels, ...).

service imposed by the system (minimum delays and losses, procedures to add/drop new TV channels, ...).

To study these problems, a simulation framework has been defined so that different protocols, in an ATM network, and its feasibility may be tested. A software distributed environment is being implemented allowing video encoders, in different machines, to communicate with the multiplexer.

Apart from the problem of solving in an efficient way the exchange of control information, it is also necessary to define the type of information and the actual place where that information is to be conveyed from encoders to multiplexer - it may be transmitted embedded in the MPEG2 flows or exchanged using a separated connection. Experiences are being performed in order to asses the best solutions for all of these matters.

4. SIMULATION RESULTS

Simulations were performed using MPEG video test sequences as "Bus", "Foreman" and "Akiyo". They were grouped into three different classes, each one exhibiting different combined levels of spatial detail and amount of movement (table 1). Each sequence consist of 124 frames (352×288 pels). We began our studies with 2 different video sources, from different classes (table 2), and progressively add more sources (table 3). The last line in table 2 & 3 (independent) refers to normal CBR video encoding (at 1.5 Mbps) and transmission.

Analysing table 2 and 3 we can observe that class C sequences (high spatial detail & medium amount of movement or vice versa) present the higher gains being nevertheless still the sequences with inferior quality. This gain is achieved at the cost of a reduction in class A sequences.

Class A	Class B	Class C	
35,8	37,4	n.a.	A B
31,9	n.a.	22,6	A C
n.a.	32,0	21,9	B C
37,8	34,6	20,0	Independent

Table 2) SNR for multiplexing 2 video sources (dB)

Class A	Class B	Class C	
32,5	33,4	23,6	A B C
26,4	36,5	n.a.	A A B
35,2	35,9	n.a.	A B B
26,2	n.a.	24,6	A A C
30,0	n.a.	21,6	A C C
n.a.	36,5	24,6	B B C
n.a.	33,9	24,3	B C C
37,8	34,6	20,0	Independent

Table 3) SNR for multiplexing 3 video sources (dB)

Simulation results shows that uniform picture quality is maintained between different video sources when multiple video sources are multiplexed.

5. DISCUSSION

This paper presents an algorithm for dynamic bandwidth allocation which allots the available bandwidth according to the needs of each video source. Video sources are grouped into three different

classes, each one exhibiting different combined levels of spatial detail and amount of movement. Simulation results show that bandwidth gains/quality improvements are more significant when heterogeneous sources are multiplexed together, specially when video sources from classes C and B are present. In our future work we will include video sequences with shot changes, sequences not GOP aligned and with variable GOP sizes. Further work needs also to be done on the implementation of the protocol to improve the management of the communication between video encoders and multiplexer.

REFERENCES

[1]-ISO/IEC IS 11172, "Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1,5 Mbit/s-video", Geneva, 1993.

[2]-ISO/IEC IS 13818: Generic coding of moving pictures and associated audio, November 1994

[3]-H. Heeke, "Statistical multiplexing gain for VBR video codecs in ATM networks", Proc. 4th International Workshop on Packet Video, Tokyo, Japan, March 1991.

[4]-J- P. Leduc, "Multiplexing Digital Television Sources on ATM Networks", Signal Processing: Image Communication 6, 1994

[5]ISO/IEC-JTC1/SC29/WG11 MPEG93/457 MPEG Video Test Model 5 (TM-5), April 1993.

Video Coding Standard Conversion in Distributed Multimedia System

K. Fazekas*, J. Turan**, I. Erenyi***,

* TU Budapest, Dept. of Microwave Telecommunications
1111 Budapest, Goldman ter 3, Hungary
Phone: +36 1 4631559, fax: +36 1 4633289
E-mail: t-fazekas@nov.mht.bme.hu

**TU Kosice, Dept. of Radioelectronics,
Park Komenskeho 13, 04120 Kosice, Slovakia
Phone: +42 95 35692, fax: +42 95 35692
E-mail: JTURAN@CCSUN.TUKE.SK

*** KFKI Research Institute for Measurement and Computer Techniques
1525 Budapest, P.O. Box 49, Hungary
Phone: +36 1 1696279, fax: +36 1 1601290
E-mail: erenyi@sunserv.kfki.hu

Abstract

Multimedia product and applications pose particularly difficult challenges for user interface designer. One of this tasks is the development of transcoding units. The general aspects of transcoding procedure in multimedia system are investigated in this paper. The related field is the development of an educational multimedia system.

1. Introduction

Recent advances in computing and communication technologies promise to create an infrastructure in which computer systems will support a wide range of interactive multimedia services in a variety of commercial and entertainment domains. Research and development efforts in multimedia computing fall into two groups. One group concentrates on the stand-alone multimedia workstation and associated software systems and tools. The other combines multimedia computing with distributed systems. Potential new applications based on distributed multimedia systems include multimedia information systems, conferencing systems, on-demand multimedia services, and distance educa-tion. In its simplest configuration of distributed multimedia system, the architecture will comprise of multimedia information server connected to clients via networks. Clients will dial-up the server and request the retrieval of information objects (consisting of audio, video, text, imagery, animation, etc.) stored at the server. Distributed multimedia

systems require continuous data transfer over relatively long periods of time, media synchronization, large storage, special indexing and retrieval techniques.

2. Transcoding

Among all the abovementioned data types, the amount of digital video that is available has increased dramatically in the last few years. Several image data compression methods and standards (JPEG, M-JPEG, H.261, MPEG-..)were created during the last decades for effective manipulation of the huge amount of such data. The several digital video sources involve several image formats and compression standards, therefore it is necessary to solve the conversion of any format and standard into any other (Fig. 1.). This conversion is named transcoding procedure. Naturally, this conversion is very complex problem and general solution not exists at present. In this presentation, we will concentrate on some partial solution of this field.

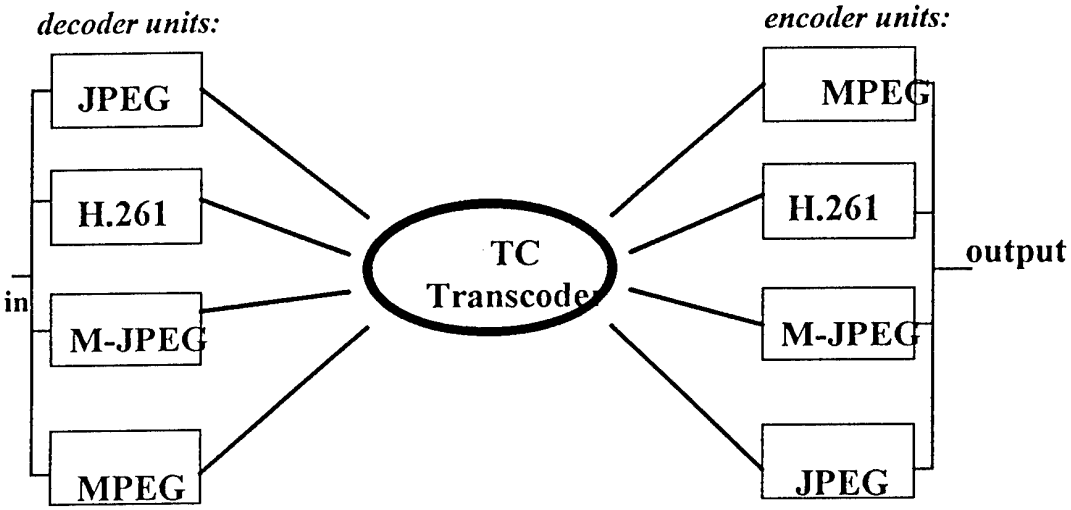


Fig. 1 Block scheme of transcoder method

In our work, the main purpose is to determine the design and implementation aspects of the transcoder unit. The abovementioned standards involve the following processing units: DCT, quantizer, motion-compemsation, lossless coding, etc. Previously, we developed simulational software (DCT, wavelet, DPCM, blockmatching motion-compensation, etc.) for the investigation of these units. At present, we are developing simulational software for the investigation of the transcoding procedures using some earlier results. Our main task is the development of an educational distributed multimedia system and also an healthcare distributed multimedia system in which the clients are multimedia PCs. The transcoding is a critical part involved into this work.

inverse quantizer and an inverse DCT.

Most of hardware-based JPEG codecs support only the 4:2:2 format decimated video, and the H.261 uses the 4:1:1 format decimated video, one possible solution of transcoding is the subsampling of the JPEG chrominance components vertically by two.

3. Implementation aspects

Naturally for implementation we need high-speed compression tools, or codecs (compression/ decompression devices). Software solutions - even in the case when running on parallel platform - offer poor results not satisfying most of the real-time requirement. Hardware solutions for accepted standards are provided by specialised inexpensive VLSI codec chips or chip families. However, due to their specialised functions, they do not provide the flexibility that may be highly desirable for advanced, much demanding applications with various or novel compression methods or just for testing the different compression algorithms. Instead of the rigid structure of the specialised codecs flexible dynamically (re)configurable fast hardware structures would be favoured in these applications.

For low-level image processing, motion analysis and image coding purposes the methods and techniques usually include a few processing steps or signal processing operations/functions, (like convolution/filtration, discrete transform, and inverse transform, effective coding methods, special effect calculations, format transforms, clustering pixels, vector quantization operations, etc.). These steps have to be chained or pipelined after each other, while the digital image data flow through the chain

The transcoding procedure is more complicated in the case of other standard pairs (e.g. JPEG - MPEG).

In the first phase of our work, we concentrate on the solving of transcoding procedure involving only the JPEG - H.261 pair encoding algorithm.

and thus the image gets transformed. In transcoding operation the type of functioning is similar at both sides, encoding and decoding.

The available and future VLSI FPGAs (Field Programmable Logic Gate Arrays) are well suited to form add-on parallel coprocessor systems for workstations, PCs, video/graphic tools. Their internal logic structure may be defined/configured (tuned to the needs of the algorithms) via downloading the configuration data (pattern) to their internal configuration RAMs. Novel developments in the fields of FPGA chips consider signal/image processing needs, offering very flexible fine-grain internal structure for the chips. For development purposes and also simulation of the behaviour of the intended internal logic convenient sophisticated software tools are available to support the configuration design [8] . Since the configuration and its operation are defined by the configuration pattern, downloaded and stored in internal RAMs, the structure can be dynamically (during run-time) changed or reconfigured, simple via downloading a new pattern. Novel

We can investigate the general problem of transcoding procedure only in limited extent such as a given matched decoder/ encoder pair. The main bottleneck in this problem is the large number of operational steps such as DCT computations and memory operations. From this point of view, first the relative simpler versions such as JPEG/H.261 and M-JPEG/H.261 pairs are under investigation using simulations. Later, it will be involved also the MPEG standard which is the most perspective general standard. Most timeconsuming operation is the processing of transform coefficients, therefore the separable version of the 2D DCT will be applied. Frame memory is needed for the intermediate processing.

In general, many 2D DCT algorithms have been proposed to reduce the computational complexity and to increase the operational speed. These algorithms can be divided into two groups, the row-column method and the direct 2D method. The row-column method computes the 2D DCT by applying the 1D DCT on the rows of the input image data frames, storing the transformed results in an intermediate matrix, transposing the matrix, and performing the 1D DCT again on the columns of the transposed matrix. Since there exist many 1D DCT algorithms, there are also many realizations for the row-column method.

In the case of JPEG, the input image is first divided into nonoverlapping blocks. Each block has 8 x 8 pels. Each block is transformed into the frequency domain by DCT. The standard does not specify a unique DCT algorithm. Consequently, users may choose the algorithm that is best suited for their

The details of this work are image format or resolution conversion, chrominance signal subsampling, intermediate processing for decoded DCT components, processing of intra- and inter-blocks of GOP supposing H.261 coding etc. The received input signal first will be decoded, then its components will be mapped into intermediate format in the transcoder unit, and finally the "new" coded signal will be created from the intermediate format. One of the most

applications. The DCT coefficients are quantized and entropy coded.

In the JPEG decoder, after extracting the coding and the quantization tables from the compressed bit stream, the compressed data passes through an entropy decoder. The DCT coefficients are first dequantized and then translated to the spatial domain via an inverse DCT. After a block-to-raster translation, the image is fully decoded.

The H.261 encoding algorithm - like the MPEG - uses a combination of DCT coding and differencial coding. The main elements of a H.261 encoder are frame prediction, DCT transformation, quantization of transform coefficients and variable length coding. The DCT coding path is similar to the one used in JPEG and MPEG. Similarly to the operations in JPEG, the DCT operates on 8 x 8 picture blocks. Four luminance (Y) blocks and one B - Y and one R - Y colour difference block are combined to form a macroblock.

In the H.261 decoder, the compressed input is buffered and processed by the variable length decoder. The decoded data are parsed and then processed by an

FPGAs allow (e.g. ATMEL AT6OOO) allow partial reconfiguration, to speed up the process. A single cell operation may be changed in 2OO ns.

The use and development of reconfigurable structures for sophisticated high-speed multimedia and image/signal processing applications is an emerging new research area. More and more efforts are concentrated to questions like what are the application requirements, how they can be solved using reconfiguration, how the task partitioning and then control pattern downloading affect the computational overhead, what performance measurement/ evaluation consideration should be, etc. As for up to date the early results and publications, however, agree in that this novel field of high-speed sophisticated systems is a promising one.

At present, the development of FPGA units is going on based on simulational results, and parallel with this, we are working on the evaluation method of the new units.

4. Conclusion

Our work is involved in the development of an educational multimedia system based on the server-client model. The work is in initial phase, first simulational software were developed for verifying our algorithms.

References

[1] ACM'95 Proceedings

[2] Borko Furcht, Stephen W. Smoliar, HongJiang Zhang,
Video and Image Processing in Multimedia Systems
Kluwer Academic Publishers 1995

[3] D. Minoli, R. Keinath,
Distributed Multimedia Through Broadband Communication Services
Artech House 1994.

[4] ITU-T Recommendation H.261: Video codec for audiovisual services at p x 64 kbit/s

[5] B. Furht, M. Milenkovic,
A Guided Tour of Multimedia Systems and Applications
IEEE Computer Society Press, Los Alamitos, California, 1995.

[6] M-JPEG Option, Handbuch
FAST - Fast Electronis GmbH, 1994

[7] Digitalvideo Computing GmbH: MPEG Development Kit Reference Manual 2.0, 1995.

[8] SPECTRUM (TM) Reconfigurable computing platform. Giga Operation Corp. Ca. USA

HUMAN FACE RECOGNITION: AUTOMATIC FACE DETECTION

G. Marcone, A. Fusi*, G. Stoppani* and G. Orlandi***

* Fondazione Ugo Bordoni, Roma, Italy
** University of Rome "La Sapienza", Italy
Tel. : +396 5480 2135; Fax: +396 5480 4401
e-mail: gmarcone@fub.it

ABSTRACT

In this paper a method to automatically locate the head of an individual in a generic image is proposed. It is based on the hypothesis that the outline of a human head can be seen as an elliptical structure. The proposed method exhaustively evaluates all the possible ellipses associate to the edge map information and selects the ellipse that best fits the head depicted in the image. The results demonstrate the robustness of the method to variation of lighting, tilt and translations in the image plane.

1. INTRODUCTION

In recent years, the problem of human face identification has attracted considerable attention because of many applications of automatic face recognition systems (to control the access of security buildings, to enhance the security of the user authentication in ATMs, to improve the information security, etc...). However, the face recognition research [1] has been mainly focused on distinguishing an input face image from a database of known face images, while the task of detecting faces in an arbitrary background is usually carried out by either hand segmenting or capturing faces against a known uniform background.

Face detection has direct relevance in the face recognition problem. In particular, identifying and locating faces in an unknown image is the first important step to implement a fully automatic human face recognizer. Moreover, face detection has potential applications in human-computer interfaces and surveillance systems. A face finder can make workstations with cameras more user friendly by turning monitors on and keeping them active whenever there is someone in front of the camera.

The problem of detecting a human face within digitized images consists of determining whether or not there is a face in an arbitrary image, and in the affirmative case, of segementing the face region from non-face regions, returning the location and the spatial extent of the face in the image plane.

There are various approaches to the problem of face detection: fixed templates approach [3]; deformable templates approach [4] and the use of the spatial image invariance. In the first approach the difference measurement between a fixed reference pattern, or a bank of reference sub-patterns, and candidate image locations is computed; then the output is threshold for matches. In the second approach a deformable template is fitted to different parts of the image and the output is then thresholded for matches. The last approach is based on a set of spatial image relationships common to all the face pattern, and a check for positive occurrences of these invariance at all candidate locations is performed.

In this work a computational methodology for detecting human faces in digitized images has been developed. It is inspired to the works [5] and [6] and uses an a-priori information: the roughly hypothesis, but quite verified in practical, that the outline of a human head can be seen as an elliptical structure. It can be roughly classified as a deformable template method because it is able to distinguish the face region by using a parametrized elliptical template. This template marks the boundary between the face region and the background.

The edge map of the image is processed to get an high level description of the scene depicted in the image. This description is based on the main edge lines of the objects contained in the scene. The edge lines are then subdivided into edge segments. These segments are paired with other edge segments and fitted to a linearized equation of the ellipse. The parameters of the ellipse (centre point, semi-axes) are found by solving a 4x4 system of linear equations. After all possible pairs of segments are considered, a set of corresponding ellipses is obtained. For each ellipse, the edge curves that intercept it are grouped, according to a suitable procedure. Then a classification task of the fitting ellipse set is carried out, according to a function that measures the fitness between the generic ellipse and the corresponding edge curves.

Further a selection procedure has been implemented to form the best fitting ellipse set with the corresponding edge curve sets.
The best fitting ellipse is calculated which is the average ellipse of the best fitting ellipse set. A second ellipse is calculated resolving a linearized over-determined system of equations, that is obtained considering all the points of the edge curves corresponding to the ellipses in the best fitting set.
Finally, a cost function based on measurements of the differences between the two calculated ellipse is defined to evaluate the quality of the detection task. This global parameter is suitable for an automatic procedure.
The paper is organized as follows: Section 2 reports a detailed description of the used detection procedure. Section 3 deals with the analysis of the results. Finally Section 4 describes several experimental results.

2. FACE IMAGE DETECTION ALGORITHM

The detection algorithm is based on the edge map information of the input face images.
The edge map is a low level description of the scene depicted in a generic input image. It contains information about the location, size and shape of the objects in the image. This information is used to accomplish the subsequent segmentation task, that usually separates objects of interest from the rest of the image content.
In the image class considered in this work, the object of interest is the face of an individual depicted in the foreground. The corresponding edge map contains the outline and the features of the face. Further, the edge map contains also the edge information of objects that belong to the background. This information may give rise to misunderstanding in the segmentation task. Depending on the application the background may be uniform (without objects) or non-uniform (with objects of various sizes and shapes).

Figure 1. Intensity face image and its edge map from Canny's edge detector

In this work, the Canny's algorithm [7] has been implemented to get the edge map of face images with non-uniform background. This algorithm has been optimized to evidence the head outline, discarding the edges corresponding to features not marked, beard, grizzled hair, etc..
Figure 1 shows an example of a generic input image and its Canny edge map.

2.1 EDGE MAP PROCESSING

The edge map is not able by itself to give a description based on the main lines of the objects contained in the image. It is needed to accomplish further processing to get the high level description suitable to the segmentation of face from the rest of the image.
The first processing task consists of the removal of the intersection points. These points in the edge map occur in relation to occlusion of different objects. It is needed to remove these points in order to preserve the integrity of the edge lines belonging to the object in the foreground. The removal procedure used in this work is inspired to [5] and consists of disjoining the intersection of two edge segments not belonging to the same object. The removal task is carried out by multiplicating (logic and) the intensity of the points in the neighbourhood of an intersection point with a proper binary mask.
The second processing step consists of researching and following the main lines of the object in the image.
The line following algorithm marks all the contiguous points that belong to the same line. A generic line ends when there are not any more contiguous points not-marked. In the case of line with thickness more than one pixel, the algorithm extracts from it all the possible lines of one pixel thick. The result of the algorithm is a set of *lines* or *curves* **c** of 1 pixel thick. In the following this set will be referenced as *curve set* **C**.
The third processing step consists of cutting each curve **c** of **C** into smaller and linear components (i.e. edge segments). This task may be seen as an approximation of a curve with a polyline. The result of the cutting procedure is a number of segments **s** corresponding to the curve **c**. All the segments **s** obtained for each curve **c** of **C** form the *segment set* **S**.

2.2 FITTING PROCEDURE

The ellipse fitting procedure take a pair of edge segments (s_i, s_j) in **S** and tries to fit them into a linearized equation of the ellipse. The parameters of the ellipse are found by resolving a 4x4 system of linear equations according to the following procedure.
Starting from the equation of the ellipse:

$$\frac{(x-x_0)^2}{a^2}+\frac{(y-y_0)^2}{b^2}=1 \qquad (1)$$

where (x_0,y_0) is the ellipse centre, **a** the horizontal semi-axe and the **b** vertical semi-axe. The equation (1) can be rearranged in :

$$2xa_0 - y^2 a_1 + 2ya_2 - a_3 = x^2 \qquad (2)$$

where:

$$a_0 = x_0 \ , \ a_1 = \frac{a^2}{b^2} \ , \ a_2 = \frac{a^2}{b^2} y_0 \ , \ a_3 = x_0^2 + \frac{a^2}{b^2} y_0^2 - a^2 .$$

Substituting the co-ordinates of the extremal points (x_1,y_1) (x_2,y_2) and (x_3,y_3)(x_4,y_4) respectively of the segment s_i and s_j in the equation (2), it is possible to form the following system of equations:

$$\begin{pmatrix} 2x_1 & -y_1^2 & 2y_1 & -1 \\ 2x_2 & -y_2^2 & 2y_2 & -1 \\ 2x_3 & -y_3^2 & 2y_3 & -1 \\ 2x_4 & -y_4^2 & 2y_4 & -1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \\ x_4^2 \end{pmatrix}$$

After finding (a1,a2,a3,a4), the parameters of the ellipse can be obtained according to the following formulas:

$$x_0 = a_0 \qquad y_0 = \frac{a_2}{a_1} \qquad (3)$$

$$\mathbf{a} = \sqrt{\left(-a_3 + a_0^2 + \frac{a_2^2}{a_1}\right)} \qquad \mathbf{b} = \sqrt{\left(\left(-a_3 + a_0^2 + \frac{a_2^2}{a_1}\right)\frac{1}{a_1}\right)}$$

The calculated parameters (3) are selected to describe a possible candidate ellipse when the centre, **a** and **b** are within the image plane, and the ratio **b/a** is bounded within a range of allowable values for most faces.
After all possible pairs of edge segments (s_i, s_j) in the *segment set* **S** are considered, a new set **E** of corresponding ellipses e_{ij} is obtained. For exposition convenience in the following a generic ellipse e_{ij} will be referenced as **e**.

2.3 GROUPING PROCEDURE

For each ellipse **e** in the *fitting ellipse set* **E** the edge curves **c** that fall within a particular area around the ellipse are grouped, according to the following procedure:

- for each ellipse **e** with parameters (x_0,y_0,a,b), two ellipses e^{in} and e^{out} are defined. The first one with parameters $(x_0,y_0,a(1-\chi),b(1-\chi))$ is internal to **e**, while the second one with parameters $(x_0,y_0,a(1+\chi),b(1+\chi))$ is external to **e**. These two ellipses have the same centre of **e** and fix the boundaries of a particular area **A** around the ellipse **e**.
- for each curve **c** of **C**, the number of points that fall within the area **A** are calculated. If these points are more than 70% of the overall curve points, the curve **c** is assumed as a *component curve* for the ellipse **e**.

Figure 2 gives a detailed explanation of the above procedure. At the end of the grouping procedure for each ellipse **e** there is a corresponding *component curve set* $C^e=(c_k^e;k=1,..,N_e)$ with N_e *component curves*.

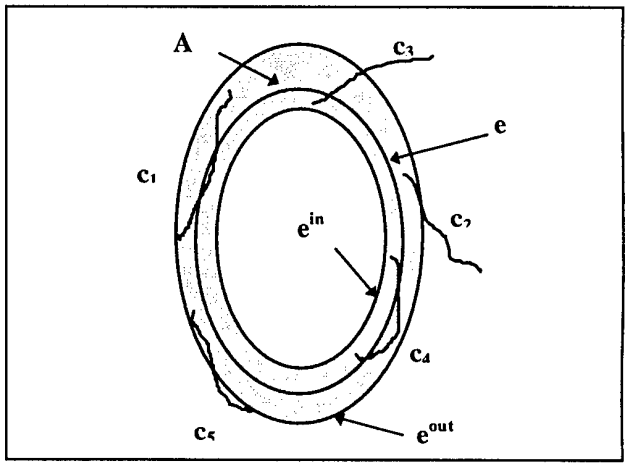


Figure 2. Ellipse **e** and its region **A** delimited by the internal and external ellipses respectively e^{in} and e^{out}. In this case the *component curve* set C^e is formed by the curves c_1, c_4 and c_5.

2.4 ORDERING PROCEDURE

At this point it is possible to classify the fitting ellipse set **E** according to a voting method. For this purpose a function **f(e)** that measures the fitness of the ellipse **e** to the edge curves is defined. The value of **f(e)** is assumed to be the sum of the lengths of the *component curves* in C^e. According to the values of **f(e)** the ellipse set **E** is ordered, obtaining the *ordered fitting ellipse set* E^{ord} .

2.5 SELECTION PROCEDURE

From the above sets, it is now possible to define a suitable subset of ellipses E^{best}, which is the *best fitting ellipse set.* This subset is formed according to the following steps:

- step 1) select the first ellipse **e** of E^{ord} with parameters (x_0,y_0,a,b).
- step 2) calculate a view centre point (x_c,y_c). In the case of the first ellipse the *view centre* is the ellipse centre (x_0,y_0). In the case of more ellipses the view

centre is the mass centre of the ellipses considered (see Figure3a).

- step 3) calculate the coverage area. Starting from the view centre point; for each *component curve* $\mathbf{c_k^e}$ the angle α_k is determined (see Figure 3b), then the sum $\zeta = \sum \alpha_k$ is performed, discarding the overlapping areas.
- step 4) test if ζ is more than a threshold τ (a percentage of the round angle). In affirmative case go to next step, otherwise take into account the *component curves* of a further ellipse, selecting the subsequent ellipse in $\mathbf{E^{ord}}$.
- step 5) define the best fitting ellipse set. The set $\mathbf{E^{best}}$ consists of **M** ellipses with *component curves* that give rise to coverage area of 60% of the round angle.

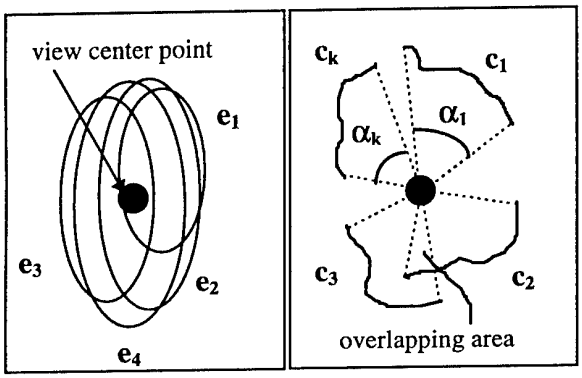


Figure 3. Two examples are reported: a) the view centre point (x_c,y_c) for the ellipses $\mathbf{e_1}$, $\mathbf{e_2}$, $\mathbf{e_3}$, $\mathbf{e_4}$; b) the angle α_1, α_2, α_3, α_k corresponding to the component curves $\mathbf{c_1}$, $\mathbf{c_2}$, $\mathbf{c_3}$, $\mathbf{c_k}$.

3. THE BEST FITTING ELLIPSE

The solution to the detection problem is to calculate the ellipse ε_1, which is the average of the **M** ellipses of $\mathbf{E^{best}}$, according to:

$$(\bar{x}_0, \bar{y}_0, \bar{a}, \bar{b}) = \frac{1}{M}\sum_{m=1}^{M}(x_0^m, y_0^m, a^m, b^m)$$

In order to evaluate the quality of the detection task, it is important to define a global function, which can be suitable for an automatic procedure. For this purpose a second ellipse ε_2 is calculated, considering the overall points of the $\mathbf{N_e}$ component curves of the **M** ellipses of $\mathbf{E^{best}}$. The co-ordinates of these points are substituted in the equation (2) obtaining the following over-determined system of equations:

$$\begin{pmatrix} 2x_1 & -y_1^2 & 2y_1 & -1 \\ 2x_2 & -y_2^2 & 2y_2 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 2x_N & -y_N^2 & 2y_N & -1 \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_N^2 \end{pmatrix} \qquad (4)$$

This system is of the form **AX= C**, where **A** is Nx4, X is 4x1 and C is Nx1. It can be solved by using the pseudo inverse method: $\mathbf{X} = (\mathbf{A^tA})^{-1}\mathbf{A^tC}$.
In the case of a curve belongs to different component curve set $\mathbf{C^e}$, its points will be considered more times in the system (4).

Further a global quality parameter $S_m = \frac{1}{128}\sum_{k=0}^{127} T_r(k)$ has been introduced, which is the average value of the *similitude trace* function $\mathbf{T_r}$. This function is based on measurements of the differences between the two ellipses and is calculated as in the following:

- calculate the average centre (x_m,y_m) of the two ellipses ε_1 and ε_2
- define a semi-axe $\mathbf{v_k}$ with origin in (x_m,y_m) and orientation $\phi_k = k*2\pi/128$, k=0,127.
- calculate the intersection points (x_{1k},y_{1k}) (x_{2k},y_{2k}) between $\mathbf{v_k}$ and the two ellipses ε_1 and ε_2

calculate the function $T_r(k)$:

$$T_r(k) = \sqrt{(x_{1k}-x_{2k})^2 + (y_{1k}-y_{2k})^2}$$

For more details see Figure 4.

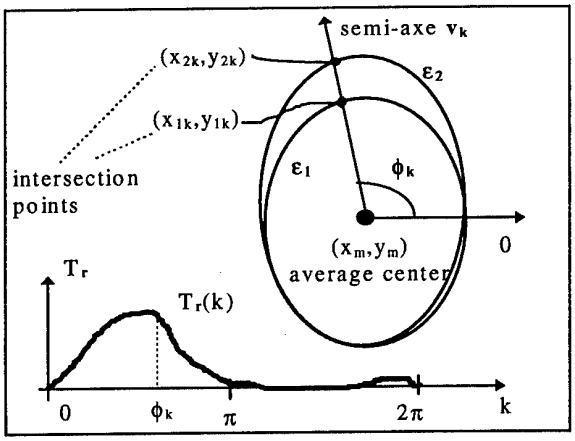


Figure 4. The similitude trace T_r of the two ellipses ε_1 and ε_2 is drawn for 0<k<127. For a given semi-axe $\mathbf{v_k}$ the intersection points (x_{1k},y_{1k}) (x_{2k},y_{2k}) and the value of $T_r(k)$ are outlined.

The similitude trace function $T_r(k)$ provides useful information about the quality of the detection task. For low values of S_m the two ellipses ε_1 and ε_2 give concordant results in terms of face detection. So it is possible to

assume that the detection task is of good quality (see the S_m values in Figure 5a and 5b).
This information could be useful used in a fully automatic segmentation system: if the detection task has a global quality parameter Sm less than a given threshold, the subsequent segmentation task may be carried out.

4. EXPERIMENTAL RESULTS

The proposed segmentation method has been tested on the face image databases of the M.I.T. consisting of 128x120 pixel images with variable imaging conditions (moderately cluttered background, variable lighting, etc.). In figure 6 some results of the segmentation procedure are reported. It is possible to see from the images in the 1st row that the segmentation algorithm is robust to lighting variations. From the images in the 2nd and 3rd row the algorithm is proved to be further independent to head tilting. The algorithm is also independent to the location of the head in the image plane as it is shown in the 4th row. Finally, it features good performance also in the case of noisy face images.

5. CONCLUSIONS

In this work a detection algorithm for automatic face detection has been implemented. It differs from similar methods in the computing methodology to get ellipses that best fit the head outline. In particular 4x4 systems of linear equations based on edge segments have been used obtaining a wider ellipse fitting set. This set has been ordered according to a measure of the fitness. Then a best fit ellipse set has been formed from the previous set. Two ellipses have been obtained from the last ellipse set. Finally a cost function that measures the quality of the detection task has been introduced. This function is suitable for an automatic procedure. The method gives good results in terms of face detection with different experimental conditions. Further work have to be done in order to generalise the method, considering also ellipses with semi-axes orientation different from the Cartesian axes. The open problem is to evaluate the behaviour of the head detection algorithm when it is interfaced with the subsequent recognition task in a fully automatic recognition system.

6. REFERENCES

[1] R. Chellappa, S. Sirohey, C.L. Wilson, C.S. Barnes., "Human and Machine Recognition of Faces: A Survey", *Computer Vision Laboratory; Center for Automation Research, University of Maryland,* CAR-TR-73,. August 1994.

[2] K. K. Sung, T. Poggio, "Example-based Learning for View-based Human Face Detection", *Massachusetts Institute of Technology Artificial Intelligence Laboratory,* A.I. Memo No. 1521, December 1994.

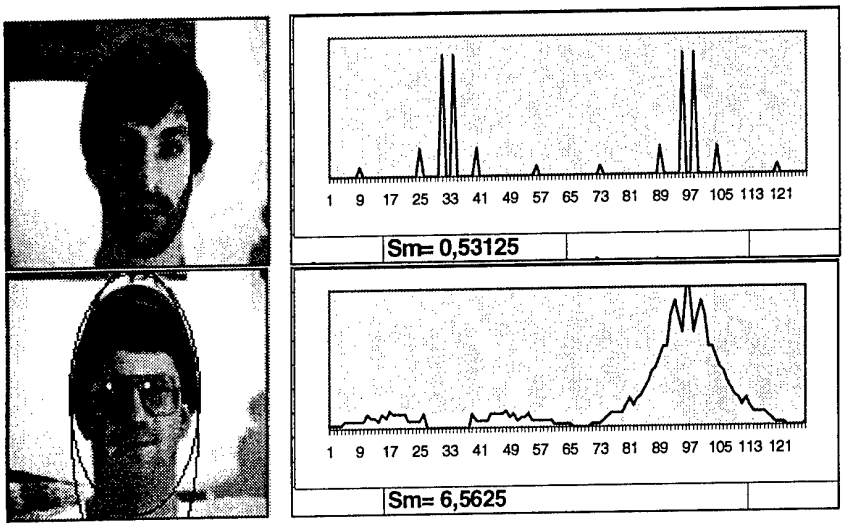
[3] R. Brunelli, T. Poggio, "Face Recognition: Feature Versus Templates", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1993, vol.15, No. 1, pp.1042-1052.

[4] A. Yuille, D. Cohen. and P. Hallinan, "Feature Extraction From Faces Using Deformable Templates", *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* pp. 104-109. 1989.

[5] A. Eleftheriadis, A. Jacquin, "Automatic Face Location Detection for Model-assisted Rate Control in H.261-compatible Coding of Video", *Signal Processing: Image Communication,* November 1995, Vol. 15, pp. 435-455.

[6] S. A. Sirohey, "Human Face Segmentation and Identification", *Computer Vision Laboratory; Center for Automation Research, University of Marylan,* CAR-TR-695. November 1993.

[7] J. Canny, "A Computational Approach to edge Detection", *IEEE Transactionson on Pattern Analysis and Machine Intelligence,* 1986, vol.8, pp.679-689.



The quality of the detection task in the case of figure5a is better than that one of figure 5b, this is comfirmed by the values of the global quality parameter Sm reported in figure 5a and 5b

Figure 5.Two face images with the two calculated ellipses and the corresponding behaviors of the similitude trace function Tr(k).

face detection with variable lighting

face detection with variable lighting and right tilt

face detection with variable lighting and left tilt

face detection with different position

face detection with different gaussian noise

Figure 5. Results of the segmentation algorithm in different experimental conditions.

Crowd Motion Estimation Using Invertible Rapid Transform

*J. Turán - *K. Fazekas - J. Gamec - L. Kövesi*

Department of Radioelectronics
Technical University of Košice
Park Komenského 13
04021 Košice
Slovakia
Tel./Fax: +42 95 6335692
E-mail: J.TURAN@CCSUN.TUKE.SK

**Department of Microwave Telecommunications*
Technical University of Budapest
Goldmann Tér 3
1111 Budapest
Hungary
Tel./Fax: +36 12 043289
E-mail: T-FAZEKAS@NOV.MHT.BME.HU

Abstract

This article presents a novel technique of crowd motion estimation using invertible rapid transform (**IRT**). The new method was used to estimate crowd motion from the image data sequences captured at railway station in large cities.

Keywords
Crowd Motion Estimation, Rapid Transform, Invertible Rapid Transform, Motion Estimation

Introduction

The understanding of crowd behaviour in semi-confined spaces is an important part of the design of new pedestrian facilities and major layout modifications to existing areas and, for the daily management of crowds at football matches, pop concerts, carnivals, airports and even in the day to day movement of commuters in and out a large cities is a substantial problem with serious consequences for human life and safety for public order if it is not managed successfully [1, 2].

Human observers of crowds particularly those experienced in the management of crowds in public places, can detect many crowd features, in some cases quite easily. Normally they can distinguish between a moving and a stationary crowd and estimate the majority direction and speed of movement of a large crowd. For facilities already in existence, there is an established practice of using extensive closed circuit television monitoring of crowds. Human observers normally positioned to watch the TV monitors of a such systems are not sufficient for obtain real time data by watching recorded video sequences. There is thus a considerable benefit from being able to develop methods for automatically collecting crowd description data by use of image processing techniques applied to the video sequences [3, 4, 5]. These methods are based on well established image processing techniques and are able to monitoring and collecting data about key features of crowds: stationarity, density and motion.

Particularly estimation of crowd motion is based on two well known methods: optical flow and block matching motion detection [3, 4]. The motion vectors calculated by these methods may be used to devise a polar plot (showing velocity magnitude and direction) for a moving crowd, where the dominant motion tendency of a crowd can be seen. Unfortunately these methods are rather complicated and not suitable for straightforward real time implementation.

This article presents a novel technique of crowd motion estimation using invertible rapid transform (**IRT**) [5]. The new method was used to estimate crowd motion from the image data sequences captured at railway station in large cities.

2. Invertible rapid transform

The rapid transform (**RT**) [6] is a fast shift invariant transform well known in the field of pattern recognition. The **RT** has some interesting properties such as invariance to cyclic shifts, to reflections of the data sequence, and to slight rotations of a two-dimensional pattern [6, 7]. The **RT** is applicable to both binary and analogue inputs and can be extended to multiple dimensions [7]. Because of the recursive nature

of the calculations and the use of very simple operators it can be easily implemented in both software and dedicated digital hardware [8]. The **RT** was used in recognition of alphanumeric characters [6, 9, 10], robotics and scene analysis [10, 11]. Even though the **RT** is a nonlinear and thus a noninvertible transform it is possible to recover the original signal **x** from its rapid transform sequence **RT{x}**, computing additional data (known as matrix of states **K** for **1D-RT** or matrices of states $\mathbf{K}_p^{(r)}$ for **2D-RT**), i.e. the invertible rapid transform (**IRT**) can be defined [12, 13]. The **IRT** may be used for signal coding, motion estimation and nonlinear filtering [14].

Signal flow graph [12, 13] for compute of the **1D-IRT** is shown in Fig.1.

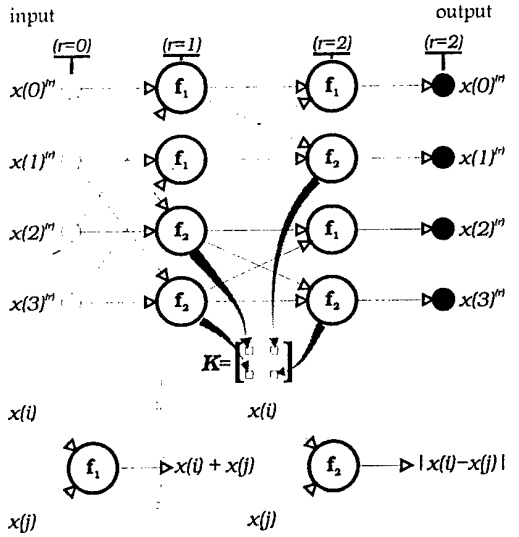


Fig.1. Signal flow graph of the **1D IRT**

The matrix of states $\boldsymbol{K}$ or system matrices of states $\boldsymbol{K}_p^{(r)}$ may be computed as follows. For one-dimensional case:

$$k(i,r)=0,\ \text{if}\ \ x(i)^{(r)} - x(i+N/2)^{(r)} < 0$$

$$k(i,r)=1,\ \text{if}\ \ x(i)^{(r)} - x(i+N/2)^{(r)} \geq 0. \quad (1)$$

The dimension of matrix $\boldsymbol{K}$ is n x $N/2$. For two dimensional case:

$$k_1^{(r)}=1,\ \text{if}\ x^{(r)}(i,j)-x^{(r)}(i+N/2,j)\geq 0$$

$$k_1^{(r)}=0,\ \text{if}\ x^{(r)}(i,j)-x^{(r)}(i+N/2,j)<0$$

$$k_2^{(r)}=1,\ \text{if}\ x^{(r)}(i,j+N/2)-x^{(r)}(i+N/2,j+N/2)\geq 0$$

$$k_2^{(r)}=0,\ \text{if}\ x^{(r)}(i,j+N/2)-x^{(r)}(i+N/2,j+N/2)<0$$

$$k_3^{(r)}(i,j)=1,\ \text{if}$$
$$\left|x^{(r)}(i,j)+x^{(r)}(i+N/2,j)\right|-\left|x^{(r)}(i,j+N/2)+x^{(r)}(i+N/2,j+N/2)\right|\geq 0$$

$$k_3^{(r)}(i,j)=0,\ \text{if}$$
$$\left|x^{(r)}(i,j)+x^{(r)}(i+N/2,j)\right|-\left|x^{(r)}(i,j+N/2)+x^{(r)}(i+N/2,j+N/2)\right|<0$$

$$k_4^{(r)}(i,j)=1,\ \text{if}$$
$$\left|x^{(r)}(i,j)-x^{(r)}(i+N/2,j)\right|-\left|x^{(r)}(i,j+N/2)-x^{(r)}(i+N/2,j+N/2)\right|\geq 0$$

$$k_4^{(r)}(i,j)=0,\ \text{if}$$
$$\left|x^{(r)}(i,j)-x^{(r)}(i+N/2,j)\right|-\left|x^{(r)}(i,j+N/2)-x^{(r)}(i+N/2,j+N/2)\right|<0 \quad (2)$$

where *(r)* is transform step of **IRT** and

$i,j = 0,1, ..., (N/2-1)$

$r = 1, 2, ..., n$

$p = 1, 2, 3, 4$

The system of matrices of states $\boldsymbol{K}_p^{(r)}$ is illustrated in Fig.2.

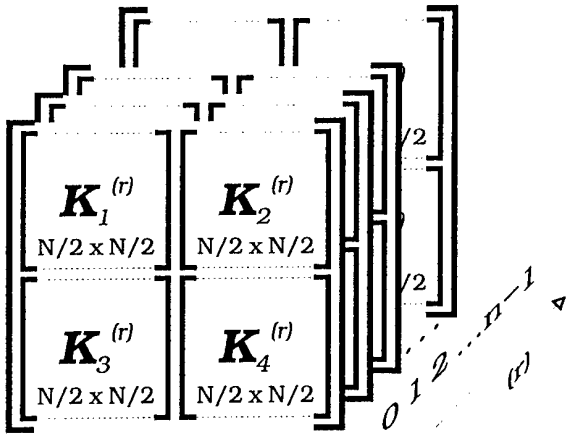


Fig.2. The system of matrices of states $\boldsymbol{K}_p^{(r)}$

3. Motion estimation algorithm with use of IRT

Motion estimation algorithms are based on presumption that in matrix $\boldsymbol{K}$ or in system $\boldsymbol{K}_p^{(r)}$ is included relevant information about the picture [14, 15, 16] and the motion in picture influence the first column of $\boldsymbol{K}$ or the first set of matrices of $\boldsymbol{K}_p^{(r)}$ (i.e. $\boldsymbol{K}_1^{(0)}, \boldsymbol{K}_2^{(0)}, \boldsymbol{K}_3^{(0)}, \boldsymbol{K}_4^{(0)}$) in maximal way. Cyclical translations in the image are deterministically and unambiguously encoded to the values of this matrices. The matrices $\boldsymbol{K}$ or $\boldsymbol{K}_p^{(r)}$ are binary matrices and can be computed

with use of simple and thus very fast algorithm using operations of comparison, addition and subtraction. This results in simplicity of the motion estimation subblock matching criterion computation, which is than based on the operations of bit by bit modulo2 additions [16].

First, the image is divided into smaller rectangular areas, which we call subblocks (Fig.3). Let U_k be an $N \times N$ size subblock of frame k and U_{k-1} be equivalent subblock of frame k-1. Let search area (**SA**) be an $(N+2d_m)\times(N+2d_m)$ size of frame k-1, centered at the same spatial location as U_k and U_{k-1} is subblock from **SA**, where d_m is the maximum displacement allowed in either direction in integer number of pixel.

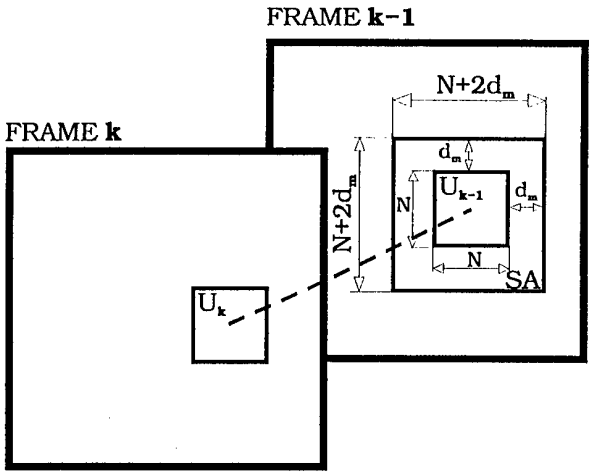


Fig. 3: Positions of subblocks U_k, U_{k-1} and **SA** at the frames k and k-1

3.1 Motion estimation algorithm with use of 1D-IRT

Let $\boldsymbol{K}_r$ and $\boldsymbol{K}_c$ are matrices of states computed by row and by column of subblock respectively.

STEP 1: Compute $\boldsymbol{K}_{r(k)}$, $\boldsymbol{K}_{c(k)}$ for subblock U_k.

STEP 2: Compute $\boldsymbol{K}_{r(k-1)}$, $\boldsymbol{K}_{c(k-1)}$ for subblock U_{k-1}.

STEP 3: Compute matching criterion

$$\sigma_{row}(u,v) = \sum_{r=0}^{N-1}\sum_{i=0}^{N/2-1}\sum_{j=0}^{\mu}\left(\boldsymbol{K}_{r(k)}(i,j) \oplus \boldsymbol{K}_{r(k-1)}(i,j)\right),$$

$$\sigma_{col}(u,v) = \sum_{r=0}^{N-1}\sum_{i=0}^{N/2-1}\sum_{j=0}^{\mu}\left(\boldsymbol{K}_{c(k)}(i,j) \oplus \boldsymbol{K}_{c(k-1)}(i,j)\right),$$

$$u,v \in \langle -d_m, d_m \rangle \qquad (3)$$

where $\oplus$ denotes bit-by-bit modulo2 addition.

Repeat steps 2, 3 for every possible positions (u, v) of subblock U_{k-1} in subblock **SA** $((2d_m+1)^2$ cycles).

STEP 4: The desired vector of motion correspond to the position (u_0, v_0) of subblock U_{k-1} with minimal value of $\sigma(u, v)$.

Modifications of the algorithm - I

μ -number of used columns of matrix $\boldsymbol{K}$

$\mu \in \{0, 1, ..., n-1\}$

Modifications of the algorithm - II

4a, $(u_0,v_0) \in \{u,v\}; \sigma_{row}(u_0,v_0) = \min(\sigma_{row}(u,v))$

4b, $(u_0,v_0) \in \{u,v\}; \sigma_{col}(u_0,v_0) = \min(\sigma_{col}(u,v))$

4c, $(u_0,v_0) \in \{u,v\}; \sigma_{row}(u_0,v_x) = \min(\sigma_{row}(u,v)) \wedge$
$\wedge \sigma_{col}(u_x,v_0) = \min(\sigma_{col}(u,v))$

4d, $(u_0,v_0) \in \{u,v\}; \sigma_{r+c}(u_0,v_0) = \min(\sigma_{r+c}(u,v)),$ (4)

where $\sigma_{r+c}(u,v) = \sigma_{row}(u,v) + \sigma_{col}(u,v)$ (5)

3.2 Motion estimation algorithm with use of 2D-IRT

STEP 1: Compute first set of $\boldsymbol{K}_p^{(r)}$ (i.e. $\boldsymbol{K}_{1(k)}^{(0)}, \boldsymbol{K}_{2(k)}^{(0)}, \boldsymbol{K}_{3(k)}^{(0)}, \boldsymbol{K}_{4(k)}^{(0)}$) for block U_k.

STEP 2: Compute first set of $\boldsymbol{K}_p^{(r)}$ (i.e. $\boldsymbol{K}_{1(k-1)}^{(0)}, \boldsymbol{K}_{2(k-1)}^{(0)}, \boldsymbol{K}_{3(k-1)}^{(0)}, \boldsymbol{K}_{4(k-1)}^{(0)}$) for block U_{k+1}.

STEP 3: Compute matching criterion

$$\sigma(u,v) = \sum_{p=1}^{\tau}\sum_{i=0}^{N/2-1}\sum_{j=0}^{N/2-1}\left(\boldsymbol{k}_{p(k)}^{(0)}(i,j) \oplus \boldsymbol{k}_{p(k-1)}^{(0)}(i,j)\right) \qquad (6)$$

$$u,v \in \langle -d_m, d_m \rangle$$

where $\oplus$ denotes bit-by-bit modulo2 addition.

Repeat steps 2, 3 for every possible positions (u,v) of subblock U_{k-1} in subblock **SA** $((2d_m+1)^2$ cycles).

Modifications of the algorithm

3a, $\tau = 1$

3b, $\tau = 2$

3c, $\tau = 3$

3d, $\tau = 4$

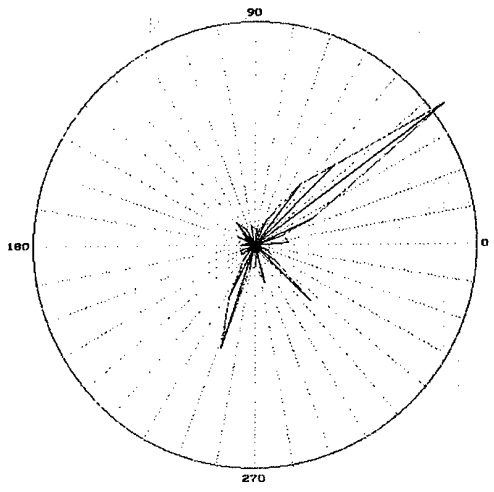
Frame 1

Frame 3

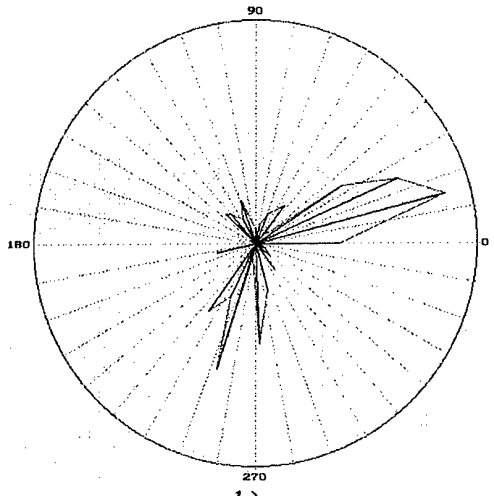
Frame 2

Frame 4

Fig.4 The image sequence



a)



b)

Fig.5 Polar plot v-s: a) Frame 1-2, b) Frame 3-4

STEP 4: The desired vector of motion corresponds to the position (u_0, v_0) of subblock U_{k-1} with minimal value of $\sigma(u, v)$, t.e.

$$(u_0, v_0) \in \{u, v\}; \sigma(u_0, v_0) = \min(\sigma(u, v)). \quad (7)$$

4. Crowd motion estimation

The new method of crowd motion estimation was used to estimate crowd motion from the image data sequences captured at railway stations in large cities. For subsequent image frames (Fig. 4) the subblocks displacement vectors using **IRT** was calculated. Then the vectors was used to devise a polar plot (showing velocity magnitude v and direction s) for moving crowd, with use of their aggregation to discrete direction 'bins' and with various bin size (Fig. 5). From these polar histograms the dominant motion tendency of the motion crowd may be clearly identificated. Irregular motion, movements of arms, legs and clothing and localized variations of brightness all cause errors in the computed motion vectors compared to the actual overall motion of the individuals in the crowd. This effect can be easily removed (filtered) from these polar diagrams and thus improved human and machine interpretation of crowd motion is achieved. The proposed experiments indicate that **IRT** motion estimation gives good results in terms of computation cost, speed and motion estimation accuracy.

5. Conclusions

This paper has shown that it is possible to use **IRT** for estimation of crowd motion. The method discussed is amenable to using simple operations suitable for real-time implementation.

References

[1] Davies, A. C. - Yin, J. H. - Velastin, S. A.: Crowd Monitoring Using Image Processing. Proceedings IEE, 1995, 105-111.

[2] Bartolini, F. - Cappellini, V. - Mecocci, A.: Counting People Getting in and out of a Bus by Real-time Image-sequence Processing. Image and Vision Computing, Vol. 12, No. 1, 1994, 36-41.

[3] Velastin, S. A. - Yin, J. H. - Davies, A. C., Vincencio-Silva, M. A. - Allsop, R. E. - Penn, A.: Automated Measurement of Crowd Density and Motion Using Image Processing. 7th IEE Inf. Conf. on Road Traffic Monitoring and Control, London, 26-28 April, 1994, 127-132.

[4] Velastin, S. A. - Yin, J. H. - Vincencio-Silva, M. A. - Davies, A. C. - Allsop, R. E. - Penn, A.: Image Processing Techniques for On-line Analysis of Crowds in Public Transport Areas. IFAC Symp. on Transp. Systems: Theory and Application of Advanced Technology, Tianjin, China, 24-26 August 1994.

[5] Turán, J. - Davies, A. C. - Velastin S. A.: Crowd Motion Detection Using Invertible Rapid Tranform. Proc. of 2nd Int. Conf. On Image and Signal Processing, Budapest, Nov. 8-10th, 1995, 73-75.

[6] Reitboeck, H. - Brody, T.P. : A Transformation with Invariance Under Cyclic Permutation for Application in Pattern Recognition. Inf. and Control, Vol.15, 1969, 130-154.

[7] Wagh, M. D. - Kanetkar, S.V. : A Class of Translation Invariant Transforms. IEEE Trans. on Acoustic, Speech and Signal Proc., Vol. ASSP-25, No.3, 1977, 203-205.

[8] Chmúrny, J. - Turán, J. : Processors for Technical Realization of Fast Translation Invariant Transforms. Computers and Art. Int. Vol.3, No.6,1984, 563-572.

[9] Wang, P. O. - Schiau, R. C.: Machine Recognition of Printed Chinese Characters via Transformation Algorithms. Pattern Recognition, Vol.5,1973, 303-321.

[10] Chmúrny, J. - Turán, J. : Two-dimensional Fast Translation Invariant Transforms and Their Use in Robotics. Electronic Horizon, Vol.15, No.5, 1984, 211-220.

[11] Schütte, H. - Frydrychowitz, S. - Schröder, J.: Scene Matching with Translation Invariant Transforms. Proc. 5IPCR, Miami, USA, 1980, 195-198.

[12] Turán, J. - Chmúrny, J. : Two-dimensional Inverse Rapid Transform. Computers and Art. Intelligence, Vol.2, No.5, 1983, 473-477.

[13] Fang, M.: Class of Invertible Shift Invariant Transforms. Signal Processing, Vol.23, No.4, 1991, 35-44.

[14] Turán, J. - Kövesi, L. - Kövesi, M. : CAD System for Pattern Recognition and DSP with Use of Fast Translation Invariant Transforms. Journal on Communications, Vol.XLV, 1994, 85-89.

[15] Gamec, J. - Turán, J.: Inverse Rapid Transform and Motion Analysis. In: Emergent Techniques in Digital Signal Processing, COST229, Bayona-Vigo, Spain, Oct. 19-21, 1994, 201-205.

[16] Gamec, J. - Turán, J.: Motion Estimation with Use of 1D Inverse Rapid Transform. In: Intelligent Terminals and Source and Channel Coding, COST 229, Budapest, Sept. 7-9, 1993, 213-218.

A BAYESIAN APPROACH TO THE SEGMENTATION OF FLAME IMAGES

Pedro M. Jorge *Jorge S. Marques* *Pires Barbosa*
INESC/ISEL *INESC/IST* *CPPE*

INESC, Rua Alves Redol n. 9, 1000 LISBOA
CPPE, Rua Mouzinho da Silveira n. 10, 1250 LISBOA

ABSTRACT

Flame images are a useful source of information to characterize combustion processes in industrial plants. However, the segmentation of flame images is difficult. For example, the intensity of the background is often higher than the intensity of the flame itself. This requires the use of a priori knowledge about the background and flame characteristics.

This paper presents a Bayesian approach to the segmentation of flame images. Flame boundary is estimated by the MAP method, using a probabilistic model of the image. Prior information about the contour shape is provided by a Markov random field. Robust estimation techniques are used to improve the performance in the presence of outliers. Experimental tests with industrial flame images from Setubal thermoelectric plant are provided to evaluate the performance of the algorithm showing that the proposed algorithm discriminates between flame and background.

1. INTRODUCTION

Flame images are a useful source of information about the characteristics of combustion processes in industrial plants. The use of this information in automatic monitoring/control systems requires the evaluation of the flame boundary. This is a difficult problem where classic image segmentation algorithms (e.g., thresholding, region growing [9]) fail. For example, in boilers with multiple flames, the intensity of the background is often higher than the intensity of the monitored flame since the background is affected by the radiation of all the flames and some of them can be directly observed. These difficulties require the use of a priori knowledge about the background and flame characteristics.

This paper describes a Bayesian approach to the segmentation of industrial flames based on Markov random fields [5] and robust statistics [3]. The flame boundary is evaluated by a MAP estimator. The posterior probability density function is computed by using an image formation model and a shape model (Markov random field). This approach is inspired by the work of Figueiredo and Leitão [4] on the estimation of ventricular contours. These ideas are extended here by the use of robust statistics, a key issue on the application of image analysis techniques to real data. Probabilistic models adapted to flame images are also presented.

The paper is organized as follows: Section 2 formulates flame boundary extraction as a MAP estimation problem. Section 3 and 4 present the image formation model and a probabilistic model of the flame boundary. Section 5 describes a flame segmentation algorithm based on MAP estimation. Experimental results are given in section 6 and section 7 presents the conclusions.

2. BAYESIAN APPROACH

Let x be a vector of parameters defining the boundary of a flame present in an observed image I. Assuming that x and I are random variables with known joint distribution, the *maximum a posteriori* (MAP) estimate of x given the image I is defined by

$$\hat{x}_{MAP} = \arg\max_{x}[p(x|I)] \tag{1}$$

The *a posteriori* probability $p(x|I)$ is obtained using the *Bayes law*

$$p(x|I) = \frac{p(I|x)p(x)}{p(I)} \tag{2}$$

where $p(x)$ is the *a priori* probability density function of the contour (prior), $p(I|x)$ is the conditional probability density function of the image, given the contour and $p(I)$ is a normalization factor. To compute $\hat{x}_{MAP}$ it is necessary to define the prior $p(x)$ and the likelihood function $p(I|x)$, and to optimize their product.

Figure 1 shows a typical flame image obtained in Setubal thermoelectric plant. Let us consider a set of L horizontal lines defined in the image and let x_i denote the abscissa of the intersection of the flame boundary

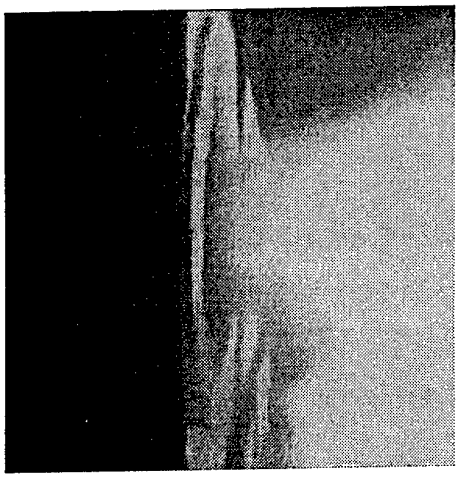


Figure 1: Original image

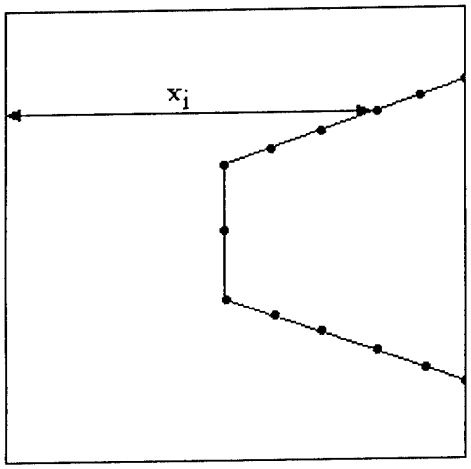


Figure 2: Boundary representation.

with the i-th horizontal line (see figure 2). The flame boundary will be described by the vector

$$x = [x_1, x_2, \ldots, x_L]^T \tag{3}$$

containing the abscissas of all intersection points.

3. IMAGE MODEL

The observed image consists of two basic components: a structured background and the flame (see figure 1). The flame region is characterized by high intensity values which saturate the camera. This may not be true near the flame boundary where image values depend on the flame and background but it is an accurate assumption inside the flame region.

Figure 3 shows the intensity profile of a row of image 1. The evolution of intensity in the interval [0,125]

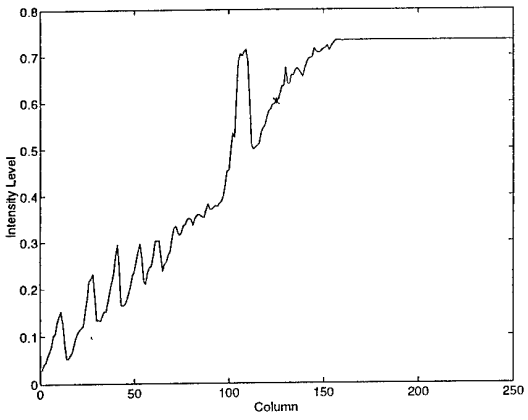


Figure 3: Intensity values of one line from the observed image.

follows the background. After column 125, intensity grows until it saturates due to the presence of the flame. Transition from background to the flame region occurs at $x_i = 125$, approximately.

We shall assume that the observed image I is the superposition of a deterministic image $\bar{I}(x)$ with a noise image W with Gaussian distribution. Let I_i, $\bar{I}_i(x_i)$ and W_i denote the i-th line of these images. Therefore

$$I_i = \bar{I}_i(x) + W_i \tag{4}$$

where $\bar{I}(x)$ is equal to the background image B, inside the background region, and equal to a constant value S in the flame region, i.e.,

$$\bar{I}_{ij}(x) = \begin{cases} B_{ij} & j < x_i \\ S & j \geq x_i \end{cases} \tag{5}$$

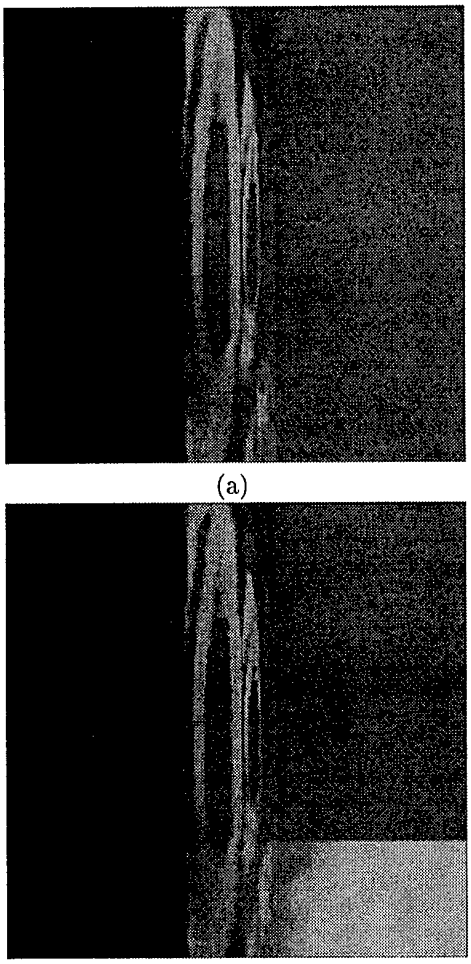
Two possible background image models are obtained by low-pass filtering the images of figure 4. These images were created by removing the main flame in figure 1.

Figure 5 displays a row of $\bar{I}(x)$ assuming that $x_i = 125$ (solid line) and the background profile (dashed line) obtained from figure 4(a).

Assuming that $W_i \sim N(0, R_i)$, it can be concluded from (4) and (5) that

$$p(I_i|x_i) \propto \exp\left\{-\frac{1}{2}[I_i - \bar{I}_i(x_i)]^T R_i^{-1}[I_i - \bar{I}_i(x_i)]\right\} \tag{6}$$

Furthermore, it will be assumed that image lines are independent and the components of W_i are uncorrelated random variables with variance σ^2. Simplifying (6) and taking logarithm of both members, one obtains



(a)

(b)

Figure 4: Background Images.

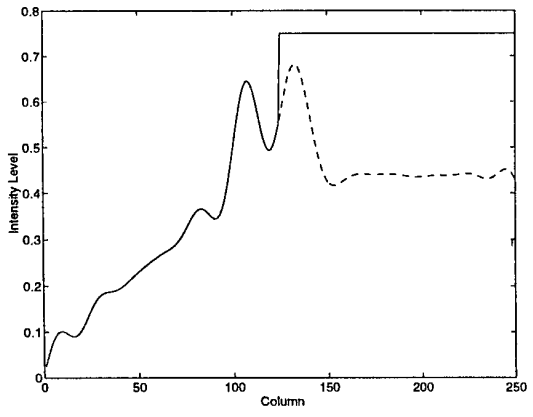


Figure 5: Image line model with $x_i = 125$ (solid line) and background model (dashed line).

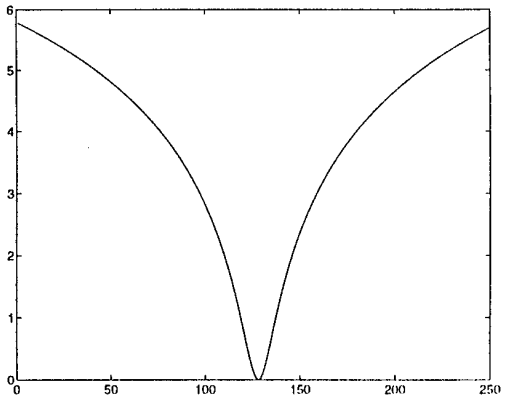


Figure 6: Lorentzian function with $\bar{x}_i = 125$ and $\beta = 5$.

$$\log\{p(I|x)\} = \sum \log\{p(I_i|x_i)\}$$
$$\log\{p(I_i|x_i)\} = K - \frac{1}{2\sigma^2}\|I_i - \bar{I}_i(x_i)\|^2 \quad (7)$$

where $\|.\|$ denotes the Euclidean norm.

4. CONTOUR MODEL

The flame boundary is modelled in this paper as a unidimensional Markov random field with Gibbs probability density function [5]

$$p(x) = \frac{1}{Z}\exp\left\{-\sum_C V_C(x)\right\} \quad (8)$$

where $Z = \sum \exp\{-\sum_C V_C(x)\}$ is the partition function and $V_C(x)$ is the potential of clique C, (clique is an isolated site or a set of sites such that any two sites in C are neighbors). Clique potentials are defined by the user or estimated from a large set of data (field realizations). In this paper, the first strategy is adopted.

Since Gibbs distribution (8) is used as a prior in the estimation of the flame boundary, it should contain the a priori knowledge available about the unknown parameters. The flame boundary is a smooth curve with a known average shape (see figure 7). We shall use zero order and second order cliques to model this information.

Zero order cliques are used to measure the distance of the estimated flame boundary with respect to the average shape. The potential of a zero order clique is defined by the Lorentzian function [3]

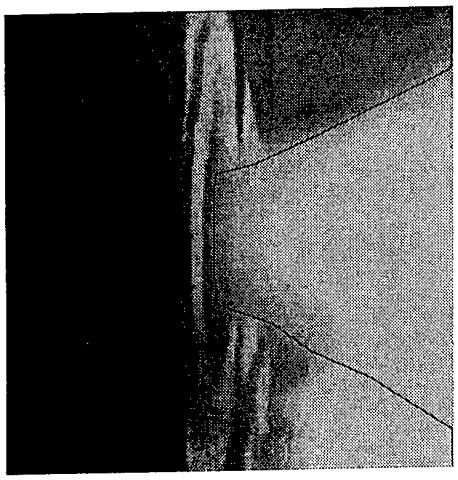


Figure 7: Mean shape.

$$V_{C_i^0}(x_i) = \log\left\{1 + \frac{1}{2}\left(\frac{x_i - \bar{x}_i}{\beta}\right)^2\right\} \quad (9)$$

as shown in figure 6, where $\bar{x}$ is the mean boundary defined by the user or computed from a large set of data.

The probability density function associated with a zero order cliques

$$p_i^0(x_i) \propto \frac{1}{1 + \frac{1}{2}\left(\frac{x_i - \bar{x}_i}{\beta}\right)^2} \quad (10)$$

has longer tails than the ubiquitous normal distribution. Therefore, outliers (boundary values far from the average) have smaller influence on the estimates, leading to robust estimation procedures.

To increase the smoothness of the estimated contours, second order cliques $\{i-1, i, i+1\}$ are used. The clique potentials are define by

$$V_{C_i^2}(x_{i-1}, x_i, x_{i+1}) = \frac{1}{12\lambda^2}(x_{i-1} - 2x_i + x_{i+1})^2 \quad (11)$$

These are regularization terms similar to the ones used in ill-conditioned computer vision problems [6] or in active contours [7], [1].

Using equations (8), (9) and (11) one obtains the prior

$$p(x) \propto \prod_{i+1}^{L} \frac{\exp\left\{-\frac{1}{12\lambda^2}(x_{i-1} - 2x_i + x_{i+1})^2\right\}}{1 + \frac{1}{2}\left(\frac{x_i - \bar{x}_i}{\beta}\right)^2} \quad (12)$$

5. MAP ESTIMATOR

Let us now discuss the computation of the MAP shape estimate. Replacing (2) and (8) in (1),

$$\begin{aligned}\hat{x}_{MAP} &= \arg\max_x [p(I|x)p(x)] \\ &= \arg\min_x \left\{-\log[p(I|x)] + \sum_C V_C(x)\right\}\end{aligned} \quad (13)$$

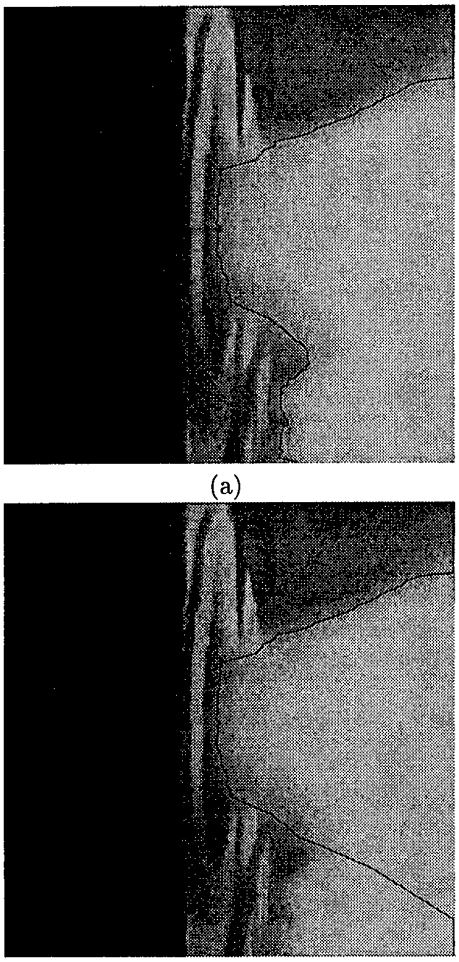
The evaluation of MAP estimates requires the optimization of a non-convex function with a large number of variables. Several methods have been proposed to tackle this problem (e.g. Metropolis algorithm [8], Gibbs sampler [5] or iterated conditional modes [2]). We have used the iterated conditional modes (ICM) algorithm proposed by Besag in [2]. ICM is a deterministic relaxation algorithm which performs a minimization with respect to a single variable in each iteration. Let $\{i_1, i_2, \ldots, i_t, \ldots\}$ be a sequence of sites (each site must occur an infinite number of times). At t-th iteration x_{i_t} is modified to minimize the cost function keeping the other variables constant. This leads to a recursive update law

$$\begin{aligned}\hat{x}_i &= \arg\min_{x_i} \left\{\frac{1}{2\sigma^2}\sum_j \left[I_{ij} - \bar{I}_{ij}(x_i)\right]^2\right. \\ &+ \log\left\{1 + \frac{1}{2}\left(\frac{x_i - \bar{x}_i}{\beta}\right)^2\right\} \\ &+ \left.\frac{1}{2\lambda^2}\left[x_i - \frac{1}{6}\left(-\hat{x}_{i-2} + 4\hat{x}_{i-1} + 4\hat{x}_{i+1} - \hat{x}_{i+2}\right)\right]^2\right\}\end{aligned} \quad (14)$$

The algorithm is initialized using the maximum likelihood estimate of the flame boundary.

6. EXPERIMENTAL RESULTS

Figure 8(a) shows the boundary estimates obtained using background image of figure 4(a) and mild *a priori* shape information. In this case, acceptable estimates of the flame boundary are obtained in the upper part and middle part of the image. However, the algorithm is not able to separate the main flame from smaller ones which are observed in the lower part of the image. This problem can be overcome by using strong shape restrictions (small β). Unfortunately, this reduces the ability to correctly estimate shape deformations. One way to alleviate this difficulty consists of including the other observed flames in the background image (see figure 4(b)). Since these regions are saturated, only shape prior information will be relevant to estimate the main flame boundary there. The results obtained with this background model are shown in figure 8(b).



(a)

(b)

Figure 8: Contour estimates using background images of figure 4.

7. CONCLUSIONS

This paper describes an algorithm for the segmentation of industrial flames in a cluttered background. The proposed method is a modified version of the algorithm developed by Figueiredo and Leitão in [4] to encompass the use of robust estimation techniques which are instrumental in industrial applications. A Bayesian approach is adopted to estimate the flame boundary. The boundary is obtained by a MAP estimator using an image model and a shape prior. The image model takes into account the information available about the structure of the background. This information is obtained from an image of the background. To model the shape prior, an unidimensional Markov random field is used. Lorentzian functions are used to define the clique potentials in order to improve the robustness of the estimates in the presence of shape outliers. This is considered as a key feature in the performance of the algorithm with real images.

Flame images obtained inside a boiler of Setubal thermo-electric plant were used to illustrate the algorithm performance.

8. ACKNOWLEDGEMENTS

This work was supported by Companhia Portuguesa de Produção de Electricidade (CPPE) under contract CHAMA. We thank Eng. Queiros dos Santos, Eng. António Gonçalves and Eng. Macário Marques of CPPE for support and technical expertise. We thank Prof. Miranda Lemos of INESC for involving us in this project.

9. REFERENCES

[1] A. Abrantes, J. Marques, A Class of Constrained Clustering Algorithms for Object Boundary Extraction, *IEEE Trans. Image Processing*, Oct. 1996.

[2] J. Besag, On the statistical analysis of dirty picture, *J. Royal Statist. Soc. B*, Vol. 48, 1986, pp. 259-302.

[3] M. Black, A. Rangarajan, The outlier process: Unifying line processes and robust statistic, *IEEE Proc. Comp. Vision and Pattern Recog.*, 1994, pp. 15-22.

[4] M. Figueiredo, J. Leitão, Bayesian estimation of ventricular contours in angiographic images, *IEEE Trans. On Medical Imaging*, Vol. 11, No. 3, Set. 92, pp. 416-429.

[5] S. Geman, D. Geman, Stochastic Relaxation, Gibbs distribution, and Bayesian restoration of images, *IEEE Trans. On Pattern Anal. Mach. Intel*, Vol. PAMI-6, No. 6, Nov. 84, pp. 721-741.

[6] B. Horn, Robot Vision, MIT Press, 1986.

[7] M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active Contour Models, *Int. J. Computer Vision*, Vol. 1, 1987, pp. 259-268.

[8] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, Equation of state calculation by fast computing machines, *J. Chem. Phys.*, Vol. 21, 1953, pp. 1087-1091.

[9] W. Pratt, Digital image processing, 2nd Edition, Wiley-Interscience.

Array Processing and Channel Identification

CANCELLATION OF EXTERNAL AND MULTIPLE ACCESS INTERFERENCE IN CDMA SYSTEMS USING ANTENNA ARRAYS

Olga Muñoz, Juan A. Fernández-Rubio

Dpt. Teoria Senyal i Comunicacions, UPC, Barcelona, SPAIN
Tel: +34-3-4016431; fax: +34-3-4016447
e-mail: olga@gps.tsc.upc.es

ABSTRACT

Multiuser detectors for CDMA systems have been an active area of research for last two decades. Most of these detectors are developed under the assumption of having no interference external to the system. This paper presents a multisensor-multiuser detector for CDMA systems able to estimate the spatial signature of all active users projected onto the subspace orthogonal to the external interference. With this information an specific beamformer can be designed for each user able to null both the external and multiple access interference.

1. INTRODUCTION

Direct-Sequence Code Division Multiple Access (DS-CDMA) is an accepted technique for future high capacity digital wireless communications systems. Nevertheless, despite of the number of desirable features, CDMA systems are interference limited and suffer from near-far problem. The near-far problem occurs when different users have dissimilar powers and their code sequences are not perfectly orthogonal. In an asynchronous DS-CDMA system it is impossible guarantee that the users' received signals are orthogonal for every possible realization of the propagation delays. Another important limitation for the performance of wireless CDMA systems is multipath fading, which induces more dissimilar powers. Furthermore, when there is a single fading path, there is no means of use the inherent temporal diversity of CDMA to overcome fading.

The standard receiver for DS-CDMA is simply a bank of matched filters, each filter matched to a particular user code. The standard receiver works fairly well in a system where we have only a few users whose codes are almost orthogonal and where the received powers are equal. However, in a near-far situation the standard receiver fails. Since multiple-access-interference is a highly structured interference and the signature waveforms of all users are available at the central receiver, this additional knowledge may be exploited in the decision process. Multiuser detection has the capability of eliminating near-far problem and providing a capacity increase in CDMA systems [1][2]. On the other hand, since spatial diversity reception combats the fading effects of the channel, some multiuser receivers incorporating explicit antenna diversity to overcome fading have been also contemplated in the literature [3][4]. The use of an antenna array has been also considered to cancel those users with higher power, avoiding the requirement of perfect power control [4][5].

Nevertheless, multiuser receivers developed so far do not consider the possibility of having interferences external to the system. We must take into account that most existing users in any given frequency band are narrowband. A certain level of out-of-band spurious emission is unavoidable and, in fact, is legally permitted. In the limit when the interference gets very close to a base site, it can significantly degrade the capacity of the entire cell. Such jamming from existing services to new mobile services should be considered in the design of a high performance receiver. Another kind of external interference is that one provocated by other-cell user, about which the centralized receiver has no information.

In this paper we present a multisensor-multiuser scheme able to overcome near-far problem, external interference and multipath fading. The basic steps of the algorithm are the following. First, the received signal is fed to a bank of matched filters. Using the outputs of the filters matched to the active users plus the output of one filter matched to an unused code, we first estimate the interference subspace. This information is then used jointly with the signal at the filters output to estimate the spatial signature of every user projected onto the subspace orthogonal to the external interference.

The rest of the paper is organized as follows: Next section describes the signal model. In section 3 the proposed method is formulated. Section 4 presents some

THIS WORK WAS SUPPORTED BY THE NATIONAL RESEARCH PLAN OF SPAIN CICYT, UNDER GRANT TIC-95-1022-C05-01.

simulation results and finally in section 5 general conclusions are drawn.

2. PROBLEM FORMULATION

The system under consideration is a K-user asynchronous DS-CDMA system using BPSK modulation and operating over a frequency non-selective channel. The baseband signal for the k-th user is given by

$$s_k(t) = \sum_m d_k[m]\, b_k(t - mT) \tag{1}$$

the data stream $d_k[m] \in \{+1, -1\}$ is pulse amplitude modulated by a period of the code waveform $b_k(t)$, with $b_k(t) = 0$ for $t \notin \{0, T\}$ and T the bit time.

$$b_k(t) = \sum_{l=0}^{L-1} c_{kl} P_{Tc}(t - lT_c) \tag{2}$$

c_{kl} is the l-th chip in the k-th code and P_{Tc} is a rectangular pulse of duration $T_c = T/L$.

The received baseband signal for the multichannel case is

$$\mathbf{x}(t) = \sum_{k=1}^{K} \sqrt{p_k} s_k(t - \tau_k)\mathbf{a}_k(t) + \mathbf{A}_I \mathbf{i}(t) + \mathbf{n}(t) \tag{3}$$

p_k, τ_k and $\mathbf{a}_k$ are respectively the transmitted power, the propagation delay and the steering vector with dimension equal the number of sensors N. All for the k-th user. Matrix $\mathbf{A}_I$ contains the steering vectors of the interferences and vector $\mathbf{i}(t)$ contains the interfering signals at time t:

$$\mathbf{A}_I = [\mathbf{a}_{K+1} \ldots \mathbf{a}_{K+I}] \tag{4}$$

$$\mathbf{i}(t) = [i_1(t) \ldots i_I(t)]^T \tag{5}$$

with I the number of directional external interferences. No assumption about the temporal structure of the interferences is made.

$\mathbf{n}(t)$ is the noise vector at the array input. The noise is considered white Gaussian, uncorrelated among different sensors and with equal variance σ^2 for all of them. As we are assuming frequency non-selective channels, $\mathbf{a}_k(t)$ may be considered as the sum of P coherent paths [6]. Then, we call $\mathbf{a}_k(t)$ the *generalized steering vector* or *spatial signature* of the k-th signal. This vector may be time-varying due to the combined effect of multipath and Doppler. Here, $\mathbf{a}_k$ is assumed to be slowly varying compared to the symbol time:

$$\mathbf{a}_k(t) \cong \mathbf{a}_k(t + T) \tag{6}$$

The model [6] and the proposed method are easily generalizable to frequency selective channels.

3. MULTIUSER SEPARATION AND INTERFERENCE SUPPRESSION

The received signal vector is fed to a bank of $K + 1$ filters. Each one of the first K filters is matched to one of the K active users of the system. The last one is matched to an unused code. The sampled output of the l-th filter is:

$$\mathbf{z}_l[n] = \frac{1}{T}\int_{nT+\tau_l}^{(n+1)T+\tau_l} \mathbf{x}(t) b_l(t - nT - \tau_l)dt \qquad l = 1 \ldots K+1 \tag{7}$$

Let be

$$\begin{aligned} c_{kl}[n] &= \frac{1}{T}\int_{nT+\tau_l}^{(n+1)T+\tau_l} s_k(t - \tau_k) b_l(t - nT - \tau_l)dt \\ &= \beta_{kl} d_k[n] + \gamma_{kl} d_k[n + sgn(\tau_{lk})] \end{aligned} \tag{8}$$

τ_{lk} is $\tau_l - \tau_k$, $sgn(\tau)$ denotes the sign function equals ± 1 depending on the sign of τ or 0 if τ=0, and

$$\beta_{kl} = \frac{1}{T} R_{b_k b_l}(\tau_{lk}) \tag{9}$$

$$\gamma_{kl} = \frac{1}{T} R_{b_k b_l}(\tau_{lk} - T sgn(\tau_{lk})) \tag{10}$$

$R_{b_k b_l}(\tau)$ denotes cross correlation function between the k-th and l-th signature signals at τ offset, that is

$$R_{b_k b_l}(\tau) = \int_0^T b_k(t + \tau) b_l(t)\, dt \tag{11}$$

Note that $\beta_{lk} = \beta_{kl}$ and $\gamma_{lk} = \gamma_{kl}$. For the k-th user $\beta_{kk} = 1$ and $\gamma_{kk} = 0$.

Finally, eq. 7 may be written as

$$\mathbf{z}_l[n] = \sum_{k=1}^{K} \sqrt{p_k} c_{kl}[n]\, \mathbf{a}_k + \mathbf{A}_I \mathbf{i}_l[n] + \mathbf{n}_l[n] \tag{12}$$

where $\mathbf{i}_l$ and $\mathbf{n}_l$ are respectively the interference and noise vector filtered by the l-th filter.

Consider now the spatial correlation matrix

$$\begin{aligned} \mathbf{R}_{zz,l} &= E\{\mathbf{z}_l[n]\, \mathbf{z}_l^H[n]\} \\ &= \sum_{k=1}^{K}\sum_{r=1}^{K} \sqrt{p_k p_r} E\{c_{kl}[n]\, c_{rl}^*[n]\}\, \mathbf{a}_k \mathbf{a}_r^H + \\ &+ \mathbf{A}_I E\{\mathbf{i}_l[n]\, \mathbf{i}_l[n]\}\, \mathbf{A}_I^H + E\{\mathbf{n}_l[n]\, \mathbf{n}_l[n]\} \end{aligned} \tag{13}$$

$E\{\cdot\}$ and H denote expectation and complex conjugate transpose, respectively.

Let's calculate different terms. Assuming that symbols are uncorrelated:

$$E\left\{c_{kl}[n]\, c_{rl}^{*}[n]\right\} = \left(\beta_{kl}^2 + \gamma_{kl}^2\right)\delta_{kr} \tag{14}$$

where δ_{kr} is the Kronecker delta.

Regarding the external interference, let be:

$$\begin{aligned}\mathbf{Q}_l &= E\left\{\mathbf{i}_l[n]\,\mathbf{i}_l^H[n]\right\} \\ &= \frac{1}{T^2}\int_0^T\int_0^T \mathbf{S}_i(u-v)b_l(u)b_l(v)dudv \end{aligned}\tag{15}$$

with

$$\mathbf{S}_i(\tau) = E\left\{\mathbf{i}(t+\tau)\,\mathbf{i}^H(t)\right\} \tag{16}$$

Working with expression 15, $\mathbf{Q}_l$ can be derived as:

$$\mathbf{Q}_l = \frac{1}{T^2}\int_{-T}^{T} \mathbf{S}_i(\lambda)R_{b_l b_l}(\lambda)d\lambda \tag{17}$$

The spatial correlation of filtered noise is:

$$E\left\{\mathbf{n}_l[n]\,\mathbf{n}_l^H[n]\right\} = \sigma_l^2\mathbf{I}_N = \frac{\sigma^2}{L}\mathbf{I}_N \tag{18}$$

$\mathbf{I}_N$ is the identity matrix with dimension NxN.

Finally, we can write the spatial correlation matrix at the output of the first K filters as:

$$\begin{aligned}\mathbf{R}_{zz,l} = p_l\mathbf{a}_l\mathbf{a}_l^H + \sum_{\substack{k=1\\k\neq l}}^{K} p_k\left(\beta_{kl}^2+\gamma_{kl}^2\right)\mathbf{a}_k\mathbf{a}_k^H + \\ +\mathbf{A}_I\mathbf{Q}_l\mathbf{A}_I^H + \sigma_l^2\mathbf{I}_N \qquad l=1,...,K\end{aligned}\tag{19}$$

and the spatial correlation matrix at the output of the $K+1$ filter as:

$$\begin{aligned}\mathbf{R}_{zz,K+1} &= \sum_{k=1}^{K} p_k\left(\beta_{k(K+1)}^2+\gamma_{k(K+1)}^2\right)\mathbf{a}_k\mathbf{a}_k^H + \\ &+\mathbf{A}_I\mathbf{Q}_{K+1}\mathbf{A}_I^H + \sigma_{K+1}^2\mathbf{I}_N\end{aligned}\tag{20}$$

Let be matrix $\mathbf{S}$, the matrix whose element at the k-th row and l-th column is $\beta_{kl}^2+\gamma_{kl}^2$. Thus, dimension of $\mathbf{S}$ is $(K+1)\,x\,(K+1)$

$$\mathbf{S} = \begin{bmatrix} 1 & \cdots & \beta_{1(K+1)}^2+\gamma_{1(K+1)}^2 \\ \vdots & \ddots & \vdots \\ \beta_{1(K+1)}^2+\gamma_{1(K+1)}^2 & \cdots & 1 \end{bmatrix} \tag{21}$$

Let's define also matrices $\mathbf{R}_{zz}$, $\mathbf{R}_I$, $\mathbf{A}$, and $\mathbf{N}$ as follows:

$$\begin{aligned}\mathbf{R}_{zz} &= \begin{bmatrix}\mathbf{R}_{zz,1}\\ \vdots \\ \mathbf{R}_{zz,K+1}\end{bmatrix} \qquad \mathbf{R}_I = \begin{bmatrix}\mathbf{A}_I\mathbf{Q}_1\mathbf{A}_I^H \\ \vdots \\ \mathbf{A}_I\mathbf{Q}_{K+1}\mathbf{A}_I^H\end{bmatrix} \\ \mathbf{A} &= \begin{bmatrix} p_1\mathbf{a}_1\mathbf{a}_1^H \\ \vdots \\ p_K\mathbf{a}_K\mathbf{a}_K^H \\ \mathbf{0}\end{bmatrix} \qquad \mathbf{N} = \frac{\sigma^2}{L}\mathbf{1}\otimes\mathbf{I}_N\end{aligned}\tag{22}$$

where the symbol $\otimes$ denotes the Kronecker product, $\mathbf{0}$ is an all zeros matrix with dimension NxN and $\mathbf{1}$ is an all ones column vector with dimension $K+1$. With previous definitions and some algebraic manipulation of eq. 19 and 20 we can write:

$$\mathbf{R}_{zz} = (\mathbf{S}\otimes\mathbf{I}_N)\mathbf{A} + \mathbf{R}_I + \frac{\sigma^2}{L}\mathbf{1}\otimes\mathbf{I}_N \tag{23}$$

Operating upon matrix $\mathbf{R}_{zz}$ we obtain a new matrix $\mathbf{M}$. This new matrix can be partitioned into $K+1$ blocks with dimension NxN:

$$\mathbf{M} = (\mathbf{S}^{-1}\otimes\mathbf{I}_N)\mathbf{R}_{zz} = \begin{bmatrix}\mathbf{M}_1^T & \cdots & \mathbf{M}_{K+1}^T\end{bmatrix}^T \tag{24}$$

The part of $\mathbf{M}$ corresponding to the k-th user is

$$\mathbf{M}_k = p_k\mathbf{a}_k\mathbf{a}_k^H + \mathbf{A}_I\tilde{\mathbf{Q}}_k\mathbf{A}_I^H + u_k\frac{\sigma^2}{L}\mathbf{I}_N \tag{25}$$

u_k is the k-th element of vector $\mathbf{u} = \mathbf{S}^{-1}\mathbf{1}$ and $\tilde{\mathbf{Q}}_k$ is calculated as the k-th part of matrix:

$$\tilde{\mathbf{Q}} = \begin{bmatrix}\tilde{\mathbf{Q}}_1 \\ \vdots \\ \tilde{\mathbf{Q}}_{K+1}\end{bmatrix} = (\mathbf{S}^{-1}\otimes\mathbf{I}_I)\begin{bmatrix}\mathbf{Q}_1 \\ \vdots \\ \mathbf{Q}_{K+1}\end{bmatrix} \tag{26}$$

$\mathbf{I}_I$ is the identity matrix with dimension IxI. Remember that I is the number of external interferences.

The last part of $\mathbf{M}$, corresponding to the $K+1$ filter, is:

$$\mathbf{M}_{K+1} = \mathbf{A}_I\tilde{\mathbf{Q}}_{K+1}\mathbf{A}_I^H + u_{K+1}\frac{\sigma^2}{L}\mathbf{I}_N \tag{27}$$

u_{K+1} is the $K+1$-th element of vector $\mathbf{u}$. The signal subspace component in matrix $\mathbf{M}_{K+1}$ contains only information about the external interferences. The other submatrices have information about the external interferences and the signal of the corresponding user. The main eigenvector in this case, depending on the signals power, may be quite different of the steering vector of

any of the involved signals. Nevertheless, it is possible to eliminate the information relative to the external interferences in each $\mathbf{M}_k$ with $k \neq K+1$. For this purpose, from the signal subspace of $\mathbf{M}_{K+1}$ we compute the projection matrix onto the subspace orthogonal to the external interferences. The projection matrix is:

$$\begin{aligned} \mathbf{P}_\perp &= \mathbf{E}_{n,K+1}\mathbf{E}_{n,K+1}^H = \mathbf{I}_N - \mathbf{E}_{s,K+1}\mathbf{E}_{s,K+1}^H \\ &= \mathbf{I}_N - \mathbf{A}_I(\mathbf{A}_I\mathbf{A}_I^H)^{-1}\mathbf{A}_I \end{aligned} \quad (28)$$

where $\mathbf{E}_{n,K+1}$ is the noise subspace of matrix $\mathbf{M}_{K+1}$. Its columns are the noise subspace eigenvectors. $\mathbf{E}_{s,K+1}$ is the signal subspace of matrix $\mathbf{M}_{K+1}$. Its columns are the signal subspace eigenvectors. Note that it is not necessary to calculate the steering vectors of the interferences individually (matrix $\mathbf{A}_I$). We can compute $\mathbf{P}_\perp$ from the global signal subspace or global noise subspace.

Next step is projecting each one of the submatrices $\mathbf{M}_k$. As resulting of the projection operation, each matrix $\mathbf{P}_\perp\mathbf{M}_k$ $(k = 1...K)$ has only one signal eigenvector: $\mathbf{a}_{P,k}$. This eigenvector is the steering vector of the corresponding user projected onto the subspace orthogonal to the external interference. With this information an specific beamformer for each user may be computed able to null external and multiple access interference. The weight vector for the k-th user is the k-th column of matrix $\mathbf{W}$:

$$\mathbf{W} = \mathbf{A}_P(\mathbf{A}_P^H\mathbf{A}_P)^{-1}\mathbf{I}_K \quad (29)$$

with $\mathbf{A}_P = [\mathbf{a}_{P,1} \ldots \mathbf{a}_{P,k} \ldots \mathbf{a}_{P,K}]$ and $\mathbf{I}_K$ the identity matrix with dimension KxK.

4. SIMULATION RESULTS

An asynchronous CDMA system was simulated in order to investigate the performance of the multiuser detection algorithm presented in this paper. The modulating signals in this system are Gold sequences with length L=31.

The characteristics of an external interference depend to a large extent on its origin. It may be categorized as being either broadband or narrowband relative to the bandwidth of the information-bearing signal. The proposed detector permits capacity increase exploiting the different spatial signature of users even in presence of external interferences from which the receiver has no information. As an example we illustrate the separation of two users (K=2) that are received by an array of six $\lambda/2$ linearly spaced sensors (N=6). We consider a jamming signal consisting of one sinusoid of frequency equal the carrier frequency of the CDMA signals. The angles of arrival are 35° and 45° from broadside for the systems users and 70° for the jamming signal. The signal to noise ratio (SNR) at each of the sensors is 10, 16 and $23dB$ respectively. The relative propagation delays for the system users are τ_1=0 and τ_2=$2T_c$. At the receiver three matched filters are used. The first and second filter are matched to the first and second user code, synchronized with the corresponding one. The third one does not need to be matched to any system user. In the simulation we have assumed that is $5T_c$ delayed with respect to the first user. The correlation matrix at the output of the matched filters bank is estimated by temporal averaging of the despreaded signal vector, using a block size of 100 symbols. The beamformer designed for the system users under these conditions are shown in figure 1.

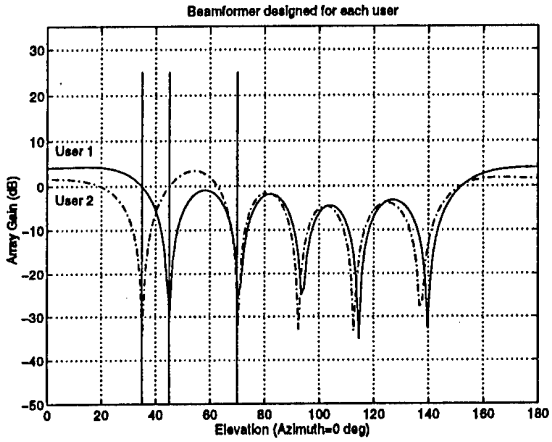


Fig. 1: Beamformer for active system users

If an exact knowledge of the correlation matrix at the output of the matched filters would be available, estimation of the steering vector would be perfect for every user, independently of the signal powers. The limitations in the estimation are not due to the method itself but to an inexact knowledge of the correlation matrix. The requirement of an accurate estimation of matrices $\mathbf{R}_{zz,k}$ limits the possible dynamic range of the powers of the received signals, since a weak signal may be masked by a strong one. Nevertheless, the powers range where the proposed receiver offers good performance is excellent. This situation is illustrated in figure 2, which plots the $SNIR$ (signal to noise plus interference ratio) at the array output (solid line) versus the power of the narrowband interference, for the first (-o-) and the second user (-+-). The $SNIR$ of each user at the output of the classical receiver (a single matched filter) is also illustrated with dashed line.

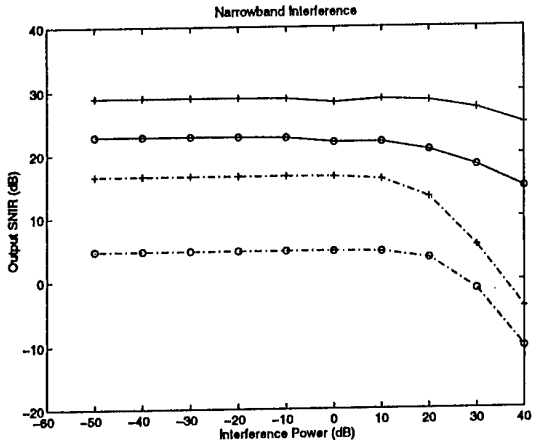


Fig. 2: Output SNIR with narrowband jamming

In previous example we considered a narrowband interference. Nevertheless, no knowledge about jamming was used by the receiver. Then, the same procedure may be applied to another kind of interference. As a second example we consider a broadband one. This is characterized as a BPSK signal with random chips. This situation is worse for the classical receiver, since the code gain is now reduced with respect to previous case. On the contrary, performance of the proposed detector is practically the same than in the narrowband case.

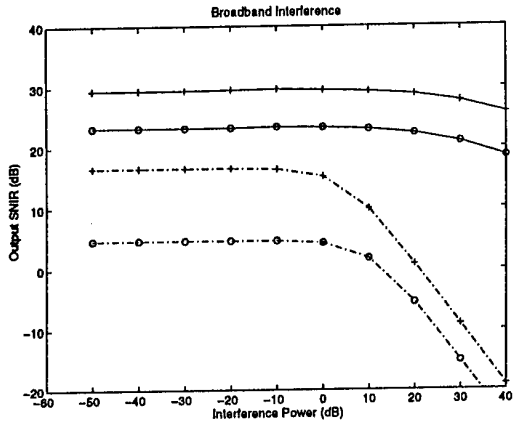


Fig. 3: Output SNIR with broadband jamming

The code gain with respect multiple access interference in both cases is smaller than that achieved in a synchronous system. Specifically, for our two users the gain with respect to each other is reduced from $30dB$ to $11dB$ for the time delays considered. As can be deduced from fig. 2 and 3, in such a situation the use of an antenna array may be very useful. The rejection of the multiple access interference yields a substantial improvement of the $SNIR$ with respect to the classical receiver. This improvement is achieved even in the presence of external interference with high power. Since the external interference is also cancelled, the improvement is more significant when the jamming signal has higher power.

5. CONCLUSIONS

A new method for steering vector estimation in CDMA systems in the presence of interference external to the system has been proposed. The most important features of this method are related below.

As can be observed from the simulations, the proposed method is almost independent on the nature of the jamming signal and can cope with a very high dynamic range of interference power. The estimations are carried out working at the symbol rate instead of the chip rate, without any temporal reference or any *a priori* spatial information. The required information is the codes waveform and timing of active users. This information is always required in order to demodulate the signals. As no model is assumed for the steering vectors the proposed solution is robust to calibration errors. In the simulations here presented we have considered a non-multipath scenario because is easier to extract conclusions from the array beam pattern having only one direction of arrival. Nevertheless, since there is no model assumption for the steering vector, the same procedure can be applied to estimate the generalized steering vector when the signals arrive from multiple reflections, while other classical DOA estimation methods as MUSIC fails in this situation.

6. REFERENCES

[1] R. Lupas and S. Verdú, "Near-Far Resistance of Multiuser Detectors in Asynchronous Channels," IEEE Trans. on Comm., vol. 38, March 1990

[2] S. Verdú, "Recent Progress in Multiuser Detecction", Multiple Access Communications (A Selected Reprint Volumen), IEEE Press, 1992

[3] Z. Zvonar, "Multiuser Detection and Diversity Combining for Wireless CDMA Systems", Wireless and Mobile Communications, J. Holtzman and D. Goodman Eds. Kluwer Academic Publishers, 1994

[4] S. Miller and S. Schwartz, "Integrated Spatial-Temporal Detectors for Asynchronous Gaussian Multiple-Access Channels", IEEE Trans. on Comm., vol. 43, Feb. 1995

[5] O. Muñoz, and J. Fernández-Rubio, "Blind Multiuser Combining at the Base Station for Asynchronous CDMA Systems", in Proc. ICASSP'96, May 1996

[6] O. Muñoz, and J. Fernández-Rubio, "Adaptive Arrays for Frequency Non-Selective and Selective Channels", in Proc. EUSIPCO'94, vol. 3, pp. 1536-1539, Sept. 1994

BLIND ADAPTIVE BEAMFORMING USING THE SPECTRAL LINE GENERATION PROPERTY OF CPFSK SIGNALS

Daniel Iglesia, Adriana Dapena, Cristina Mejuto, Luis Castedo

Departamento de Electrónica y Sistemas, Universidad de La Coruña
Campus de Elviña s/n, 15.071 La Coruña, Spain
Tel.: ++ 34-81-132552, e-mail: luis@des.fi.udc.es

ABSTRACT

This paper presents a technique for adaptive beamforming that exploits the cyclostationary properties of CPFSK modulations. The method is based on the ability of this type of modulations to generate spectral lines when they are raised to a fractional number which is the inverse of its modulation index. A stochastic gradient algorithm is proposed to compute the coefficients that maximize the output SINR. The algorithm is blind because it does not need to know the transmitted symbols: only the carrier frequency, the symbol rate and the modulation index is required.

1. INTRODUCTION

It is well known that many digital modulated signals generate spectral lines when they pass through certain nonlinear transformations. As an example, linear digital modulations like ASK, PSK, QAM, etc., produce spectral lines at frequencies related with the carrier frequency and the symbol period when they are raised to a integer power (usually 2 or 4). This property has been sucessfully used in [1] to develop a blind adaptive beamforming technique that only requires to know one of the frequencies of the spectral lines generated by the desired signal. This technique consists on adjusting the beamformer coefficients to minimize the Mean Square Error between the array ouput after the nonlinearity and a complex exponential. The advantages of this technique are remarkable: it is not necessary to know the desired signal steering vector (therefore it is very robust to array calibration errors) [2], it does not require a reference signal [2], it does not suffer from capture problems as the Constant Modulus (CM) beamformer [3] and it can be easily implemented with a stochastic gradient algorithm without solving generalized eigenvalues problems [4].

This work has been supported by CICYT, Spain, grant # TIC96-0500-C10-02

The beamforming technique proposed in [1] only considers integer powers and therefore its applicability is limited to linear modulations. The main objective of this paper is to explain how to extend this approach to a CPFSK (Continous Phase Frequency Shift Keying) [5] modulation which belongs to the class of nonlinear memory modulations. CPFSK signals do not generate spectral lines when they are raised to an integer number but it will be demonstrated that they do if they are raised to a fractional number which is the inverse of its modulation index. Therefore, beamformer coefficients can also be selected to minimize the Mean Square Error between the array output after the nonlinearity and a complex exponential. The differences with the integer case is that implementation of the adaptive algorithm requires a phase unwrapping algorithm and analysis of stationary points becomes extremly difficult because fractionally order moments appear. Nevertheless, simulations show that performance is very similar to the integer case.

This paper is organized in five sections. Section 2 presents the spectral line generation property of CPFSK signals. This property is used in section 3 to develope an optimization criterion for blind adaptive beamforming. In section 4 simulation results are presented. Finally, section 5 is devoted to the conclusions.

2. SPECTRAL LINE GENERATION

A zero-mean complex signal $x(t)$ generates a spectral line at a frequency α after passing through the nonlinearity $(.)^r$ if and only if the *r-th* order cyclic moment defined as

$$\begin{aligned} m_{rx}^{\alpha} &= \left\langle x^r(t) e^{-j2\pi\alpha t} \right\rangle \\ &= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x^r(t) e^{-j2\pi\alpha t}\, dt \end{aligned} \quad (1)$$

exists and is nonzero [6]. The operator $\langle\cdot\rangle$ denotes the time average operation and r is a real number not

necessarily integer.

The existence of these cyclic moment arises from the cyclostationarity properties if $x(t)$. More specifically, it can be demonstrated [6] that, under certain conditions, the cyclic moments of $x(t)$ correspond to the Fourier series coefficients of its statistical moments

$$m_{rx}(t) = E[x^r(t)] = \int_{-\infty}^{\infty} x^r(t) f(x,t)\, dx \qquad (2)$$

where $f(x,t)$ is the first order density function of $x(t)$ and $E[\cdot]$ denotes the statistical average operation.

In the sequel we calculate the statistical moment of order $r = 1/h$ for a CPFSK signal with modulation index h and its Fourier series coefficients. A CPFSK signal is represented as

$$x(t) = Acos(2\pi f_c t + 2\pi h \int_{-\infty}^{t} d(\tau)d\tau + \phi_0) \qquad (3)$$

where A is the signal amplitude, f_c is the carrier frequency, h is the modulation index, ϕ_0 is the initial phase of the carrier and $d(t)$ is a PAM signal given by

$$d(t) = \sum_k I_k g(t - kT) \qquad (4)$$

where $g(t)$ is a rectangular pulse of amplitude $1/2T$ and the symbols I_k are the amplitudes which result of mapping digits of the information sequence to the amplitude levels $\{\pm 1, \pm 3, ..., \pm(M-1)\}$. In the interval $nT \leq t \leq (n+1)T$ the complex representation of $x(t)$ can be written as [5]

$$s(t) = Aexp\left\{ j\left(2\pi f_c t + \pi h \phi_n + \pi h I_n \frac{t - nT}{T} + \phi_0 \right) \right\} \qquad (5)$$

where

$$\phi_n = \sum_{k=0}^{n-1} I_k, \qquad (6)$$

Using this expression, the $\frac{1}{h}th$ order moment of $s(t)$ is given by

$$E[s^{\frac{1}{h}}(t)] = A^{\frac{1}{h}} e^{j 2\pi \frac{f_c}{h} t} e^{j \frac{\phi_0}{h}} E[e^{j\pi\phi_n} e^{j\pi I_n (\frac{t-nT}{T})}] \qquad (7)$$

Assuming that the symbols I_n are independent and identically distributed, ϕ_n is independent from I_n and therefore

$$E[e^{j\pi\phi_n} e^{j\pi I_n \frac{t-nT}{T}}] = E[e^{j\pi\phi_n}] E[e^{j\pi I_n \frac{t-nT}{T}}] \qquad (8)$$

Now recall that the term ϕ_n is the sum of n odd numbers I_k. Therefore, ϕ_n is an even number when n is even and it is odd when n is odd and as a consequence $e^{j\pi\phi_n} = e^{j\pi n}$. This fact enables us to express the first average in (8) as follows

$$\begin{aligned} E[e^{j\pi\phi_n}] &= \sum_{\phi_n} e^{j\pi\phi_n} Prob(\phi_n) = \\ &= e^{j\pi n} \sum_{\phi_n} Prob(\phi_n) = e^{j\pi n} \end{aligned} \qquad (9)$$

Let us calculate the second average in (8). The symbols I_n take values inside a set of equiprobable points, i.e., $Prob(I_n = M - 1 - 2k) = 1/M$, $k = 0, ..., M-1$. Therefore, this average takes the following form

$$\begin{aligned} E[e^{j\pi I_n \frac{t-nT}{T}}] &= \frac{1}{M} \sum_{k=0}^{M-1} e^{j\pi(M-1-2k)\frac{t-nT}{T}} \\ &= \frac{e^{-j\pi n}}{M} \sum_{k=0}^{M-1} e^{j\pi(M-1-2k)\frac{t}{T}} \end{aligned} \qquad (10)$$

Finally, substituting (9) and (10) into (7) we obtain the $\frac{1}{h}th$ order moment of $s(t)$

$$E[s^{\frac{1}{h}}(t)] = A^{\frac{1}{h}} e^{j(2\pi\frac{f_c}{h}t + \frac{\phi_0}{h})} \frac{1}{M} \sum_{k=0}^{M-1} e^{j2\pi(M-1-2k)\frac{t}{2T}} \qquad (11)$$

It is apparent that this moment is a periodic function of time since it has the form of a Fourier series expansion. As already mentioned, the coefficients of this expansion are the cyclic moments. This implies that a CPFSK signal has nonzero $\frac{1}{h}th$ order cyclic moments at the following frequencies

$$f_i = \frac{f_c}{h} + \frac{M - 1 - 2i}{2T} \qquad i = 0, \cdots, M-1 \qquad (12)$$

and their value is

$$m_{\frac{1}{h}s}^{\alpha=f_i} = A^{\frac{1}{h}} e^{j\frac{\phi_0}{h}} \qquad i = 0, \cdots, M-1 \qquad (13)$$

This result demonstrates that after passing the signal $s(t)$ through the nonlinearity $(\cdot)^{\frac{1}{h}}$ it is obtained M spectral lines at the frequencies f_i given in (12) with amplitude $\frac{1}{M} A^{\frac{1}{h}} e^{j\frac{\phi_0}{h}}$. To illustrate this property, figure 1 shows the power spectral density (PSD) of a four-level CPFSK signal with modulation index $h = 0.75$ before and after the nonlinearity $(\cdot)^{\frac{1}{h}}$. It can be seen that four spectral lines appear after applying the nonlinearity.

In the following section it is shown how this property can be used to optimally extract a signal from the array output minimizing the noise and interferences effect.

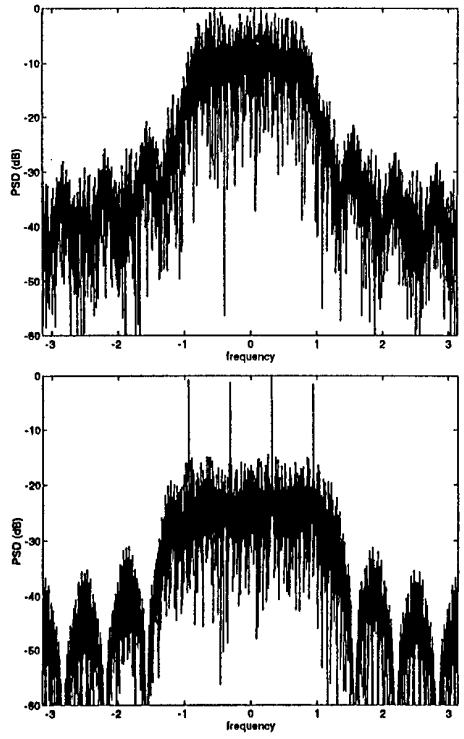


Figure 1: PSD of CPFSK with $h = 0.75$ before and after the nonlinearity

3. OPTIMIZATION CRITERION

Let us consider a narrowband beamformer which processes an input vector $\mathbf{x}(t)$ to produce an output that can be expressed as

$$y(t) = \mathbf{w}^H \mathbf{x}(t) \tag{14}$$

where $\mathbf{w}$ is a complex-valued coefficients vector and ^{H} denotes the conjugate transpose operator.

We propose to adjust the coefficients $\mathbf{w}$ according to the minimization of the following cost function

$$J = < |e^{j2\pi\alpha t} - y^{\frac{1}{k}}(t)|^2 > \tag{15}$$

where $< \cdot >$ denotes the time average operator and α is one of the frequencies f_i where the desired signal generates a spectral line after the nonlinearity $(.)^{\frac{1}{k}}$. Similarly to [1] we claim that minimization of this cost function yields to optimal extraction of the desired signal in the sense that the Signal to Interference and Noise Ratio (SINR) is maximized.

A simple and reasonable way to compute the optimum coefficients $\mathbf{w}$ is the steepest descent method

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu \nabla_{\mathbf{w}} J(n) \tag{16}$$

where μ is the algorithm step size and $\nabla_{\mathbf{w}} J$ represents the complex gradient of J with respect to $\mathbf{w}$ in the instant n. In our particular case $\nabla_{\mathbf{w}} J$ is

$$\nabla_{\mathbf{w}} J = -\frac{1}{h} < \mathrm{e}^*(n) y^{(\frac{1}{k}-1)}(n) \mathbf{x}(n) > \tag{17}$$

where $\mathrm{e}(n) = e^{j\pi\alpha n} - y^{\frac{1}{k}}(n)$ is the error signal whose variance we want to minimize and * denotes the conjugate operator. Substituing the time average in (17) by its instantaneuos estimate we obtain the following stochastic gradient algorithm

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu \mathrm{e}^*(n) y^{\frac{1}{k}-1}(n) \mathbf{x}(n) \tag{18}$$

It is interesting to note that implementation of this algorithm only requires the modulation index h and the frequency α. Since from (12) the value of α depends on h, f_c and T, only these three parameters are required to extract the desired signal. The algorithm is blind because the knowledge of the transmitted symbols is not required.

Implementation of (18) typically requires raising a complex number to a fractional exponent. This operation is carried out as follows

$$z = \rho(z) e^{j\,arg(z)} \Rightarrow z^r = \rho^r e^{j\ arg(z)\ r} \tag{19}$$

where r is any real number. It should be mentioned that when computing $arg(z)$ we cannot use a conventional arctangent subroutine. This type of subroutines give us the principal value of a complex number, $ARG(z)$, which is in the interval $[-\pi, \pi]$ interval. The relationship between the true value of the phase, $arg(z)$, and its principal value is

$$arg(z) = ARG(z) + 2\pi k \tag{20}$$

where k is an integer. However note that for an arbitrary real number r

$$e^{j\ arg(z)\ r} \neq e^{j\ ARG(z)\ r}\ \ e^{j2\pi k\ r} \tag{21}$$

Therefore, the principal value of the phase cannot be used. To compute $arg(z)$ it is necessary an unwrapping pahse algorithm such as the one described in [7].

The cost function that we are minimizing is not a quadratic form of $\mathbf{w}$. This raises the question of wether there are undesirable stationary points that may impair the convergence of the adaptive algorithm. The analysis in [1] shows that for $h = 0.5$ the cost function (15) is free of undesirable minima except when the interferences generate a spectral line to the same frequency that the desired signal. This particular case of CPFSK is known as Minimum-Shift Keying (MSK) [5] and it is very common in communications due to

its greater bandwidth efficience. Analysis for other modulation indices involves fractionally order statistics and turns out to be much more complicated and it has not been performed. However, simulations did not show the existence of undesirable minima.

4. SIMULATIONS

Several computer simulations were carried out to illustrate the performance of the proposed method. We considered a uniform linear array with 10 sensors equispaced half wavelength. The array input signals are sampled at a rate ten times faster than the symbol rate.

In the first simulation example, a simple environment with one binary CPFSK signal and Gaussian noise was considered. Its input SNR is 5 dB and its direction of arrival (DOA) 0°. Figure 2 plots the time evolution of the SINR for different values of the modulation index. In all cases the algorithm step-size is set to 2×10^{-4}. It can be seen that in all cases it converges to the maximum SNR solution. However, rate of convergence strongly depends on the value of h. Simulations showed that convergence is faster for values of h close to 0.5 and 1.

In the second simulation example we considered an environment with three incoming binary CPFSK whose parameters are reflected in table 1. The three have the same modulation index ($h = 0.6$) but different carrier frequencies. Figure 4 shows the time evolution of the SINR in this environment with an algorithm step-size $\mu = 5 \times 10^{-6}$. It can be seen, again, that the algorithm converges to the maximum SINR solution.

Signal	Carrier Freq.	SNR	DOA
Desired	0.0	0 dB	0°
Interf # 1	0.1	10 dB	30°
Interf # 2	0.2	20 dB	−30°

Table 1: Environment parameters

To illustrate the ability of the algorithm to cancel interferences that generate spectral lines at the same frequency as the desired signal, a third computer experiment was carried out in which we considered the same environment as before but with the interferences having the same carrier frequency as the desired signal. Figure 5 shows the time evolution of the SINR for this case. It can be seen that the misadjustment noise of the algorithm has increased but still converges to the maximum SINR solution.

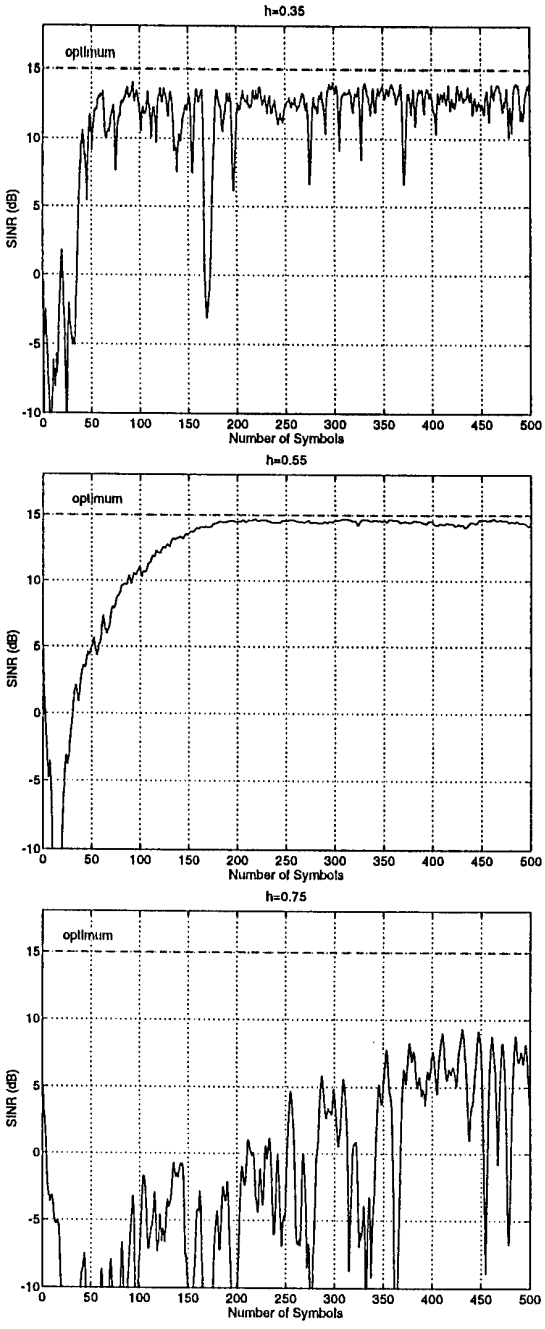


Figure 2: Time evolution of SNR for different values of mdulation index.

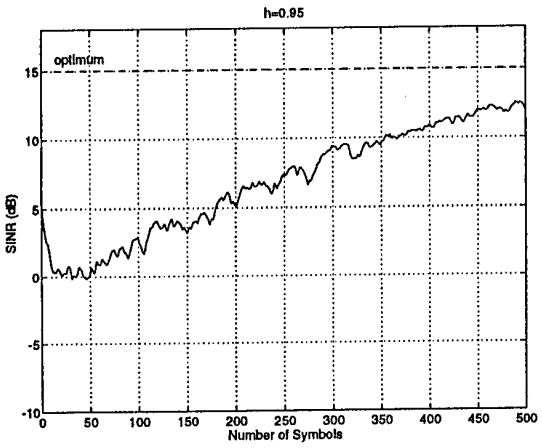


Figure 3: Time evolution of SNR for different values of mdulation index (cont.).

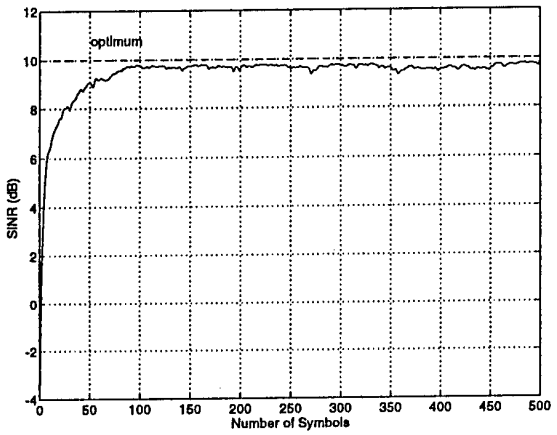


Figure 4: Algorithm performance in an interference environment (different carrier frequences).

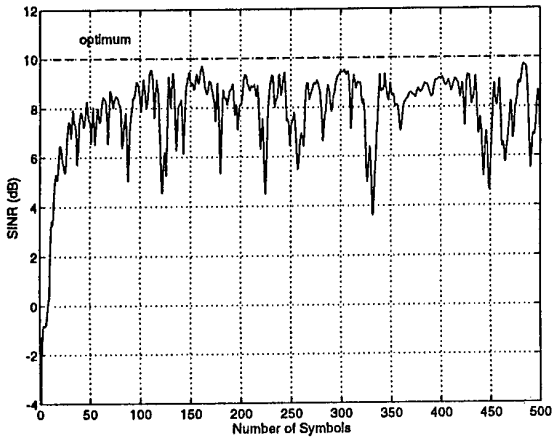


Figure 5: Algorithm performance in an interference environment (same carrier frequences).

5. CONCLUSIONS

This paper presents a technique for adaptive beamforming that exploits the cyclostationary property of CPFSK signals. The method uses the ability of this kind of modulated signals to generate spectral lines when they are raised to a fractional number which is the inverse of its modulation index. The approach is based on minimizing a cost function defined as the Mean Square Error between the array output raised to the inverse of the modulation index and a complex exponential. The frequency of this exponential is one of the frequencies of the spectral lines generated by the desired signal. The approach is blind because the transmitted symbols are not required: only the carrier frequency, the symbol rate and the modulation index is needed to extract the desired signal.

The analysis of the stationary points in the proposed cost function is very complicated because it implies dealing with fractional order statistics and it has not been performed. However, simulations have shown that in the fractional case the behavior is similar to the integer case.

6. REFERENCES

[1] L. Castedo and A. R. Figueiras-Vidal, "An Adaptive Beamforming Technique Based on Cyclostationary Signal Properties", *IEEE Trans. on Signal Processing*, vol. 43, no. 7, pp. 1637-1650, July 1995.

[2] R.T. Compton, Jr.,*Adaptive Antennas: Concepts and Performance*, Englewood Cliffs, New Jersey, Prentice-Hall, 1988.

[3] J. Lundell, B. Widrow, "Application of the Constant Modulus Adaptive Beamformer to Constant and Non-Constant Modulus Signals", *Proc 22nd Asilomar Conf. Signals, Systems and Computers*, pp. 432-436, Pacific Grove, CA, Nov. 1987.

[4] M. D. Zoltowski, J. F. Ramos, "Blind Adaptive Beamforming for Narrow-Band Cochannel Digital Communications Signals", submitted to *IEEE Trans. Signal Processing*, 1995.

[5] J. G. Proakis, *Digital Communications*, McGraw-Hill. Singapore, 1995.

[6] W. A. Gardner, *Statistical Analysis: A Non Probabilistic Theory*, Englewood Cliffs, NJ: Prentice-Hall, 1987.

[7] A. V. Oppenheim, R. W. Schafer, *Discrete-Time Signal Processing*, Englewood Cliffs, New Jersey, Prentice-Hall, 1989.

ON THE PERFORMANCE OF THE CONSTANT MODULUS ARRAY RESTRICTED TO THE SIGNAL SUBSPACE†

J.R. Cerquides and J.A. Fernández-Rubio

Signal Theory and Communications Department, Polytechnic University of Catalonia
Módulo D5, Campus Nord UPC, C/Gran Capitán s/n, 08034 Barcelona, SPAIN
Tel.:+34-3-4015938, Fax:+34-3-4016447
E-mail: ramon@tsc.upc.es

ABSTRACT†

The Constant Modulus Array has a slow rate of convergence mainly due to both the nonconvex nature of its cost function and the well known behavior of stochastic steepest descent algorithms for environments with a large eigenvalue spread. In this paper we analyze the solutions of the Constant Modulus Cost Functions, showing that the weight vectors associated to its minima lie on the signal subspace. From that information we develop a modified version of the Constant Modulus Array, which speeds up the convergence and reduces the final misadjustment error. The proposed method is specially useful for arrays having a large number of sensors and low Signal to Noise Ratio for the Source of Interest.

1. INTRODUCTION

The Constant Modulus Array (CMA) was introduced in 1986 by Gooch and Lundell [1], who suggest to apply the Constant Modulus (CM) criterion originally designed by Godard [2] to the field of adaptive antenna. Due to its interesting properties (low computational load, independence of the array manifold, etc.) it has become probably the most popular blind beamforming scheme. However, it is far from being free of drawbacks. One of its main inconvenient is its slow convergence, which sometimes makes it inapplicable in practical environments, specially when the Signal to Noise Ratio of the incoming signals is poor. This is the problem we address in our paper. As we will show, the analysis of the extrema of the cost function will reveal useful information about the location of the minima and its relationship with the signal subspace. This fact can be exploited to speed up the convergence of the algorithm.

The paper is organized as follows: in section 2 we review the constant modulus cost functions, introducing the appropriated vectorial notation; section 3 is devoted to analyze the nature of the solutions and its consequences. In section 4 we propose a novel technique exploiting the results of the performed analysis and having in mind to avoid some of the undesirable properties of the original proposal. Section 5 shows the relationship between the proposed technique and the Generalized SideLobe Canceller (GSLC). Simulation results of the suggested algorithm compared with other approximations are shown in section 6. Finally we end the paper by paying attention to specific scenarios where the proposed algorithm can be specially useful.

2. THE CONSTANT MODULUS ARRAY

The CMA is one of the simplest blind beamforming scheme. Although, as happens with Sato and Decision Directed (DD) algorithms, it can be seen as a particular case of the more general family of Bussgang algorithms, it was developed independently by Godard and later applied to array processing by Gooch and Lundell. The algorithm tries to minimize a nonconvex cost function designed to penalize deviations in the envelope of the array output signal:

$$J \stackrel{\Delta}{=} E\left\{\left(|y[n]|^2 - 1\right)^2\right\} \qquad (1)$$

being y[n] the output of the array, obtained as a linear combination of the received signals,

† This work has been partially supported by the National Research Plan of Spain CICYT, under Grant TIC-95-1022-C05-01

$$y[n] = \mathbf{w}^H \mathbf{x}[n] \quad (2)$$

where $\mathbf{x}[n]$ is the received snapshot and $\mathbf{w}$ is the weight vector, which describes the spatial response of the array. Superindex ^{H} hold for the hermitian operator, i.e.: transpose and conjugate.

The underlying idea under the mathematical formulation is very simple: the Source Of Interest (SOI) is assumed to have the constant envelope property[1]. This property is lost due to the contribution of noise and/or interfering processes to the output signal y[n]. The algorithm tries to indirectly remove noise and interferences by restoring the loss property.

The optimization procedure follows a single, LMS like, stochastic steepest descent algorithm which yields a weight vector adaptation equation given by:

$$\mathbf{w}[n+1] = \mathbf{w}[n] - \mu\left(\left|y[n]\right|^2 - 1\right) y^*[n]\mathbf{x}[n] \quad (3)$$

where μ is the step-size parameter, whose choice is a commitment between convergence speed and misadjustement noise. Probably the more useful result about the optimum value of μ is given by Katia and Duhamel[3], who suggest to select it as:

$$\mu = \mu_0 \frac{1}{\|\mathbf{x}[n]\|^2} \frac{1}{|y[n]|\left(|y[n]|+1\right)} \quad (4)$$

where μ_0 is the normalized step-size parameter, which must lie in the open (0,1]. The resulting algorithm is called the Normalized CMA (NCMA).

3. SOLUTIONS OF THE CONSTANT MODULUS COST FUNCTION AND ITS NATURE

In a general environment, the snapshot $\mathbf{x}[n]$ will be composed of two terms,

$$\mathbf{x}[n] = \sum_{m=1}^{M} \mathbf{d}_m s_m[n] + \mathbf{n}[n] \quad (5)$$

where M is the number of incoming signals, $\mathbf{d}_m$ is the generalized steering vector (including possible multipath effects), $s_m[n]$ represents the m-th signal and $\mathbf{n}[n]$ models the noise (usually thermal noise) generated at the sensors. Under this assumption, y[n] can be rewritten as:

$$y[n] = \sum_{m=1}^{M} \mathbf{w}^H \mathbf{d}_m s_m[n] + \mathbf{w}^H \mathbf{n}[n] = \sum_{m=1}^{M} g_m s_m[n] + n_y[n] \quad (6)$$

The set of definitions $g_m = \mathbf{w}^H \mathbf{d}_m$ (m = 1..M) and $n_y[n] = \mathbf{w}^H \mathbf{n}[n]$ is implicit in the above equation. Susbtituting (6) into (1) and manipulating the expression, we can finally find:

$$J = f\left(g_1 \ldots g_M, k_1 \ldots k_M, \sigma_1 \ldots \sigma_M, P_n\right) \quad (7)$$

where k_m (m = 1..M) represents the kurtosis[2] of the m-th signal, σ_m (m = 1..M) is the standard deviation of the m-th signal, and P_n is the noise power at the array output,

$$P_n = \mathbf{w}^H \mathbf{R}_{nn} \mathbf{w} \quad (8)$$

To determine the behavior of the CM algorithm we need to find the extrema of J, solving the following vectorial equation:

$$\nabla_{\mathbf{w}} J = \mathbf{0} \quad (9)$$

However, taking into account expression (7) and the relationship between the set of coefficients $(g_1 \ldots g_M, P_n)$ and the weight vector $\mathbf{w}$, it is possible to apply the chain rule to equation (9) obtaining:

$$\nabla_{\mathbf{w}} J = \sum_{m=1}^{M} \frac{\partial J}{\partial g_m} \nabla_{\mathbf{w}} g_m + \frac{\partial J}{\partial P_n} \nabla_{\mathbf{w}} P_n = \sum_{m=1}^{M} \frac{\partial J}{\partial g_m} \mathbf{d}_m + \frac{\partial J}{\partial P_n} \mathbf{R}_{nn} \mathbf{w} = \mathbf{0} \quad (10)$$

Equation (10) has two sets of solutions:

1. If $\partial J / \partial P_n = 0$, then the first term must also be zero, but if we assume that the generalized steering vectors $\mathbf{d}_m$ are linearly independent, the only valid solution is then given by $\partial J / \partial g_m = 0$ for all m. It is possible to demonstrate that this condition implies $g_m = 0$ for all m, and consequently, the weight vector associated to this solution lies completely in the noise subspace. Also, it is not difficult to demonstrate that this solution is a minimum if and only if all the incoming signals show a kurtosis larger than two, which is not the case for communications signals. Thus, this solution may be catalogued as an unwanted one.
2. If $\partial J / \partial P_n \neq 0$, then we can rewrite eq. (8) to read as follows:

$$\mathbf{w} = -\frac{\sum_{m=1}^{M} \frac{\partial J}{\partial g_m} \mathbf{R}_{nn}^{-1} \mathbf{d}_m}{\frac{\partial J}{\partial P_n}} = \sum_{m=1}^{M} c_m \mathbf{R}_{nn}^{-1} \mathbf{d}_m \quad (11)$$

[1] This property is shared by many manmade communications signals (i.e.: PSK and FSK modulations, among others).

[2] The kurtosis of a signal is defined as the quotient between its fourth order momentum and the squared second order momentum.

From observation of eq. (9) we can conclude that this set of solutions result in a linear combination of the eigenvectors associated to the signal subspace. A special case of interest is encountered under the classical hypothesis $\mathbf{R}_{nn} = \sigma_n^2\mathbf{I}$. In this situation, the desired solutions of the CM algorithm are linear combinations of the generalized steering vectors of the incoming signals.

4. THE PROPOSED ALGORITHM

From the performed analysis it is obvious that we can directly avoid unwanted solutions if the number of sources present in the signal scenario is "approximately" known. We have quoted the word "approximately" because our proposal does not need to know exactly the number of incoming sources. It is enough to provide information about the maximum expected number of simultaneous sources. We will denote this number by S ($S \geq M$).

Under this condition the proposed algorithm is summarized as follows:

1. Estimate $\mathbf{R}_{xx}=E\{\mathbf{x}[n]\mathbf{x}^H[n]\}$ from the received data.
2. Solve the generalized eigenvalue problem described by $\mathbf{R}_{xx}\mathbf{v}_i = \lambda_i\mathbf{R}_{nn}\mathbf{v}_i$ for all i=1..N, being N the number of sensors.
3. Extract the N-S less significant eigenvectors and form the new matrix:
$$\mathbf{V} = [\mathbf{v}_{S+1}, \mathbf{v}_{S+2}, \ldots, \mathbf{v}_N] \qquad (12)$$
where we assume that the eigenvectors $\mathbf{v}_i$ have been previously normalized in step 2.
4. Solve the linearly constrained problem:
$$\min_{\mathbf{w}} J = E\left\{\left(|y[n]|^2 - 1\right)^2\right\} \text{ subject to } \mathbf{w}^H\mathbf{R}_{nn}\mathbf{V} = \mathbf{0} \qquad (13)$$
where the introduction of a set of N-S restrictions will improve the convergence rate of the adaptive process.

The developed algorithm is designed to work over blocks of data rather than on a sample by sample basis, although it is possible, if required, to follow an adaptive procedure for obtaining the eigenvectors of the data correlation matrix [4].

5. REFORMULATION OF THE PROPOSED TECHNIQUE AND THE GSLC

Equation (13) yields a constrained optimization problem. A first approach is to employ the Frost algorithm, preprocessing all snapshots to remove components lying in the subspace spanned by V. However, this approach does not exploit the subspace rank reduction to reduce the number of computations. A general solution to the optimization of the Constant Modulus Cost Function given some linear restrictions was given by Griffiths[5]. However, in our special case, as all restrictions are equal to zero the formulation can still be simplified. By having in mind the structure of the GSLC beamformer, shown in figure 1, and taking into account how it works, it is clear that the simplest possible choice for $\mathbf{w}_0$ is:

$$\mathbf{w}_0 = \mathbf{0} \qquad (14)$$

avoiding the need to perform any computation related with the upper branch of the beamformer. The blocking matrix **B** must be orthogonal to the restrictions. If, under typical conditions, the noise is assumed to be uncorrelated between sensors, having equal power in all of them, the orthogonality condition for **B** can be written in terms of **V** as:

$$\mathbf{B}^H\mathbf{V} = \mathbf{0} \qquad (15)$$

Thus, the columns of **B** must lie in the signal subspace, and **B** can be chosen as:

$$\mathbf{B} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_S] \qquad (16)$$

where $\mathbf{v}_i$ is the i-th most significant eigenvector of $\mathbf{R}_{xx}$.

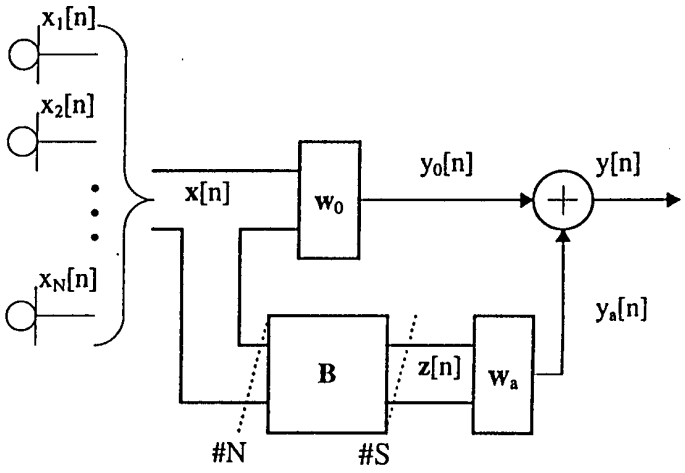


Figure 1 - Structure of the GSLC beamformer

Steps 3 and 4 of the algorithm proposed in section 4 must be modified according to the new formulation.

3. Extract the S more significant eigenvectors and form the matrix **B** as described in eq. (16).

$$\mathbf{B} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_S] \quad (17)$$

4. For every new snapshot, **x**[n],
   Project it over the signal subspace to obtain **z**[n],

$$\mathbf{z}[n] = \mathbf{B}^H \mathbf{x}[n] \quad (18)$$

   Update the weight vector $\mathbf{w}_a[n]$ following:

$$\mathbf{w}_a[n+1] = \mathbf{w}_a[n] - \mu_0 \frac{(|y[n]-1|)}{\|\mathbf{z}[n]\|^2} \frac{y^*[n]}{|y[n]|} \mathbf{z}[n] \quad (19)$$

where the normalized version of the CMA is preferred for the adaptation process.

6. SIMULATIONS

The signal scenario chosen is shown in table 1. The selected array is linear, having 30 omnidirectional sensors. Distance between two of them is half wavelength.

# of signal	Signal type	Angle of arrival	Input SNR
#1	4-PSK	30°	-5 dB
#2	8-PSK	0°	-5 dB
#3	Tone, f=0.1	20°	10 dB

Table 1 - Signal scenario for the simulations

In figure 2 we can observe the evolution of the output SINR for both algorithms, proposed and classical NCMA, when they are optimally initialized. The signal subspace method is several times faster than its unconstrained version.

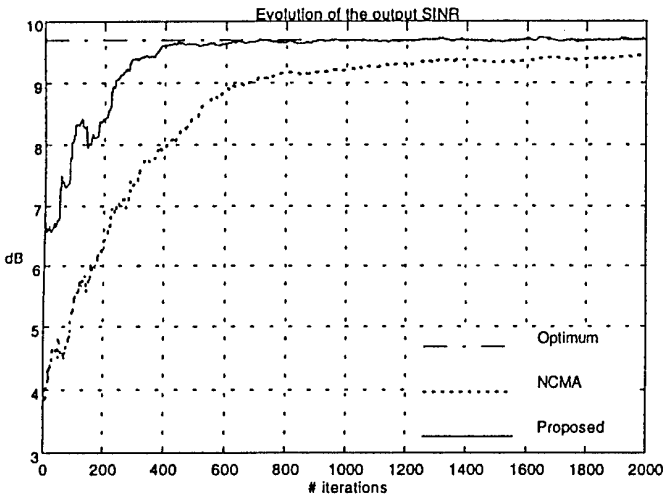


Figure 2 - Evolution of the output SINR for both algorithms: proposed vs. NCMA

It is difficult to notice, in the representation of the output SINR, the evolution of the weight vector and the final misadjustement. The proposed technique also achieves more precise reception patterns than the usual NCMA. This fact is shown in figure 3, where both algorithms are initialized to the optimum weight vector. The plot shows the error, computed as the distance between the optimum and the actual weight vectors for both algorithms.

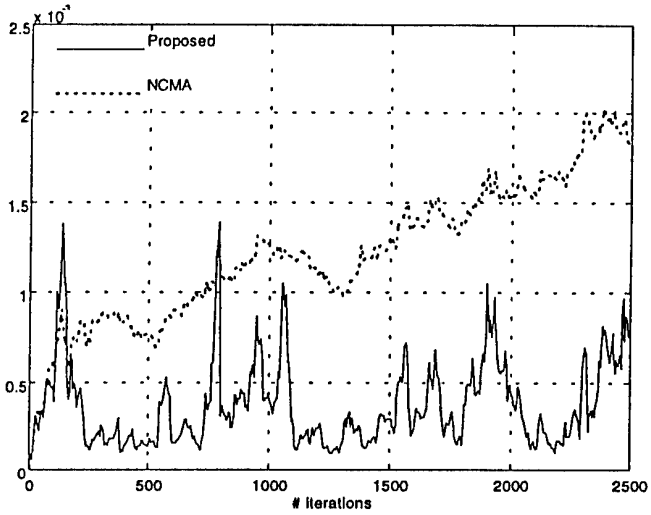


Figure 3 - Distance between the weight vectors and the optimum

6. CONCLUSIONS

Through the analysis of the solutions of the Constant Modulus Cost functions we have developed a modified version of the NCMA algorithm, which exploits the fact that the optimum weight vector lies on the signal subspace of the autocorrelation matrix of the snapshots. The proposed technique speeds up the evolution of the adaptive beamformer towards the optimum solution, showing better properties once the algorithm has converged.

Although the computational load of the proposed method is higher than in NCMA, there are several interesting cases where the suggested algorithm is specially useful:

- when the data set available for beamforming is small
- for arrays composed by a great number of sensors
- when convergence speed becomes a critical parameter in spite of computational load

In a forthcoming paper we will make a full detailed computational balance of both methods, obtaining

expressions for the excess of error introduced in each case.

7. REFERENCES

[1] Godard, D.N., "Self recovering equalization and carrier tracking in two- dimensional data communication systems", *IEEE Transactions on Communications*, vol. COM-28, pp. 1867-1875, Nov. 1980.

[2] Gooch, R.P. and Lundell, J.D., "The CM array: an adaptive beamformer for constant modulus signals", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Tokyo, pp. 2523-2526, Apr. 1986.

[3] Hilal, K. and Duhamel, P., "A convergence study of the constant modulus algorithm leading to a normalized CMA and a block-normalized CMA", *Proceedings of the European Signal Processing Conference*, Bruselas, Belgica, pp. 135-138, Aug. 1992.

[4] Yang, J.F., Kaveh, M., "Adaptive eigensubspace algorithms for direction or frequency estimation and tracking", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-36, pp. 241-251, Feb. 1988.

[5] Rude, M.J. and Griffiths, L.J., "A linearly constrained adaptive algorithm for constant modulus signal processing", *Proceedings of the European Signal Processing Conference*, vol. I., pp. 237-240, Barcelona, Sep. 1990.

IMPACT OF NON IDEAL FADING CHANNEL ESTIMATION IN A NARROW BAND SATELLITE MOBILE COMMUNICATION SYSTEM

Jan Erik Håkegård, Marie-Laure Boucheret

Ecole Nationale Supérieure des Télécomunications (ENST), site de Toulouse
10 Avenue Edouard Belin, BP 4004
31 028 Toulouse-FRANCE
E-mail: haakegaa@tlse.enst.fr, bouchere@tlse.enst.fr

ABSTRACT

The performance of the recently developed multi-user receiver and conventional receiver for narrow-band communication over fading channels with co-channel interference (CCI) and diversity is investigated with non perfect channel state information (CSI) estimation. The CSI estimates are derived using pilot symbols. Two estimation strategies are used, one based on interpolation, one on per-survivor-processing (PSP).
The performances of the receivers are assessed by computer simulations with uncoded and trellis-coded modulation. Simulations with perfectly coherent detection are also shown for comparison sake.
The results demonstrate that much of the advantage of the multi-user receiver is lost with this kind of CSI estimation. However, it still outperforms the conventional receiver, and it is less sensitive to the estimator used.

1. INTRODUCTION

In narrow-band communication with multi-beam coverage organization, frequency reuse (FR) is a key concept. The same frequency spectrum is used in different beams which are sufficiently spaced apart in order to maximize the number of users. The drawback of FR is that it introduces co-channel interference (CCI) which is generally the major source of impairment in cellular systems. Among the system solutions that have been proposed to counteract these channel impairment are channel coding, diversity and a combination of both.
Recently, some concepts from multi-user communication have been investigated in the context of coded transmission over fading channels with diversity (see e.g. [1],[2] and [10].) Until now, the channel state information (CSI), i.e. the fading process affecting the signals, has been assumed available. In a real system, the CSI has to be estimated, introducing errors which degrade the system performances.
In this paper we show how the performances are affected by the channel estimation using pilot assisted modulation. These methods have recently received much attention. Two different approaches are based on interpolation (see e.g. [3] and [4]) and on per-surviving processing (see [5], [6], [7] and [8]). The prior generally use only the periodically inserted known pilot symbols. The latter exploit both data and pilot symbols. It is based on a trellis algorithm and on linear prediction. It involves making separate fading estimates for the data sequences associated with the survivors. As a result, the channel estimator has two outputs; data decisions and corresponding channel estimates.
The paper is organized as follows. A description of the system model is given in section 2. The receiver structures are presented in section 3 and the channel state estimators in section 4. Some results from simulations are presented in section 5. Both coded and uncoded PSK are covered. Finally, the conclusions of this study are discussed in section 6.

2. SYSTEM MODEL

The block diagram of the transmitter is shown in Figure 1a. Random data are encoded and past through an interleaver. A multiplexer inserts known pilot symbols at the rate 1:nn, i.e. one pilot symbol for each group of $nn-1$ information bearing symbol. The resulting sequence is fed to the pulse shaping filter with unit impulse response and transmitted over M channels. We consider a q-PSK based communication system, i.e. the outputs of modulator takes on values from the set $\mathcal{X}_q = \{e^{j2i\pi/q} : i = 0, 1, \ldots, q-1\}$.
The transmitted signal $x(t)$ passes through M fading channels with CCI and additive white gaussian noise. The different channels are assumed uncorrelated, and

This work was in part supported by the Human-Capital and Mobility Program of the European Union.

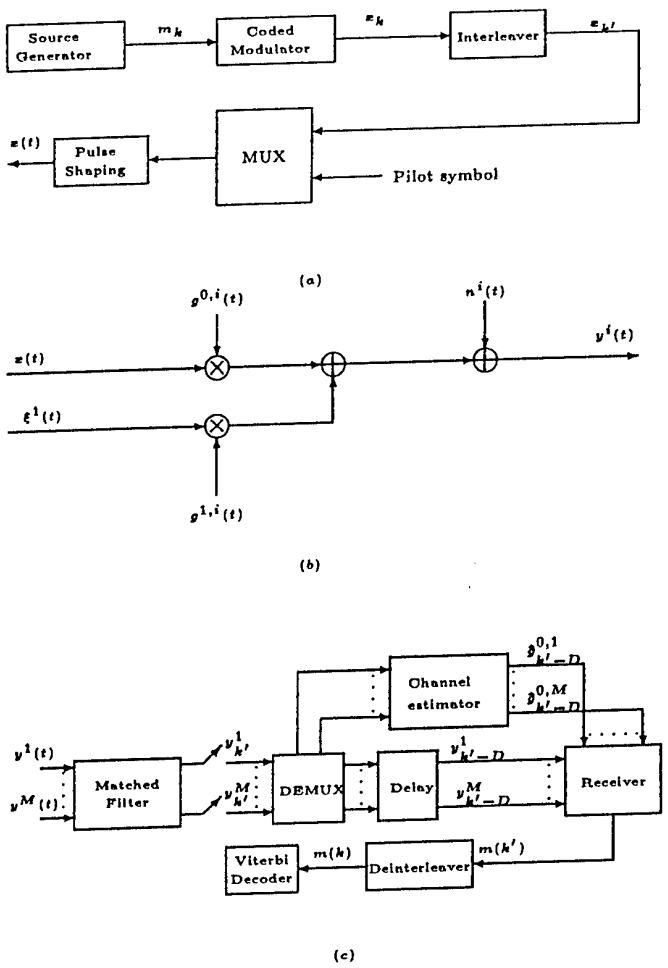


Figure 1: System model (a) The transmitter, (b) the channel and (c) the receiver

the CCI signals are assumed to be of the same kind as the useful signal. In Figure 1b a single signal is indicated for simplicity.

At the receiving end, the signals are captured by M antennas and goes through matched filters. The sampled M-vector output at time k can be expressed as:

$$\mathbf{y}_k = \mathbf{g}_k^0 x_k + \sum_{j=1}^{N} \mathbf{g}_k^j \xi_k^j + \mathbf{n}_k \tag{1}$$

where $\mathbf{n}_k$ is the additive noise and $\mathbf{g}_k^j$ is the fading affecting the useful ($j = 0$) and the interfering channels ($j = 1, \ldots, N$). These are complex Gaussian random vectors with mean zero and covariance $\frac{1}{2}\mathbf{I}_M$ and $\frac{1}{2}\gamma_j \mathbf{I}_M$, respectively (where $\mathbf{I}_\ell$ denotes the $\ell \times \ell$ identity matrix). ξ_k^j (for $j = 1, \ldots, N$) denotes the co-channel transmitted interfering signals. The useful signal is not affected by intersymbol interference (ISI). In this paper, we have assumed that the CCI is symbol synchronous with the useful signal.

We consider normalized diversity ([9]). This consists of splitting the total energy among the M diversity branches for the useful signal as well as for the interfering signals. We then have the expression:

$$\gamma_j = \Gamma_j / (M\sigma^2) \tag{2}$$

where Γ_j is the total average energy of the j-th signal. The signal-to-interference ratio (SIR) is for the jth interfering channel given by:

$$\beta_j = \gamma_0 / \gamma_j = \Gamma_0 / \Gamma_j \tag{3}$$

The received vector $\mathbf{y}_k$ can be written in a more compact form as:

$$\mathbf{y}_k = \mathbf{G}_k \mathbf{b}_k + \mathbf{n}_k \tag{4}$$

where $\mathbf{b}_k = (x_k, \xi_k^1, \ldots, \xi_k^N)^T$ and $\mathbf{G}_k$ is a $M \times (N+1)$ matrix whose j-th column is $\mathbf{g}_k^j$.

Depending on the channel estimator, the pilot symbols might be separated by a demultiplexer. This is only the case using an interpolator. Using the PSP estimator, the received signals follow two parallel lanes. One branch is fed to a channel state estimator which generates estimates $\hat{g}_k^{0,i}$ $(i = 1, \ldots, M)$ of the fading samples. In the case of the multi-user receivers, it is necessary to estimate the fading affecting the interfering signals as well. This is not shown in the figure for simplicity. In order to be able to obtain the estimates of the fading affecting the interfering signals, we assume that the SIR is inversed in the corresponding estimator. In practical terms, we might think of two beam-formers, each pointing at users in different locations.

The other branch is delayed to be in step with the channel state estimates, and after deinterleaving fed to the receivers together with the fading estimates. The outputs of the receiver are branch metrics which are fed to the Viterbi decoder.

3. RECEIVER STRUCTURES

3.1. CONVENTIONAL RECEIVER

This receiver is formed by a linear maximal ratio combiner followed by a branch-metric computer and by a Viterbi decoder matched to the coded modulation scheme chosen. The combined channel output ate instant k is given by [9]:

$$r_k = \left(\hat{\mathbf{g}}_k^0\right)^\dagger \mathbf{y}_k \tag{5}$$

where † denotes Hermitian transpose. The combined output r_k is fed to a metric computer, based on the Euclidian metric:

$$m(r_k, \widehat{x}_k) = 2\,\mathrm{Re}\,\{r_k \widehat{x}_k^*\} \tag{6}$$

which is maximum likelihood (ML) in the absence of CCI. The set of branch metrics $\{m(r_k, \widehat{x}_k),\ \widehat{x}_k \in \mathcal{X}_q\}$ is finally fed to the Viterbi decoder. This receiver requires CSI, timing and carrier phase recovery for the useful signal only.

3.2. MULTI-USER RECEIVER

This receiver structure needs the CSI, the timing and the carrier recovery for the interfering signals as well as for the useful signal. For each $a \in \mathcal{X}_q$, we define the set of $(N+1)$-vectors

$$\mathcal{S}(a) = \{\mathbf{b} = (a, \xi^1, \ldots, \xi^N)^T\} \tag{7}$$

Under the assumption of ideal interleaving, given the value of its first component a the random vector $\mathbf{b}$ is conditionally uniformly distributed over $\mathcal{S}(a)$. It is shown [10] that the ML branch metric for the Viterbi decoder is given by:

$$m(r_k, \widehat{x}_k) = \log \sum_{\widehat{\mathbf{b}}_k \in \mathcal{S}(\widehat{x}_k)} \exp(\Omega_k(\widehat{\mathbf{b}}_k)) \tag{8}$$

where:

$$\Omega_k(\widehat{\mathbf{b}}_k) = 2\,\mathrm{Re}\,\{\widehat{\mathbf{b}}_k^\dagger \mathbf{G}_k^\dagger \mathbf{y}_k\} - \widehat{\mathbf{b}}_k^\dagger \mathbf{H}_k \widehat{\mathbf{b}}_k, \tag{9}$$

and $\mathbf{H}_k = \mathbf{G}_k^\dagger \mathbf{G}_k$ is the instantaneous $(N+1)\times(N+1)$ correlation matrix of the fading vectors. Metric (8) can be well approximated, for sufficiently high SNR, by the simpler metric

$$m'(r_k, \widehat{x}_k) = \max_{\widehat{\mathbf{b}}_k \in \mathcal{S}(\widehat{x}_k)} \Omega(\widehat{\mathbf{b}}_k) \tag{10}$$

With perfect CSI, this receiver achieves the same BER curve slope as interference-free transmission.

4. CHANNEL ESTIMATION

4.1. INTERPOLATION

In a general manner, the estimated fading samples can be defined as the output of a filter bank:

$$\hat{g}_k = \sum_{i=-K+1}^{K} h_k^*(i) g_{i\cdot nn} = \mathbf{h}_k^\dagger \mathbf{g}_{nn}, 0 < k < nn \tag{11}$$

where $\mathbf{h}_k = (h_k(-K+1)\cdots h_k(K))^T$ are the filter coefficients, and $\mathbf{g}_{nn} = (g_{(-K+1)\cdot nn}\cdots g_{K\cdot nn})^T$ is the vector of the fading samples affecting the pilot symbols.

We consider two different interpolation techniques; sinc interpolation and optimal interpolation. The filter coefficients are given by:

$$\mathbf{h}_k^C = \mathbf{R}_{g\cdot nn}^{-1} \mathbf{w}_k \tag{12}$$

$$\mathbf{h}_k^O = (\mathbf{R}_{g\cdot nn} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{w}_k \tag{13}$$

respectively, where $\mathbf{R}_{g\cdot nn} = E[\mathbf{g}_{nn}\mathbf{g}_{nn}^\dagger]$ and $\mathbf{w}_k = E[\mathbf{g}_{nn} g_k^*]$. We see that for the optimal interpolator, we need to estimate the autocorrelation function of the fading samples affecting the pilot symbols. For the sinc interpolator, this is not the case. However, the length of the filters are infinite. In the simulations, the filters are truncated to the length 512, giving negligible degradation in performance.
In the calculation of the optimal interpolator, we need the noise variance as well. As pointed out in [4], the CCI can be approximated as Gaussian noise. The resulting σ^2 is therefore the sum of the interference power and the additive Gaussian noise.

4.2. PER-SURVIVOR-PROCESSING (PSP)

Recently, channel estimation using per-survivor-processing (PSP) has received some attention [6], [7]. Several variations has been proposed. In this paper we limit ourselves to the technique called Basic Algorithm (BA) in [6].
The estimates of the channel state are obtained using linear prediction and tentative symbol decisions:

$$\hat{g}_k^i = \sum_{n=1}^{L} c_n y_{k-n} \hat{x}_{k-n}^i, i = 0, \ldots, q-1 \tag{14}$$

where $\mathbf{c} = (c_0 \cdots c_{L-1})^T$ minimize the mean square estimation error. They must satisfy the normal equations [12]:

$$\sum_{i=0}^{L-1} \mathbf{R}_g(m-1)c_i + \sigma_n^2 c_m = \mathbf{R}_g(m+1), 0 \le m \le L-1 \tag{15}$$

where $\mathbf{R}_g(k)$ is the autocorrelation function of the fading samples. The index i in (14) corresponds to one particular survivor in the trellis. The $\hat{g}_k^i$'s are then used in the branch metric computation:

$$m(x_{k-1}^i \to x_k^j) = |y_k - \hat{g}_k^i \hat{x}_k^j|^2 \tag{16}$$

where $m(x_{k-1}^i \to x_k^j)$ denote the branck metric from state i to the state j in the trellis.

Instead of using the channel estimates directly, we use the tentative symbol decisions and derive new channel estimates from them. This is done with a Wiener

filter:

$$\hat{g}_k^i = \sum_{n=-K}^{K} d_n y_{k-n} \hat{x}_{k-n}^i, i = 0, \ldots, q-1 \qquad (17)$$

where d_n satisfy the equations:

$$\sum_{i=-K}^{K} \mathbf{R}_g(m-1)d_i + \sigma_n^2 d_m = \mathbf{R}_g(m+1), -K \leq m \leq K \qquad (18)$$

The reason why this is done, is to involve an interpolation over past and future channel measurements and not only a prediction from past measurements.
We have assumed that the predictor/interpolator coefficients can be optimally chosen according to (15) and (18). As they depend on the noise variance and the autocorrelation function of the fading which are non-stationary, these parameters have to be estimated as a function of time. Solutions to this problem are indicated in [6]. The PSP algorithm implemented performs well for QPSK signals. For 8PSK signals the performance is not so good. This is due to the fact that the error rate in the tentative symbol decisions is higher, perturbing the prediction of the fading samples. Other more complex algorithms are reported to work well even on 8PSK. They have a more complex trellis structure, more complex branch metric computation or a combination of both.

5. SIMULATION RESULTS

In this section we report on the bit-error-rate (BER) performance of the conventional and the multi-user receiver with non-perfect CSI. We compare these results with the ones corresponding to perfect coherent detection (perfect CSI). The two cases of uncoded QPSK and coded 8PSK are considered. The coding used is Ungerboek's 8PSK TCM scheme with 8 states. The diversity order is 2, and the Doppler bandwidth $B_dT = 0.01$. The frame length nn is 5, giving a loss of effective $E_b/N0$ of 0.97 dB. This has been taken into account by shifting the BER curves 0.97 dB to the right. In order to obtain a benchmark for the performances, the curves corresponding to perfect CSI are shifted to right as well. When coding is implemented, a block interleaver of size 15×15 is used.
For the uncoded case, the signal to interference ratio (SIR) is set to 20 dB. The estimators used are the sinc interpolator and the PSP estimator. The performances of the two receivers are very close for small E_b/N_0 in the ideal case (see Figure 2 and 3). However, the conventional receiver exhibits an error-floor, which is not the case for the multi-user receiver. This is true for any finite SIR. With non perfect CSI estimation, both receivers exhibit error-floors. With the PSP-estimator, the performances are very close for the two receivers. We see however that the conventional receiver is more sensitive to the estimator. The error-floor of conventional receiver is twice as high as the one of the multi-user receiver using the sinc interpolator.
For the coded case, the SIR is 10 dB. The estimators used are the sinc interpolator and the optimal interpolator. With the optimal interpolator, the performances of the two receivers are very close (see Figure 4 and 5. With the sinc interpolator however, the multi-user receiver performs much better, showing that it is less sensitive to bad channel estimates.

6. CONCLUSIONS

The performances of the conventional and the multi-user receiver for PSK signals transmitted over flat Rayleigh channels with CCI and diversity have been shown with non perfect CSI estimation using pilot symbols. For perfect CSI the multi-user receiver outperforms the conventional receiver and is especially efficient for low SIRs. The simulation results show that much of this advantage is lost when the CSI is estimated using the proposed techniques. It can be noted however, that the multi-user receiver is less sensitive to bad channel estimates than the conventional receiver.
It is clear that the channel estimates using pilot symbols will be poor for low SIRs. This is due to the fact that the interference is seen as additive Gaussian noise by the CSI estimator. One way to avoid this problem is to use non overlapping pilot tones. This strategy introduces other problems as spectral lines in the transmission band, but should be investigated for this kind of systems.

7. REFERENCES

[1] J. Ventura-Traveset, G. Caire, and E. Biglieri, "A multi-user approach to combating co-channel interference in narrowband mobile communications". *The 7th Tyrrhenian International Workshop on Digital Communications*, Italy, Sept., 1995.

[2] E. Biglieri, G. Caire, J. E. Håkegård, G. Taricco, J. Ventura-Traveset, "System capacity enhancement by using multi-user processing techniques in narrowband satellite mobile communications." *IEE 5th International Conference on Satellite Systems for Mobile Communications and Navigations*, London, May 1996.

[3] J. K. Cavers, "An Analysis of Pilot Symbol Assisted Modulation for Rayleigh Fading Channels", *IEEE Trans. Veh. Technol.*, vol. 40, No. 4, Nov. 1991.

[4] J. K. Cavers, "Cochannel Interference and Pilot Symbol Assisted Modulation",*IEEE Trans. Veh. Technol.*, vol. 42, No. 4, Nov. 1993.

[5] R. Raheli, A. Polydoros, C. K. Tzou, "Per-Survivor Processing: A General Approach to MLSE in Uncertain Environments", *IEEE Trans. Commun.*, vol. 43, No. 2/3/4, Feb./Mars/April 1995.

[6] A. N. D'Andrea, A. Diglio, U. Mengali, "Symbol-Aided Channel Estimation With Nonselective Rayleigh Fading Channels", *IEEE Trans. Veh. Technol.* Vol. 44, No. 1, Feb. 1995.

[7] G. M. Vitetta, D. P. Taylor, "Maximum Likelihood Decoding of Uncoded and Coded PSK Signal Sequences Transmitted over Rayleigh Flat-Fading Channels",*IEEE Trans. Commun.*, vol. 43, No. 11, Nov. 1995.

[8] G. M. Vitetta, D. P. Taylor, "Multisampling Receivers for Uncoded and Coded PSK Signal Sequences Transmitted Over Rayleigh Frequency-Flat Fading Channels",*IEEE Trans. Commun.*, vol. 44, No. 2, Feb. 1996.

[9] J. Ventura-Traveset, G. Caire, E. Biglieri and G.Taricco, "Impact of diversity reception on fading channels with coded modulation. Part I: Coherent detection," submitted to *IEEE Trans. Commun.*, Sept. 1995.

[10] J. Ventura-Traveset, G. Caire, E. Biglieri and G.Taricco, "Impact of diversity reception on fading channels with coded modulation. Part III: Co-Channel Interference," submitted to *IEEE Trans. Commun.*, Nov. 1995.

[11] J. Ventura-Traveset, G. Caire, E. Biglieri and G.Taricco, "A multi-user approach to narrowband cellular communications," submitted to *IEEE Trans. Inf. theory*, Nov. 1995.

[12] C. Therrien, *Discrete Random Signals and Statistical Signal Processing.* Englewood Cliffs, NJ: Prentice Hall, 1992.

[13] G. Ungerboeck, "Channel coding with multilevel/phase signals", *IEEE Trans. Inf. Theory*, Vol. IT-28, Jan. 1982.

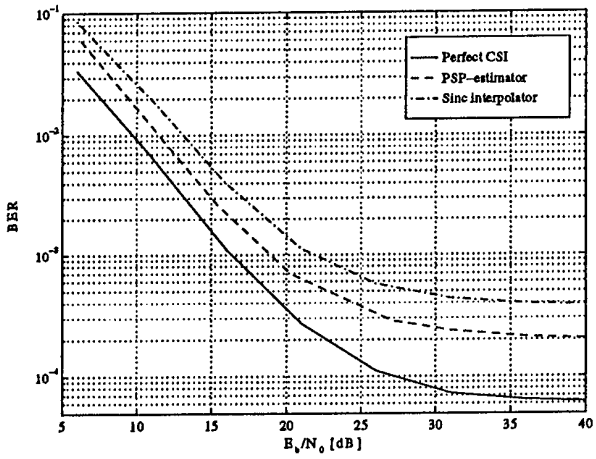


Figure 2: BER for the conventional receiver with uncoded QPSK, SIR=20 dB, $B_dT = 0.01$ and $nn = 5$.

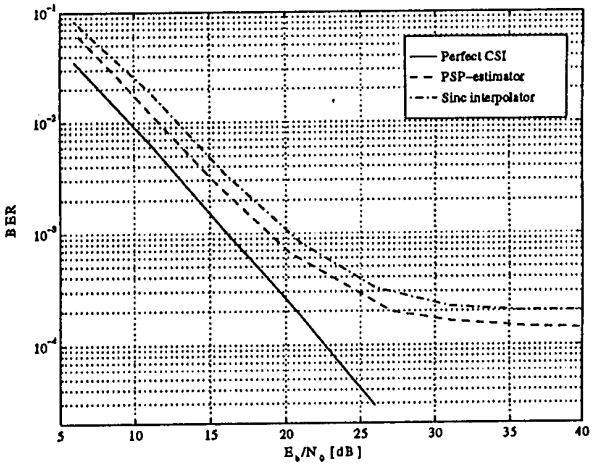


Figure 3: BER for the multi-user receiver with uncoded QPSK, SIR=20 dB, $B_dT = 0.01$ and $nn = 5$.

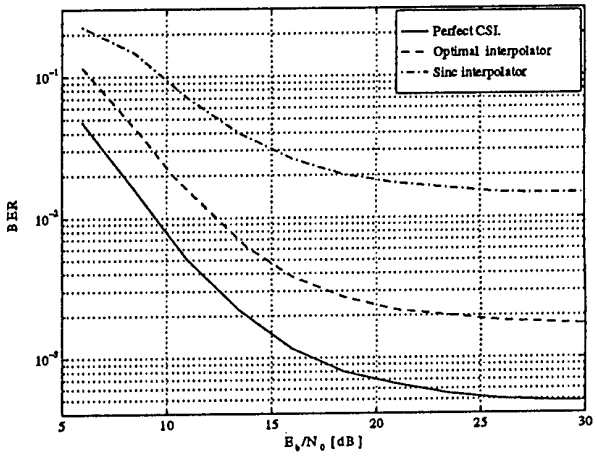


Figure 4: BER for the conventional receiver with coded 8PSK, SIR=10 dB, $B_dT = 0.01$ and $nn = 5$.

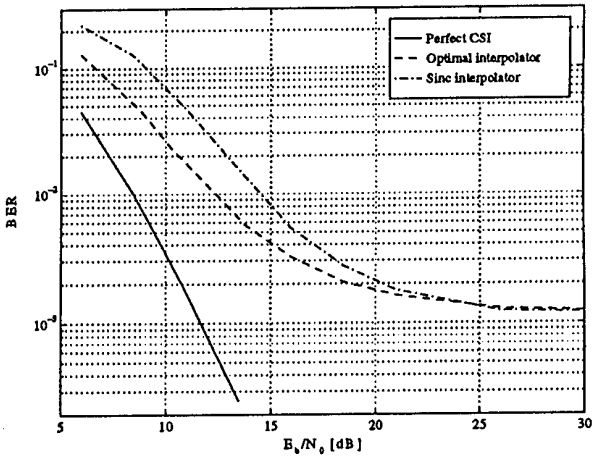


Figure 5: BER for the multi-user receiver with coded 8PSK, SIR=10 dB, $B_dT = 0.01$ and $nn = 5$.

THE SURFACE MODELLING CAPABILITIES AND LIMITATIONS OF CMAC AND GCMAC

Francisco J. González-Serrano[1] *and Antonio Artés-Rodríguez*[2]

[1]DTC - ETSI Telecomunicación. Universidad de Vigo. 36200 VIGO-SPAIN
e-mail : frank@tsc.uvigo.es
[2]DSSR - ETSI Telecomunicación. UPM. Ciudad Universitaria. 28040 MADRID-SPAIN.
e-mail : antonio@gtts.ssr.upm.es

ABSTRACT

The Cerebellar Model Articulation Controller (CMAC) is attracting a great deal of interest due to its on-line rapid learning, generalisation properties and simplicity. In this paper an analysis of the abilities of CMAC and the Generalized GCMAC (GCMAC) to approximate an arbitrary input/output mapping is addressed. An expression for the dimension of the null space spanned by the set of functions provided by the GCMAC is derived. From this expression it is possible to measure the space of functions that can be exactly modelled by the GCMAC. A set of local restrictions on the surfaces that can be exactly approximated is given. As it was expected, the local restrictions imply that only smooth surfaces are adequately stored.

1. INTRODUCTION

The CMAC network [1] was proposed as a control method based on the principles of the cerebellum's motor behavior. The simplicity of the CMAC learning algorithm and its ability to generalise from sparse input-output data pairs has led to its widespread use in many engineering applications, e.g. prediction of chaotic time series [2], real-time robotic control [3] and nonlinear deconvolution [4]. However, in many applications in digital communications such as data predistortion [5, 6], electrical echo cancellation [7] or nonlinear equalisation, the accuracy in the approximation is a more important factor than generalisation. Consequently, evaluation of the function representation capabilities of CMAC is an important issue. Previously, only preliminary studies about the interpolation capabilities of CMAC have been reported [8]. One approach is to investigate the hyper-surfaces that can, and cannot, be modelled exactly. Thus, in this paper we present a study of the dimension of the space spanned by the basis functions of the Generalized CMAC (GCMAC) [9]. This can be considered as a bound on the space of the range of possible functions which GCMAC can model. Also, we identify certain features of the functions which the GCMAC is unable to model.

2. DISCRETE HYPER-SURFACE APPROXIMATION

General approximation theory deals with the problem of approximating a multivariate function $\underline{z} = \underline{\mathbf{F}}(\underline{x})$ by a function $\hat{\underline{z}} = \underline{\mathbf{H}}(\underline{w}, \underline{x})$ having a fixed number of parameters (weights), $\underline{w}$. Spline interpolation and many approximation schemes, such as expansions in series of orthogonal polynomials, are included in this representation. In particular, the CMAC network provides a set of basis functions which, when suitably adjusted in amplitude, approximate the desired hyper-surface.

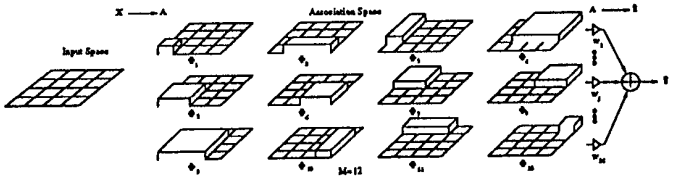


Figure 1: CMAC Input/Output Mapping using binary basis functions.

The GCMAC internally transforms every training input into a higher dimensional space so that the desired output can be made approximately linear to the transformed input. In this way, the output of GCMAC is formed by a linear combination of overlapping basis functions which are distributed in an n-dimensional subspace of $\mathcal{Z}^n$,

$$\hat{z} = \sum_{j=1}^{\rho} w_{c(j)} \Phi_{c(j)} \qquad (1)$$

where $c(j)$ contains the indices of the local functions activated by the input vector (e.g., in Fig. 2 if $\underline{x} = (1,1) \rightarrow \underline{c} = (1,5,9)$).

Since the approximation used by the GCMAC network is linear in the unknown coefficients $\underline{w}$, there always exists a choice of weights that approximates $\underline{\mathbf{F}}(\bullet)$ better than all other possible choices [10]. In this sense, simple instantaneous learning laws can be used, for which convergence can be established subject to well-understood restrictions.

3. THE GCMAC SCHEME : DEFINITION OF THE SET OF BASIS FUNCTIONS

The set of basis functions shown in Fig. 2 can be geometrically decomposed into a set of ρ_{MAX} overlays. An overlay is defined as the union of basis functions with non-overlapping hyper-rectangular receptive fields which exactly covers the discretized input space. The size of hyper-rectangles is defined by the generalisation vector $\underline{\rho} = [\rho_1, \ldots, \rho_n]$, being $\rho_{MAX} = \max(\underline{\rho})$. These overlays have different partitioning of the receptive fields so that the same input maps to different elements of the set of basis functions in different overlays. An example of the overlay structure is given in Fig. 2

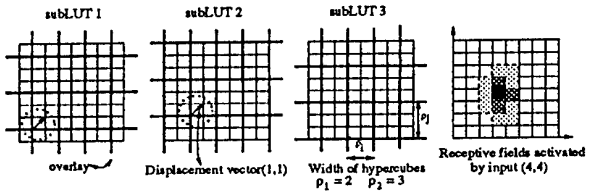


Figure 2: CMAC overlay structure: $\underline{\rho} = [2, 3]$

The generalisation vector $\underline{\rho}$ significantly affects the approximation capability and the rate of convergence of the network. When elements of $\underline{\rho}$ are chosen too large, the network is slow to learn a function containing high spatial Fourier components. On the other hand, when the elements of $\underline{\rho}$ are chosen too small, the network is unable to generalize between neighboring training samples. An heuristic rule to determine the generalisation vector can be stated as follows. A small width in the receptive fields must be used when the function varies significantly, while a larger one should be used when the function is approximately constant. In addition, the previous strategy implies a substantial improvement of the rate of convergence. As an example, Fig. 3.a shows a deterministic surface and Fig. 3.b some of the functions employed in the approximation with a generalisation vector $\underline{\rho} = [2, 15]$. The convergence rate obtained with different configurations of GCMAC is shown in Fig. 4.

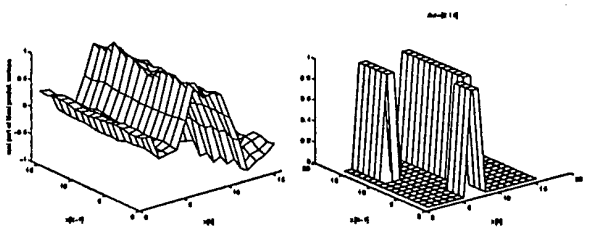


Figure 3: a) Original surface. b) Some basis functions of a GCMAC with $\underline{\rho} = [2, 15]$

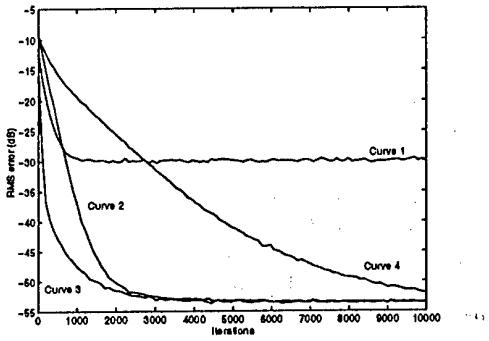


Figure 4: Curve 1: Standard CMAC ($\underline{\rho} = [15, 15]$). Curve 2: Standard CMAC ($\underline{\rho} = [2, 2]$). Curve 3: Proposed GCMAC ($\underline{\rho} = [2, 15]$). Curve 4: Proposed GCMAC ($\underline{\rho} = [15, 2]$). Traces are ensembled average of 15 convergence curves.

From Fig. 4 it can be inferred that the approximation capabilities of GCMAC improve when the generalisation vector is reduced. In particular, for a two dimensional input space, the approximation capabilities of GCMAC only depends on the minimum element (ρ_{MIN}) of the generalisation vector whereas the position of ρ_{MIN} within $\underline{\rho}$ does not affect to the final approximation error but to the rate of convergence.

4. REPRESENTATION CAPABILITIES OF GCMAC

Many neural networks can approximate, under some typically very reasonable conditions, an input-output map arbitrarily well (given infinite resources). In this work, we are interested in answering to the next question: for a given GCMAC structure (finite resources), what kind of functions can be modelled exactly? In general, GCMAC cannot reproduce an arbitrary multivariate Look-Up-Table (LUT) $\underline{\mathbf{z}} = \underline{\mathbf{F}}(\underline{\mathbf{x}})$. This can be considered as an upper bound in the approximation ca-

pabilities of the GCMAC. Brown *et al.* [8] have proved that linearly combined univariate piecewise constant LUTs ($\underline{z} = \sum_{i=1}^{n} F_i(x_i)$) can be exactly approximated by the CMAC, for any value of $\underline{\rho}$. This class of functions are the lower bound of the set of functions which are exactly modelled by a GCMAC. Between these two bounds is the answer to our question.

In discussing the modelling capabilities of GCMAC it is necessary to answer the question of whether a multivariate function can, or cannot, be exactly represented. For this reason, it is useful to specify the number of degrees of freedom wasted for select the best set of weights or, in other words, the number of linearly dependent basis functions of GCMAC. Once this bound on the approximation capabilities is set, the next step is to derive a set of relationships which must exist in the data for an exact representation. Having described the set of restrictions in the input data, it is then possible to construct functions which fail to satisfy any of the restrictions. These functions are said to be orthogonal to each of the basis functions of GCMAC.

5. SPACE OF FUNCTIONS

A matrix interpretation of GCMAC Input/Output mapping is given by:

$$\underline{\underline{A}}\,\underline{w} = \underline{\hat{z}}, \tag{2}$$

where $\underline{\underline{A}}$ is a $N \times M$ matrix for which each row is the assigned association vector, $\underline{a}_c^T$, to every input vector $\underline{x}$, N is the cardinality of the set of input integer vectors (if the number of integer values along i dimension is L_i, then $N = \prod_{i=1}^{n} L_i$), and M is the number of basis functions of the GCMAC. Starting with the matrix $\underline{\underline{A}}$, which has entries in $\mathcal{B} = [0, 1]$, one may want to know how many columns (or rows) of this matrix are non-parallel or independent of each other. The GCMAC solution matrix $\underline{\underline{A}}$ is rank-deficient when $\rho_i \geq 2$ [8] . This means that the set of functions provided by the GCMAC is linearly dependent. Assuming that the rank of $\underline{\underline{A}}$ is $R \leq M \leq N$, a unitary matrix $\underline{\underline{U}}$ can be chosen in $\mathcal{R}^{N \times N}$ such that the R-dimensional column space of $\underline{\underline{A}}$ is spanned by a subset of R columns of $\underline{\underline{U}}$, say the first R columns, which together form the matrix $\underline{\underline{\hat{U}}}$, then

$$\underline{\underline{U}} = \left(\underline{\underline{\tilde{U}}}\ \underline{\underline{\tilde{U}}}^{\perp}\right) \tag{3}$$

Since $\underline{\underline{U}}$ is unitary, any vector $\underline{z}$ can be decomposed into two mutually orthogonal vectors $\underline{\tilde{z}}$ and $\underline{\tilde{z}}^{\perp}$ in the spaces spanned by the columns of $\underline{\underline{\tilde{U}}}$ and their orthogonal complement $\underline{\underline{\tilde{U}}}^{\perp}$. In this sense, the space of functions that can be exactly modelled by the GCMAC is spanned by the columns of matrix $\underline{\underline{\tilde{U}}}$, and the functions that cannot be modelled by the GCMAC are in the space spanned by the columns of $\underline{\underline{\tilde{U}}}^{\perp}$.

Same comments hold for the row space of matrix $\underline{\underline{A}}$, and a unitary matrix $\underline{\underline{V}}$, of size $M \times M$, can be similarly found and decomposed into two orthogonal matrices:

$$\underline{\underline{V}} = \left(\underline{\underline{\tilde{V}}}\ \underline{\underline{\tilde{V}}}^{\perp}\right) \tag{4}$$

The columns of $\underline{\underline{\tilde{V}}}^{\perp}$ span the null space (or kernel) of $\underline{\underline{A}}$, i.e., the space of weight vectors $\underline{w}$ for which $\underline{\underline{A}}\,\underline{w} = \underline{0}$.

5.1. DIMENSION OF THE NULL SPACE OF GCMAC

The dimension of the null space of the linear operator $\underline{\underline{A}}$ determines the number of degrees of freedom wasted for the choice of the best set of weights. The computation of the dimension of the null space of $\underline{\underline{A}}$ is tedious and strongly dependent on the parameters which specify the GCMAC and, in particular, on the generalisation vector $\underline{\rho}$ and the number of levels in each axis, L_i. As an example, for $n = 2$ (two-dimensional input space), and a generalisation vector $\underline{\rho} = [\rho_1, \rho_2]$ $(\rho_1 \leq \rho_2)$, the dimension is given by:

$$\dim(\ker(\underline{\underline{A}})) = \sum_{j=1}^{\rho_2 - \rho_1} \left(\lceil \frac{L_1 - d_{1,j}}{\rho_1} \rceil \right) + \rho_2 - 1, \tag{5}$$

where $d_{1,j} = \mathrm{mod}(j-1, \rho_1) + 1$. When $\rho_1 = \rho_2 = \rho$ (standard CMAC), the dimension of the null space is $\rho - 1$. This result is in agreement with the fact that the number of common weights shared by adjacent input points is just $\rho - 1$.

5.2. DEPENDENCE RELATIONSHIPS OF THE DATA

Once the number of linearly dependent basis functions of GCMAC is determined, the rank of $\underline{\underline{A}}$ can be evaluated:

$$R = \mathrm{rank}(\underline{\underline{A}}) = M - \dim(\ker(\underline{\underline{A}})). \tag{6}$$

It is evident that the GCMAC is only able to exactly model hyper-surfaces with a certain number of restrictions on their values. In particular, there will exist $N - \mathrm{rank}(\underline{\underline{A}})$ non-redundant equations which specify the relationships which must satisfy the desired function to be approximated. For clarity, a geometric interpretation of the previous relationships is presented below.

5.2.1. GEOMETRIC INTERPRETATION

Consider an n-dimensional lattice which is composed of n-dimensional hyper-rectangles (receptive fields). The width of the hyper-rectangles along the ith axis is ρ_i. Let $\underline{F}(\underline{x})$ a function which has a zero value for all the inputs lying outside of the squared block composed of

four adjacent points of the n-dimensional input space as shown in Fig. 5.

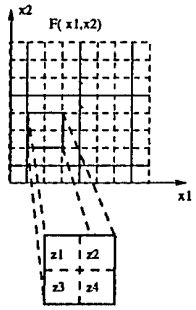


Figure 5: A two-dimensional block of data.

Depending on the relative position of the block within the lattice two possible situations can occur:

1. When both $(n-1)$-dimensional hyper-planes which cut the block lie on the same overlay (the block lies on a knot of the lattice), there are enough degrees of freedom to choose the weights of the GCMAC. Therefore, the GCMAC is able to exactly represent the desired output for the region defined by the block (see Fig. 6).

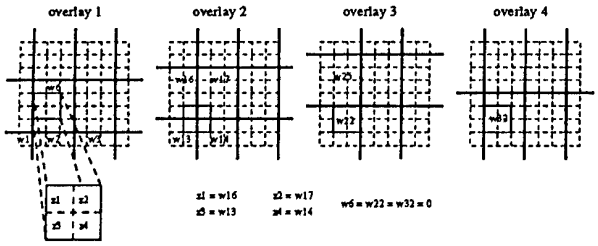


Figure 6: A two-dimensional block of data lying on a knot of the GCMAC overlay structure. $\underline{\rho} = [3,4]$, $L_i = [8,8]$.

2. However, when each of the two $(n-1)$-dimensional hyper-planes which cut the block lie on different overlays, the GCMAC is only able to exactly model the desired hyper-surface when certain restrictions in the data are imposed. The outputs corresponding to the block showed in figure 7 are given by:

$$z_1 = w_6 + w_{16} + w_{25} + w_{32}$$
$$z_2 = w_6 + w_{17} + w_{25} + w_{32}$$

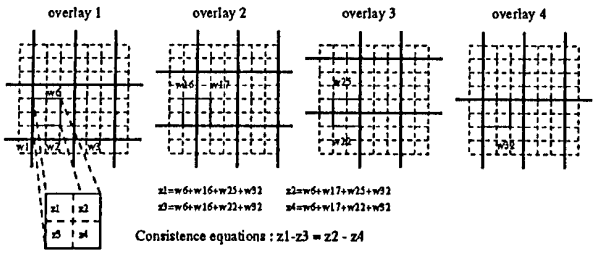


Figure 7: The block of data does not lie on a knot of the GCMAC overlay structure.

$$z_3 = w_6 + w_{16} + w_{22} + w_{32}$$
$$z_4 = w_6 + w_{17} + w_{22} + w_{32} \qquad (7)$$

From this expression, the following relationship must be verified by the desired output:

$$z_1 - z_3 = z_2 - z_4 \Leftrightarrow z_1 + z_4 = z_2 + z_3 \qquad (8)$$

The number of blocks which are not corner-aligned with the knots of the lattice determines the number of restrictions required to exactly model the desired hyper-surface. This number is equal to $N - \text{rank}(\underline{\underline{A}})$.

It is important to note that when the desired hyper-surface does not satisfy the equation 8, the GCMAC is not able to exactly model the function, i.e. some amount of error is produced when GCMAC performs the approximation (the GCMAC reproduces the desired function in a least squares sense). The restrictions required to exactly model a specific hyper-surface can be roughly condensed in the following heuristic rule: the slope of the hyper-surface at adjacent points (difference between neighbor points) must be (nearly) the same. This rule has an evident connection with the well known smoothness hypothesis required for performing general function approximation.

5.3. ORTHOGONAL FUNCTIONS

As stated above, the GCMAC is only able to reproduce hyper-surfaces with certain restrictions. Restrictions stem from the overlap between the receptive fields of the basis functions. It is obvious that, when the smoothness hypothesis is violated, the GCMAC will be unable to exactly reproduce the required hyper-surface. For this reason, it seems reasonable that the functions orthogonal to the basis functions of GCMAC, must have large spatial-frequency components. Following this line, we have found certain conditions that the orthogonal functions must verify. If we construct a function which has a zero value for all the inputs lying

outside the block shown in figure 3 and with values at the input points lying inside the block satisfying

$$z_1 = z_4 = 1\ ,\ z_2 = z_3 = -1, \qquad (9)$$

then the GCMAC is unable to reproduce the desired function. After some straightforward algebraic manipulations it is easily derived that the best (in a least squares sense) vector of weights is identically zero, giving the GCMAC a null output. Whenever a block satisfies equation 9, and whenever it is placed completely within the lattice, such that the two $(n-1)$-dimensional hyper-planes which cut the block do not lie on the same overlay, the GCMAC will be unable to represent the desired function. Any linear combination of two or more functions like those specified in equation 9, will also be orthogonal to the basis functions of GCMAC.

If the degree of generalisation is reduced along some determined direction, then the knot density on the lattice is increased and, therefore, the number of orthogonal functions to the basis functions of GCMAC is reduced. For particular choices of the generalisation vector $\underline{\rho}$, the number of function basis, M, may be greater than the number of input vectors, N. In this case, the linear system described in equation 2 is underspecified, that is, it has more unknowns (weights) than equations. In this context, the rank of $\underline{\underline{A}}$ is equal to N and the GCMAC exactly models the desired hypersurface. Even though the number of adjustable weights is greater than N, i.e. there are more weights than in a full Look-Up-Table (LUT), the generalisation capability of GCMAC speeds up the rate of convergence with respect to the (less complex) LUT.

6. CONCLUSIONS

This paper has considered some features of the class of functions which the Generalized CMAC can and cannot model. As the GCMAC network generalises, a multivariate LUT cannot be exactly approximated. A set of restrictions of the hyper-surface were derived for a perfect approximation. These restrictions can be interpreted as a simple condition of smoothness. Therefore, the GCMAC network, like the CMAC, is only able to represent smooth hyper-surfaces. However, the degrees of freedom of the hyper-surfaces that can be modelled by GCMAC are greater than those approximated by the standard CMAC. This result stems from the flexibility gained with a variable generalisation vector. Finally, a set of orthogonal functions to the basis functions of GCMAC were found.

7. REFERENCES

[1] J. Albus, "A New Approach to Manipulator Control: the Cerebellar Model Articulation Controller," *Journal on Dynamic Systems, Measurements and Control*, pp. 220–227, September 1975.

[2] J. Moody, "Fast learning in multi-resolution hierarchies," *Advances in Neural Information Processing*, pp. 29–39, 1989.

[3] W. T. Miller, F. H. Glanz, and L. G. Kraft, "CMAC: An associative Neural Network alternative to Backpropagation," *Proc. of IEEE*, vol. 78, pp. 1561–1567, Oct 1990.

[4] F. H. Glanz and W. T. Miller, "Deconvolution and Nonlinear Inverse Filtering using a Neural Network," in *Proc. of Int. Conf. on Accoust., Speech and Signal Processing*, vol. E7.3, pp. 2349–2352, IEEE, 1989.

[5] A. Artés-Rodríguez, F. González-Serrano, A. Figueiras-Vidal, and L. Weruaga-Prieto, "Compensation of Bandpass Nonlinearities by Look-Up-Tables and CMAC," in *Proceedings of the International Workshop on Applied Neural Networks to Telecommunication*, IWANNT, May 1995.

[6] F. J. G. Serrano, A. A. Rodríguez, and A. R. F. Vidal, "A Fast LUT+CMAC Data Predistorter," in *Proc. of the VIII European Signal Processing Conference. EUSIPCO'96*, (Trieste, Italy), EURASIP, Sept 1996.

[7] L. W. Prieto, *Cancelación no Lineal de Ecos en Transmisión Digital.* PhD thesis, E.T.S.I. Telecomunicación - Universidad Politécnica de Madrid, 1994.

[8] M. Brown, C. J. Harris, and P. C. Parks, "The Interpolation Capabilities of the Binary CMAC," in *Neural Networks*, pp. 429–440, Pergamon Press Ltd., 1993.

[9] F. J. G. Serrano, A. A. Rodríguez, and A. R. F. Vidal, "The Generalized CMAC," in *Proc. of International Symposium on Circuits and Systems*, (Atlanta), IEEE, May 1996.

[10] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, pp. 1481–1497, 1990.

CHANNEL IDENTIFICATION AND FAILURE DETECTION IN DIGITAL SATELLITE COMMUNICATIONS

M. IBNKAHLA[1], *J. SOMBRIN*[2], *and F. CASTANIE*[1]

[1]National Polytechnics Institute of Toulouse

ENSEEIHT, 2, Rue Camichel, 31071 Toulouse, France

[2]French Space Agency (CNES) 18, Av. E. Belin, 31055 Toulouse Cedex, France

E-mail: ibnkahla@len7.enseeiht.fr

ABSTRACT

The paper proposes a neural network technique to adaptively model and characterize digital satellite channels. The neural network model allows to identify each component of the channel by the use of the channel input-output signals as learning data. This technique can be applied to failure detection in digital satellite links, especially those arising in on-board devices. The paper gives some simulation examples of changes which occured in the on-board filters. Our adaptive method allows to determine the origin of the changes and gives the new channel characteristics.

1. INTRODUCTION

Digital satellite channels are composed of linear (e.g. linear filters) and non linear (e.g. travelling wave tubes (TWT)) devices. Classical adaptive techniques used to identify these channels (such as Volterra series approaches [1]) can give only a model for the channel input-output relationship, and are not able to characterize each component of the channel. When a failure happens in the satellite link (which may concern one or more on-board devices), it is impossible to determine its origin if we can not identify each component of the channel.

In a recent paper [5], we have proposed a general structure, the adaptive non linear enhancer (ANLE), for non linear channel identification. The ANLE is an adaptive neural network structure which allows not only to model the global non linear channel input-output relationship, but also to characterize each component of the channel (the learning process is performed by using the channel input-output signals). [3] and [4] present other applications of neural networks to satellite communications.

THIS WORK HAS BEEN SUPPORTED IN PART BY THE FRENCH SPACE AGENCY (CNES) UNDER CONTRACT 962/94/CNES/1232/00.

In this paper, we use an ANLE structure to model satellite channels equipped with TWT amplifiers. We analyze the capability of the neural network to model the channel components. This technique is then applied to failure detection. We give some simulation examples of changes in the on-board filters characteristics. Our adaptive method allows to locate the origins of the changes and gives the new characteristics of the channel.

The paper is organized as follows. Part 2 describes digital satellite channels. In part 3 we present the ANLE structure and its application to the identification problem. The application to failure detection is given in section 4.

2. DIGITAL SATELLITE CHANNELS

A satellite channel consists of two earth stations connected by a repeater (satellite) through two radio links (uplink and downlink). As an example, consider the simplified scheme of figure 1 modelled in the complex base band. The transmission filter F0, the IMUX (input multiplexing) filter F1, and the OMUX (output multi-plexing) filter F2 are linear. The TWT acts as a memoryless nonlinearity with a complex transfer function which depends only on the input complex envelope. It exhibits two kinds of non linearities, amplitude distortion (AM/AM conversion) and phase distortion (AM/PM conversion) [1, 7].

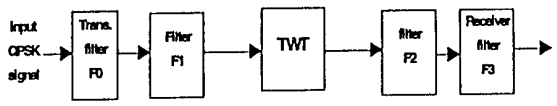


Figure 1: A simplified satellite channel.

The satellite channel used in this paper (figure 1) has the following characteristics. The signals are QPSK modulated, the transmission filter F0 is a four-pole Chebychev with $3dB$ bandwidth $\frac{1.66}{T}$, F1 is a four-pole Chebychev with $3dB$ bandwidth $\frac{2}{T}$, and F2 is a four-pole Chebychev with $3dB$ bandwidth $\frac{3.3}{T}$. The TWT AM/AM and AM/PM conversions are represented by Saleh model [7],

$$A(r) = \frac{\alpha_a r}{1+\beta_a r^2} \ , \ \phi(r) = \frac{\alpha_p r^2}{1+\beta_p r^2}$$

where r is the TWT input amplitude, $\alpha_a = 2, \beta_a = 1, \alpha_p = 4.0033$, and $\beta_p = 9.104$.

3. CHANNEL IDENTIFICATION

3.1. IDENTIFICATION STRUCTURE

In a recent paper [5], we have proposed a general structure, the adaptive non linear enhancer (ANLE), for non linear system identification. The ANLE is a neural network structure which allows not only to model the global non linear system input-output relationship, but also to identify each component of the system (the learning process is performed using the system input-output signals).

In this paper, we use an ANLE structure to model satellite channels equipped with TWT amplifiers. We analyze the capability of the neural network to model the channel components.

The ANLE structure of figure 2 is used to model the block F1-TWT-F2 (the satellite itself). The ANLE copies the structure to be identified (a memoryless non linear system between two linear systems): It is composed of a linear subnetwork (PL1, with 60 weights), a non linear subnetwork (PNL, with 18 scalar neurons), and a second linear subnetwork (PL2, with 60 weights). Note that the amplitude and phase conversions of the non linear subnetwork depend only on the input signal amplitude (as in the TWT). The learning procedure is performed by presenting to the neural net at each iteration a pair of the channel input-output complex signals.

The ANLE works as follows. The first linear part (PL1) filters the complex-valued input $x(n)$ (real FIR filtering), its output is then written as:

$$y(n) = \sum_{k=1}^{N_1} w_{1k} x(n-k+1)$$

The two non linear subnetworks correspond to the gain (G) and phase (P) conversions, respectively. The squared amplitude ρ of the output y of PL1 is presented to both non linear subnets. Their outputs G and ϕ can be written as:

$$\begin{aligned} G(\rho(n)) &= \sum_{k=1}^{N_G} w_{G2k} f(w_{G1k}\rho(n) + b_{G1k}) + b_{G2} \\ \phi(\rho(n)) &= \sum_{k=1}^{N_P} w_{P2k}(f(w_{P1k}\rho(n) + b_{P1k}) \\ &\quad - f(b_{P1k})) \end{aligned}$$

where $\rho(n) = r^2(n) = \|y(n)\|^2$. Note that the origin of the phase is 0 by construction ($\phi(0) = 0$).
The PNL output can be written as:

$$z(n) = G(\rho(n))\, e^{j\phi(\rho(n))} y(n).$$

We present to PL2 the vector $\mathbf{z}(n) = [z(n), z(n-1) \ldots z(n-N_2+1)]^t$,where N_2 is the memory of PL2. The output $s(n)$ of the ANLE is then:

$$s(n) = \sum_{i=1}^{N_2} w_{2i}(n)\, z(n-i+1)$$

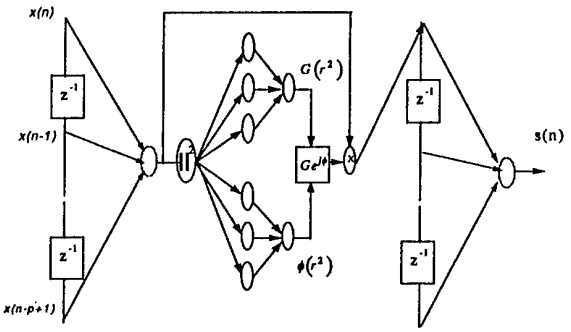


Figure 2: Identification structure

3.2. ALGORITHM

We adjust the network weights by using a gradient descent algorithm which minimizes the squared error between the channel output $d(n)$ and the ANLE output. The squared error is written as:

$$\|e(n)\|^2 = \|d(n) - s(n)\|^2 = e_R^2(n) + e_I^2(n),$$

where R and I denote the real and imaginary parts, respectively.
The PL2 weights are updated as:

$$\begin{aligned} w_{2i}(n+1) &= w_{2i}(n) + \mu(z_R(n-i+1)\,\delta_{SR}(n) \\ &\quad + z_I(n-i+1)\,\delta_{SI}(n)), \end{aligned}$$

where $\delta_{SR}(n) = e_R(n)$ and $\delta_{SI}(n) = e_I(n)$.
We present now the updating rule of subnetwork G.

The second layer weights of subnetwork G are updated as:

$$w_{G2k}(n+1) = w_{G2k}(n) + \mu x_{G1k}\delta_{G2},$$

where x_{G1k} is the output of neuron k of the first layer:

$$x_{G1k} = f(net_{G1k}), net_{G1k} = w_{G1k}\rho + b_{G1k},$$

and δ_{G2} is an error term:

$$\delta_{G2} = 2w_{21}(e_R z_R + e_I z_I).$$

The second layer bias term is updated as:

$$b_{G2}(n+1) = b_{G2}(n) + \mu\delta_{G2}.$$

The first layer weights are updated as:

$$w_{G1k}(n+1) = w_{G1k}(n) + \mu w_{G2k}\delta_{G2}\rho f'(net_{G1k}).$$

The first layer bias vector is updated as:

$$b_{G1k}(n+1) = b_{G1k}(n) + \mu w_{G2k}\delta_{G2} f'(net_{G1k}).$$

We now present the updating rule of subnetwork P. The second layer weights of subnetwork P are updated as:

$$w_{P2k}(n+1) = w_{P2k}(n) + \mu x_{P1k}\delta_{P2},$$

where x_{P1k} is the output of neuron k of the first layer:

$$x_{P1k} = f(w_{P1k}\rho + b_{P1k}) - f(b_{P1k}),$$

and δ_{P2} is an error term:

$$\begin{aligned}\delta_{P2} = {} & 2Gw_{21}(e_I(-\sin(\phi)y_R - \cos(\phi)y_I) \\ & + e_R(\cos(\phi)y_R - \sin(\phi)y_I)).\end{aligned}$$

The first layer weights are updated as:

$$w_{P1k}(n+1) = w_{P1k}(n) + \mu w_{P2k}\delta_{P2}\rho f'(w_{P1k}\rho + b_{P1k}).$$

The first layer bias vector is updated as:

$$\begin{aligned}b_{P1k}(n+1) = {} & b_{P1k}(n) + \mu w_{P2k}\delta_{P2} \\ & (f'(w_{P1k}\rho + b_{P1k}) - f'(b_{P1k})).\end{aligned}$$

Finally, the first linear part (PL1) weight vector is updated as:

$$\begin{aligned}w_{1i}(n+1) = {} & w_{1i}(n) + 4\mu(x_R(n-i+1)(2y_R\delta_y \\ & + Gw_{21}(\cos(\phi)e_R + \sin(\phi)e_I)) \\ & + x_I(n-i+1)(2y_I\delta_y \\ & + Gw_{21}(-\sin(\phi)e_R + \cos(\phi)e_I))),\end{aligned}$$

where

$$\begin{aligned}\delta_y = {} & \delta_{G2}\sum_{k=1}^{N_G} w_{G2k}w_{G1k}f'(net_{G1k}) \\ & + \delta_{P2}\sum_{k=1}^{N_P} w_{P2k}w_{P1k}f'(w_{P1k}\rho + b_{P1k}).\end{aligned}$$

It is worth noting that the above algorithm has the same properties as the classical backpropagation algorithm [2, 6] (error backpropagation, parallelism, etc.). Note that the two non linear subnetworks can be adjusted in parallel and independently.

3.3. SIMULATION RESULTS

In figure 3 we compare the neural network output (generalization) to that of the channel. The generalization MSE was 3.310^{-4} ($RMS = 0.018$, $SNR = 33.3\,dB$, the TWT working at saturation). Figures 4-7 show that each element of the channel has been correctly characterized by the corresponding part of the ANLE.

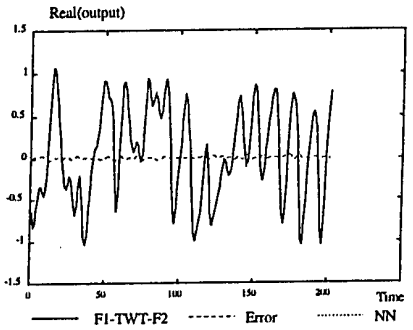


Figure 3: Generalization performance.

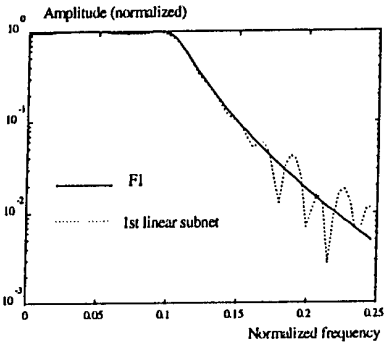


Figure 4: Frequency response of F1 and the NN model.

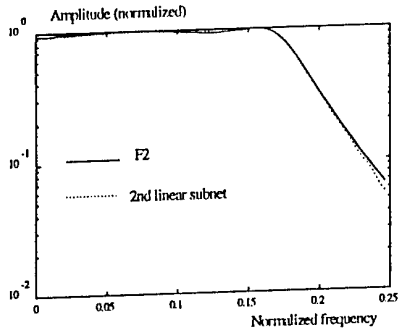


Figure 5: Frequency response of F2 and the NN model.

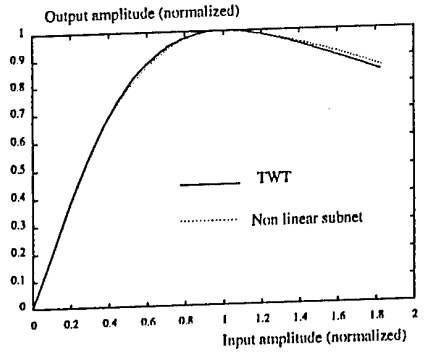


Figure 6: AM/AM conversion (TWT and NN).

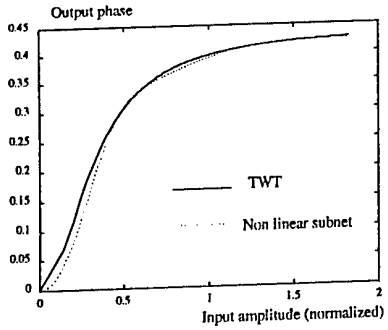


Figure 7: AM/PM conversion (TWT and NN).

4. DETECTION OF CHANGES IN SATELLITE CHANNEL CHARACTERISTICS

The identification technique is now applied to the detection of changes in satellite channels characteristics which may occur because of a failure.

In the simulation below, a 'big' change occured in filter F1 (the other channel components were taken unchanged). Figure 8 shows the frequency response of F1 before and after the change. By using the same ANLE structure as the above section, the adaptive system determined the origin of the change (filter F1) and gave the new filter frequency response (figure 9).

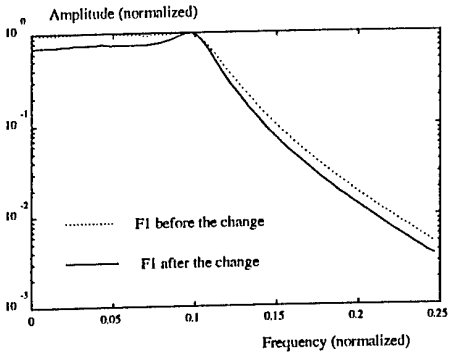


Figure 8: F1 before and after the change.

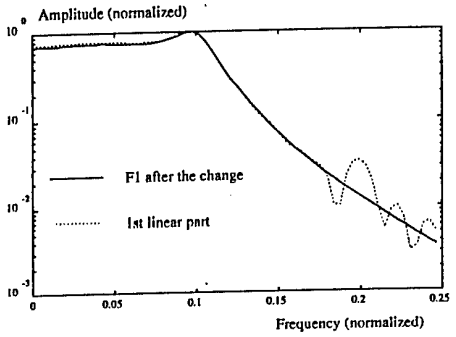


Figure 9: New filter characterized by PL1.

In the simulation below, a 'small' change occured in filter F2 (the other channel components were taken unchanged). Figure 10 shows the frequency response of F2 before and after the change. By using the same ANLE structure as the above section, the adaptive system determined the origin of the change (filter F2) and gave the new filter frequency response (figure 11).

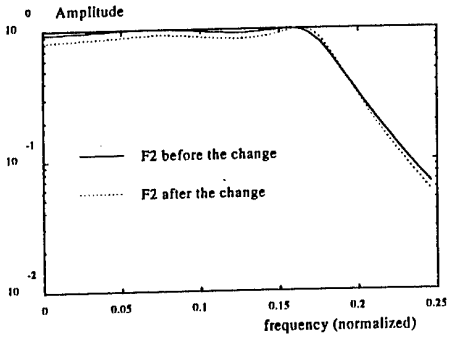


Figure 10: F2 before and after the change.

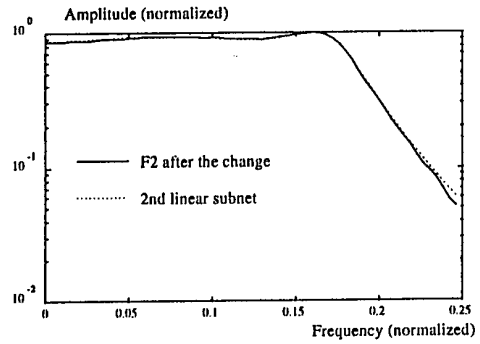


Figure 11: New filter characterized by PL2.

5. CONCLUSION

The paper proposed a new approach for channel identification and failure detection in digital satellite communications. For the identification process, we used the adaptive non linear enhancer (ANLE) structure [5] which allows to characterize the channel linear and non linear devices. The learning process is performed by using the channel input and output signals (i.e. we do not have access to the different input-output signals of the channel components). The ANLE has few parameters (e.g. 18 scalar neurons were sufficient to model the non linear part). This technique was applied to failure detection in digital satellite links. The paper gave some simulation examples of changes which occurred in the on-board filters. Our adaptive method was able to determine the origins of the changes and gave the new characteristics of the satellite channel.

6. REFERENCES

[1] S. Benedetto, E. Biglieri, and V. Castellani, *Digital Transmission Theory*, Printice Hall International, Englewood Cliffs, New Jersey, 1987.

[2] S. Haykin, *Neural Networks: a Comprehensive Foundation*, IEEE Press, 1994.

[3] M. Ibnkahla, J. Sombrin, F. Castanié, and N. J. Bershad, "Neural networks for modelling nonlinear memoryless communication channels", submitted to IEEE Trans. Communications, 1995.

[4] M. Ibnkahla and F. Castanié, "Neural networks for digital communications and signal processing: overview and new results", In E. Biglieri and M. Luise Eds., *Signal Processing for Digital Communications*, Springer Verlag, London, 1996.

[5] M. Ibnkahla and F. Castanié, "Neural network identification of non linear channels: The adaptive non linear enhancer", In proceedings of IEEE International Conference on Neural Networks (ICNN'96), Washington, D.C., June 1996.

[6] R. P. Lippmann, "An introduction to computing with neural nets", IEEE ASSP Magazine, pp. 4-22, April 1987.

[7] A. Saleh, "Frequency-independent and frequency-dependent nonlinear models of TWT amplifiers", IEEE Trans. Communications, Vol. Com-29, No. 11, November 1981.

ON THE USE OF DERIVATIVE CONSTRAINTS TO CONTROL BEAMFORMING RESPONSE SHAPES AGAINST INTERFERING DIRECTIONS

Jacques Fois Pelayo, José M. Páez Borrallo*

E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid
Ciudad Universitaria s/n, 28040 Madrid
e-mail : "jfp@serv01.rpi.ses.alcatel.es"
*e-mail : "paez@gaps.ssr.upm.es"

I. INTRODUCTION

The use of beamformers with derivative constraints has been studied by several authors over the last two decades. These derivative constraints have been aimed at obtaining a flat magnitude response near the direction of interest, and the authors have given either sufficient conditions or, as for instance in [1], sufficient as well as necessary conditions in order to achieve the flat main beam response. The purpose of this work is to show that derivative constraints can be also applied to interfering signal locations as well as to the main beam in order to avoid that possible fluctuations of the interferers may significantly affect their locations. As a matter of fact, in some applications the desired direction is normally nearly fixed while the interfering locations might suffer variations. It is needless to say that the problem can be solved from another viewpoint by increasing the number of spatial null constraints over a region where it is supposed the interferers may come from. Although the results can be extended to the broadband case [1], the paper will consider only the narrowband beamformer structure.

The structure of this paper is described as follows. After the introduction, section II contains a brief review of the beamformer problem basic concepts introducing the Generalized Sidelobe Canceller which is going to be used in the remaining sections. Next, in section III, the equations involved in obtaining the derivative constraints are mentioned and finally section IV describes the derivative constraints performance by the aid of computer simulations.

II. BACKGROUND

Let us consider an array of L isotropic elements distributed at known locations over the xyz space. The beamformer structure is such that the output at any time n is given by

$$y(n) = \mathbf{w}^H \mathbf{x}(n) \tag{1}$$

where $\mathbf{x}(n)$ is the data vector at time n and $\mathbf{w}$ the weights vector, while H denotes the matrix is complex conjugate transposed. The derivative constraints will be given with respect to the magnitude response $F(\theta,\phi)$, with θ and ϕ being the elevation and azimuth angles respectively in the xyz space. The magnitude response is defined as

$$F(\theta,\phi) = \mathbf{w}^H E\left\{\mathbf{x}\mathbf{x}^H\right\}\mathbf{w} = \mathbf{w}^H E\left\{\begin{bmatrix} 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 1 & & & \cdots & & \cdots \\ \cdots & & \cdots & & r_{ij} & & \cdots \\ \cdots & & & \cdots & \cdots & & \cdots \\ \cdots & \cdots & r_{ji} & \cdots & \cdots & & \cdots \\ \cdots & & & & & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1 \end{bmatrix}\right\}\mathbf{w} \tag{2}$$

L is the number of array sensors and τ_i the delay in sensor i (i=1,2,...,L) due to the signal propagation with respect to the axis reference. The elements of the matrix are $r_{ij} = r_{ji}^* = \exp(j\omega(\tau_i-\tau_j))$ and $\tau_i = (x_i\sin\theta\cos\phi+y_i\sin\theta\sin\phi+z_i\cos\theta)/c$ for any i, where (x_i,y_i,z_i) defines the position of the sensor i and c is the speed of propagation of the wavefront detected on the array.

The minimum variance beamforming problem with constraints is formulated minimizing the variance of the output $y(n)$ in (1)

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{\mathbf{XX}} \mathbf{w} \tag{3}$$

subject to the following set of linear constraints

$$\mathbf{C}^H \mathbf{w} = \mathbf{f} \tag{4}$$

where $\mathbf{R}_{\mathbf{XX}}$ is the covariance matrix of the data vector $\mathbf{x}$, $\mathbf{C}$ the constraint matrix and $\mathbf{f}$ the vector containing the magnitude and phase wanted at the output for every constraint.

As it is well known, the solution to the problem is given by

$$\mathbf{w}_{\mathbf{opt}} = \mathbf{R}_{\mathbf{XX}}^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_{\mathbf{XX}}^{-1}\mathbf{C}\right)^{-1}\mathbf{f} \tag{5}$$

The idea of controlling the magnitude response to achieve a flat main beam response is not new. Sufficient derivative constraints have been derived by Er & Cantoni [2] in obtaining this flat main beam response, and later, further

studies by Er [3] and Tseng [1] have been appearing to get the conditions be also necessary.

There is another way in depicting a beamformer structure different from the conventional shape. It can be easily found realizing that a particular solution to the minimum variance beamforming is the so-called *quiescent* solution extracted from (5) for the presence of only uncorrelated noise, and expressed as

$$\mathbf{w}_q = \mathbf{C}\left(\mathbf{C}^H\mathbf{C}\right)^{-1}\mathbf{f} \qquad (6)$$

Figure 1 shows the beamformer structure known as *Generalized Sidelobe Canceller (GSC)*, which was introduced by Griffiths & Jim in [4]. The optimum

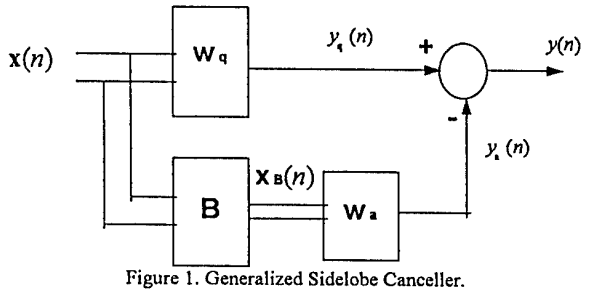


Figure 1. Generalized Sidelobe Canceller.

solution $\mathbf{w}_{opt}$ is then the sum of two vectors, one is the quiescent (only depending upon the constraints) from the upper branch and the other one is that resulting vector (depending on the data) from the lower branch, always referring to figure 1. The GSC structure with derivative constraints has been also studied in the past by Buckley & Griffiths [5].

The intention of this work is to make use of the derivative constraints specified not in the look or main directions as usually but in the interfering steering locations too.

III. OBTAINING THE SUFFICIENT LINEAR DERIVATIVE CONSTRAINTS

Sufficient null derivative constraints of the magnitude response are derived for both linear array and circular array cases, and we will deal each case with two subcases, namely zero-plus-first (ZF) derivative constraints and zero-plus-first-plus-second (ZFS) derivative constraints.

III.1 Linear array case.

Let us suppose the array is located along the z axis and the sensors are equally spaced by a half wavelength $\lambda/2$. In this case, the data in each sensor are independent upon the azimuth ϕ and so the magnitude response.

III.1.a ZF derivative constraints.

Consider the signal scenario is such that there is an interferer coming from θ_I. Taking the first derivative function with respect to θ of the magnitude response given by (2) yields the expression

$$\frac{\partial F}{\partial \theta} = -j\omega\left(\mathbf{w}^H\Lambda_\theta\mathbf{R}_{xx}\mathbf{w} - \mathbf{w}^H\mathbf{R}_{xx}\Lambda_\theta\mathbf{w}\right) \qquad (7)$$

which we make equal to zero to get the sufficient ZF linear derivative constraint, resulting

$$\mathbf{C}^H(\theta)\Lambda_\theta\mathbf{w}\Big|_{\theta_I} = 0 \qquad (8)$$

where

$$\Lambda_\theta = \mathrm{diag}\left[\frac{\partial\tau_1}{\partial\theta}\ \ \frac{\partial\tau_2}{\partial\theta}\ \ \ldots\ \ \frac{\partial\tau_L}{\partial\theta}\right] \qquad (9)$$

We would like to comment a point before going on with the analysis. Equation (8) has been considered as a sufficient condition to make the magnitude response first derivative function equal to zero but the fact is that it is not really sufficient. In effect, we know that $\mathbf{C}^H(\theta)\mathbf{w}\Big|_{\theta_I} = 0$ holds already because it corresponds to one of the spatial constraints, and this makes, specifying to θ_I, equation (7) be zero without any other additional constraint (unless we use spatial constraints holding values different from zero). Nevertheless, using the equation (8) as a constraint leads us to a flat response over the interfering region as desired.

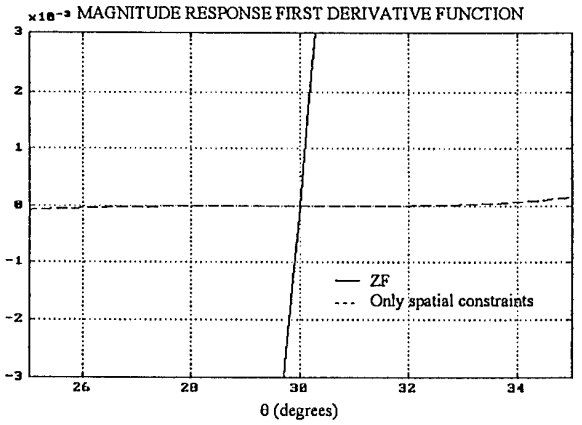


Figure 2. Zoom of the power response first derivative function for a linear array with an interferer in 30°.

The reason is rather simple if we look at the figure 2. The figure represents two first derivative functions of the magnitude response for a linear array, one corresponds to the use of ZF derivative constraints and the other to the use of only spatial constraints. It is an easy task to realize

that the derivative constraints behave as a *'second zero'* in the first derivative function, and this is why the magnitude response gets flat.

III.1.b ZFS derivative constraints.

Deriving now once again with respect to θ we obtain the equation

$$\frac{\partial^2 F}{\partial \theta^2} = 2\omega^2\left(\mathbf{w}^H \Lambda_\theta \mathbf{R}_{xx} \Lambda_\theta \mathbf{w}\right) - \\ -\omega^2\left(\mathbf{w}^H \Lambda_\theta^2 \mathbf{R}_{xx}\mathbf{w} + \mathbf{w}^H \mathbf{R}_{xx} \Lambda_\theta^2 \mathbf{w}\right) - \\ -j\omega\left(\mathbf{w}^H \frac{\partial \Lambda_\theta}{\partial \theta}\mathbf{R}_{xx}\mathbf{w} - \mathbf{w}^H \mathbf{R}_{xx}\frac{\partial \Lambda_\theta}{\partial \theta}\mathbf{w}\right) \quad (10)$$

from which, making equal to zero, it is possible to get the following equations

$$\mathbf{C}^H(\theta)\Lambda_\theta \mathbf{w}\Big|_{\theta_1} = 0 \quad (11a)$$

$$\mathbf{C}^H(\theta)\Lambda_\theta^2 \mathbf{w}\Big|_{\theta_1} = 0 \quad (11b)$$

$$\mathbf{C}^H(\theta)\frac{\partial \Lambda_\theta}{\partial \theta}\mathbf{w}\Big|_{\theta_1} = 0 \quad (11c)$$

Notice that equation (8) derived for the ZF case shows up again in (11a), and also that equation (11c) is linearly dependent on (8) due to the linear array shape. Consequently we consider equation (11b) as the linear derivative constraint for the ZFS case.

III.2 Circular array case.

Now we assume the sensors are equally spaced along a circular ring of radius d on the xy plane. This time the data is depending on both the spatial angles θ and ϕ, hence the magnitude response must be derived with respect to θ as well as with respect to ϕ.

III.2.a ZF derivative constraints.

Proceeding in the same way as for the linear case we obtain

$$\frac{\partial F}{\partial \theta} = -j\omega\left(\mathbf{w}^H \Lambda_\theta \mathbf{R}_{xx}\mathbf{w} - \mathbf{w}^H \mathbf{R}_{xx}\Lambda_\theta \mathbf{w}\right) \quad (12)$$

$$\frac{\partial F}{\partial \phi} = -j\omega\left(\mathbf{w}^H \Lambda_\phi \mathbf{R}_{xx}\mathbf{w} - \mathbf{w}^H \mathbf{R}_{xx}\Lambda_\phi \mathbf{w}\right) \quad (13)$$

so that the sufficient constraints are

$$\mathbf{C}^H(\theta,\phi)\Lambda_\theta \mathbf{w}\Big|_{(\theta_1,\phi_1)} = 0 \quad (14a)$$

$$\mathbf{C}^H(\theta,\phi)\Lambda_\phi \mathbf{w}\Big|_{(\theta_1,\phi_1)} = 0 \quad (14b)$$

where the couple (θ_1,ϕ_1) defines the interfering direction.

III.2.b ZFS derivative constraints.

In this case the development leads to the following set of equations

$$\frac{\partial^2 F}{\partial \theta^2} = 2\omega^2\left(\mathbf{w}^H \Lambda_\theta \mathbf{R}_{xx} \Lambda_\theta \mathbf{w}\right) - \\ -\omega^2\left(\mathbf{w}^H \Lambda_\theta^2 \mathbf{R}_{xx}\mathbf{w} + \mathbf{w}^H \mathbf{R}_{xx} \Lambda_\theta^2 \mathbf{w}\right) - \\ -j\omega\left(\mathbf{w}^H \frac{\partial \Lambda_\theta}{\partial \theta}\mathbf{R}_{xx}\mathbf{w} - \mathbf{w}^H \mathbf{R}_{xx}\frac{\partial \Lambda_\theta}{\partial \theta}\mathbf{w}\right) \quad (15)$$

$$\frac{\partial^2 F}{\partial \phi^2} = 2\omega^2\left(\mathbf{w}^H \Lambda_\phi \mathbf{R}_{xx} \Lambda_\phi \mathbf{w}\right) - \\ -\omega^2\left(\mathbf{w}^H \Lambda_\phi^2 \mathbf{R}_{xx}\mathbf{w} + \mathbf{w}^H \mathbf{R}_{xx} \Lambda_\phi^2 \mathbf{w}\right) - \\ -j\omega\left(\mathbf{w}^H \frac{\partial \Lambda_\theta}{\partial \phi}\mathbf{R}_{xx}\mathbf{w} - \mathbf{w}^H \mathbf{R}_{xx}\frac{\partial \Lambda_\phi}{\partial \phi}\mathbf{w}\right) \quad (16)$$

$$\frac{\partial^2 F}{\partial \theta \partial \phi} = \omega^2\left(\mathbf{w}^H \Lambda_\theta \mathbf{R}_{xx} \Lambda_\phi \mathbf{w} + \mathbf{w}^H \Lambda_\phi \mathbf{R}_{xx} \Lambda_\theta \mathbf{w}\right) - \\ -\omega^2\left(\mathbf{w}^H \Lambda_\theta \Lambda_\phi \mathbf{R}_{xx}\mathbf{w} + \mathbf{w}^H \mathbf{R}_{xx} \Lambda_\phi \Lambda_\theta \mathbf{w}\right) - \\ -j\omega\left(\mathbf{w}^H \frac{\partial \Lambda_\theta}{\partial \phi}\mathbf{R}_{xx}\mathbf{w} - \mathbf{w}^H \mathbf{R}_{xx}\frac{\partial \Lambda_\theta}{\partial \phi}\mathbf{w}\right) \quad (17)$$

$$\frac{\partial^2 F}{\partial \phi \partial \theta} = \omega^2\left(\mathbf{w}^H \Lambda_\theta \mathbf{R}_{xx} \Lambda_\phi \mathbf{w} + \mathbf{w}^H \Lambda_\phi \mathbf{R}_{xx} \Lambda_\theta \mathbf{w}\right) - \\ -\omega^2\left(\mathbf{w}^H \Lambda_\phi \Lambda_\theta \mathbf{R}_{xx}\mathbf{w} + \mathbf{w}^H \mathbf{R}_{xx} \Lambda_\theta \Lambda_\phi \mathbf{w}\right) - \\ -j\omega\left(\mathbf{w}^H \frac{\partial \Lambda_\phi}{\partial \theta}\mathbf{R}_{xx}\mathbf{w} - \mathbf{w}^H \mathbf{R}_{xx}\frac{\partial \Lambda_\phi}{\partial \theta}\mathbf{w}\right) \quad (18)$$

From the equations (15)-(18) we can derive the following linearly independent sufficient derivative constraints

$$\mathbf{C}^H(\theta,\phi)\Lambda_\theta^2 \mathbf{w}\Big|_{(\theta_1,\phi_1)} = 0 \quad (19a)$$

$$\mathbf{C}^H(\theta,\phi)\Lambda_\theta \Lambda_\phi \mathbf{w}\Big|_{(\theta_1,\phi_1)} = 0 \quad (19b)$$

IV. COMPUTER STUDIES

Software simulations have been carried out in order to show the impacts the derivative constraints yield on the beamformer output pattern. The examples illustrated in this section shall consider a very simple scenario consisting of one desired look direction and one interferer, assuming spatial null constraints applied in the interferer as well as in the look direction and the derivative null constraints applied only in the interferer.

We begin considering a linear array of 8 antenna elements equally spaced by a half wavelength and a scenario consisting of a look direction at broadside and an interferer arriving at θ_I=30°. Figure 3 shows the beamformer response where it is possible to notice how the response broadens in the interfering location as we insert derivative constraints. A zoom of the region close to the interference is depicted in figure 4, where the flat behaviour can be appreciated. Notice that any mistuning effect in the interferer is very harmful indeed to the magnitude response if we do not use derivative constraints, leading clearly to an important attenuation loss. This matter is better illustrated in figure 5, which shows the effect suffered by the attenuation performance for the three different cases treated here when the interfering frequency has changed.

Next we consider the case of the circular array. We assume the array sensors are located along a circular ring of radius d such that $kd=L$, with $k=2\pi/\lambda$.

Figure 6 depicts the beamformer response for the circular array case, which is symmetric due to the circular nature. Similar deductions derived for the linear case can be also applied for the circular case. The flat behaviour is also shown in figure 7 which is extracted from figure 6 by zooming in. A figure like figure 5 can be also given with a very similar shape but it is not provided in this paper.

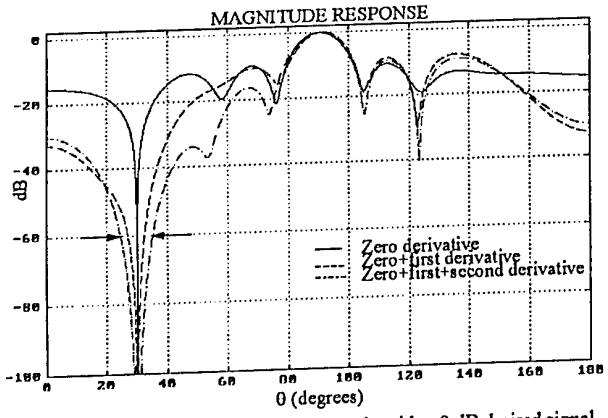


Figure 3. Beamformer response for a scenario with a 0 dB desired signal coming from the broadside direction and a 7 dB interference at θ=30° together with a -20 dB uncorrelated white noise.

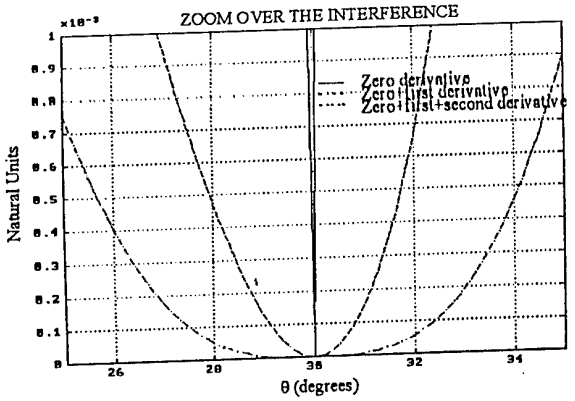


Figure 4. Zoom of figure 2.

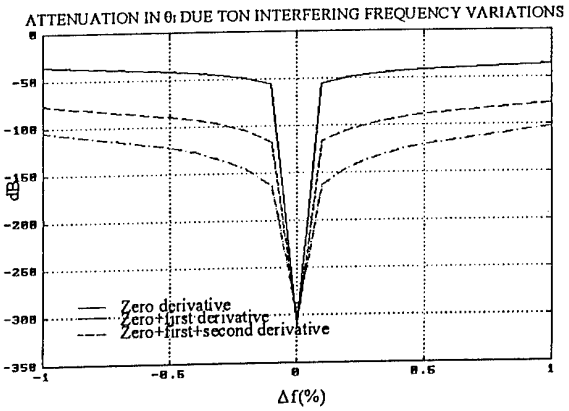


Figure 5. Attenuation in θ_I when changes affect the interfering frequency. 0% means the interfering frequency has not been affected.

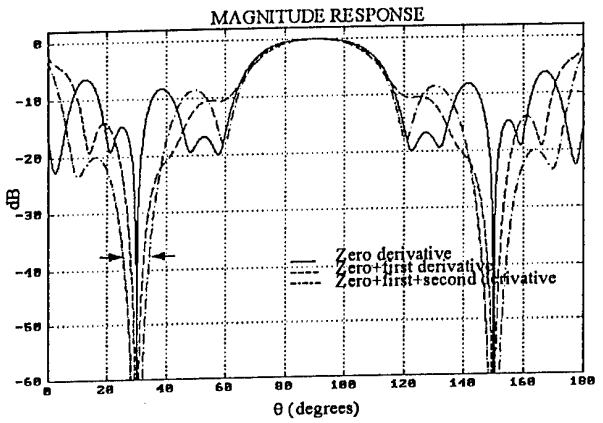


Figure 6. Beamformer pattern, fixing ϕ=90°, for a circular array of 8 isotropic elements with a look direction steered at (θ,ϕ)=(90°,90°), an interference at (θ,ϕ)=(30°,90°) and a -20 dB uncorrelated white noise.

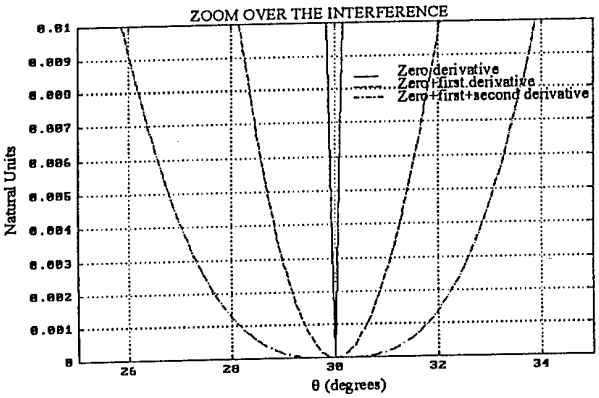


Figure 7. Zoom of figure 4. ϕ is also fixed to 90°.

We have seen thus far how the use of derivative constraints affect the shape of the beamformer output response. We might suggest now to look for an *equivalent* beamformer response without using any derivative constraint but only spatial constraints, where the term "*equivalent*" stands here for "*using the same number of constraints in matrix C*". It is clearly an alternative solution to the use of derivative constraints that spreads out the magnitude response where the interferer is located,

as depicted in figure 8 which represents the same example set forth for the linear array in figure 3 but considering the ZF case (in solid line) compared to a response (dashed line) with two spatial constraints, and consequently obtaining two equivalent responses. In the first instance, note that it is possible to distinguish two different behaviours, i. e., the derivative constraints yield a maximally flat behaviour whereas only spatial constraints produce a ripple behaviour.

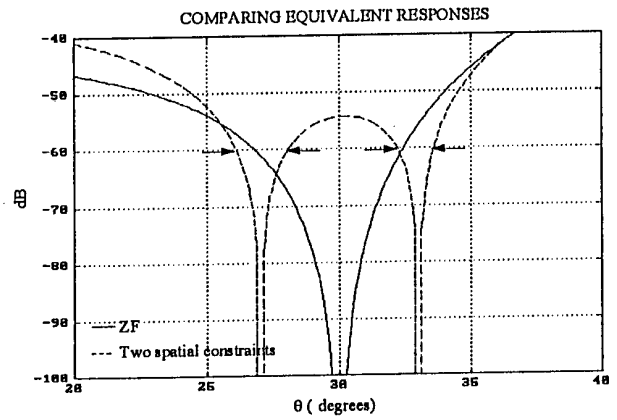


Figure 8. Linear array with an interferer in θ_I=30°. ZF case vs. response with spatial constraints in θ_1=27° and θ_2=33°.

If we look at the figure 8 we see that the ripple behaviour can make the response unsatisfactory if, for instance, we want to have a response below -60 dB in the interferer although in the edges we get a better performance. This matter can be overcome changing the spatial constraints so that the ripple gets lower as shown in figure 9 but, nevertheless, the derivative constraints give now a better response.

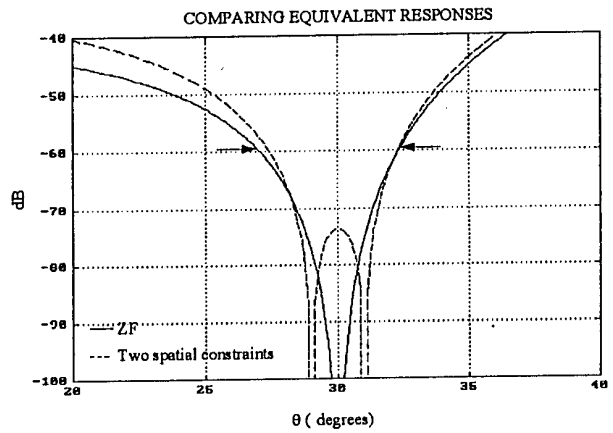


Figure 9. Linear array with an interferer in θ_I=30°. ZF case vs. response with spatial constraints in θ_1=29° and θ_2=31°.

V. CONCLUSIONS

Although derivative constraints have been widely studied in the literature, the fact of applying them to the interferers has not appeared yet so far. The reason is that equation (7) is zero in the interferers without any condition but, nevertheless, the constraints derived show the important role they play against mistuning conditions due, for example, to band extreme interferers, Doppler effects or estimated interfering directions.

Specifying derivative constraints to the interfering directions improves the magnitude performance of a beamformer and it can be seen as an alternative of adding null spatial constraints in the nearest region of the interference.

Besides, it provides a way of controlling the response by mixing the constraints specified in the main beam as well as in the interferers.

The examples shown in section IV are particular because the beamformer response changes if we have a different signal scenario or even designing the array with another number of antenna elements but in general they are good enough to give an overall view of what can be expected if we use derivative constraints.

V. REFERENCES

[1] C.-Y. Tseng, "Minimum variance beamforming with phase-independent derivative constraints", IEEE Trans. Antennas & Propagation, vol. 40, no. 3, pp. 285-294, Mar. 1992.

[2] M. H. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processor" IEEE Trans. ASSP, vol. 31, no. 6, pp. 1378-1393, Dec. 1983.

[3] M. H. Er, "Adaptive antenna array under directional and spatial derivative constraints", Proc. IEE, vol. 135, pt. H, no. 6, pp. 414-419, Dec. 1988.

[4] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming", IEEE Trans. Antennas & Propagation, vol. 30, pp. 27-34, Jan. 1982.

[5] K. M. Buckley and L. J. Griffiths, "An adaptive generalized sidelobe canceller with derivative constraints" IEEE Trans. Antennas & Propagation, vol. 34, no. 3, pp. 311-319, Mar. 1986.

Adaptive Systems in Communications

BLIND ADAPTIVE MUD WITH SILENCE LISTENING

Enrico Del Re and Luca Simone Ronga

Dipartimento di Ingegneria Elettronica
Universitá degli Studi di Firenze
Via di Santa Marta 3
50133 Florence - ITALY

ABSTRACT

In communication systems adopting CDMA as multiacces protocol, the *multiuser detection* (MUD) is the optimal choice for both a better utilization of the communication medium and a reduction of the near-far effects. Several papers [4] [5] [3] in the literature suggest some adaptive approaches to multiuser detection with various degree of blindness with respect to the knowledge of the characteristics of the channel and of the signals adopted.

The weakest aspect found in those techniques is the need for the knowledge of a close approximation of the signal adopted by the user whose information is seeked. In this paper it is presented an adaptive multiuser detector which tries to estimate the user's signal by listening to the channel when the user is not transmitting.

The resulting detector is expected to behave well even when the time varying channel changes continuosly the shape of the user of interest.

Part I
Introduction

Recent works [4] [5] [6] have shown that Multi-User Detection exerts a radical improvement on the performance of a CDMA receiver. An attractive model of a blind adaptive MUD receiver has recently been presented by Madhow, Honig and Verdú [3], a model for which the knowledge of the signature waveform of the desired user, along with the timing of all the users, are the only requirements.

Even if the signal waveform of the desired user is not exactly known, the detector performs fairly well if a limitation to the cancellation of interference is imposed.

In this paper it is presented a blind MUD with an additional adaptive branch which is designed to correct the wrong estimate of the wanted user's signature caused by a multipath channel. To reach this goal without training sequences from the trasmitter, some additional information are supposed to be available to the receiver: the knowledge of the time periods when the desired user is silent, i.e. it is not trasmitting any signal in the channel.

This work was supported under the fynancial support of ASI and MURST

Part II
"Listen to the silence!"

The inspection of the common characteristics of the information flowing in wireless systems reveals the substantially dicontinuous nature of the information flow. In voice channels over 30% of time is not used to transmit any information. In data channels discontinuances are present as well, depending on the nature of the connected system.

We shall now prove that in CDMA systems, the knowledge of a user's silence period is useful to the correct estimate of that user's signature waveform. Unlike the training sequences, which may be used to correct that estimate as well, silence periods are always present during the trasmission and so their use may be easily integrated in the communication devices.

The *silence/not silence* information is a very slowly signal compared to the spreading signatures and may be transmitted in a very narrow portion of the medium spectrum without any sensible loss of capacity.

To get into details the following signal is received by a CDMA environment:

$$\mathbf{y} = \sum_{k=1}^{K} A_k b_k s_k + n \qquad (1)$$

where

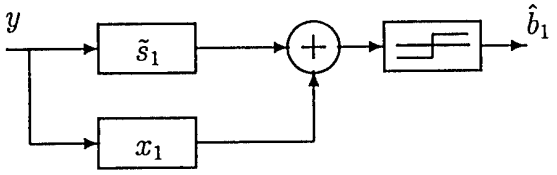


Figure 1: The general structure of the MUD receiver.

K is the number of users,

A_k is the received amplitude of the k-th user's signal,

b_k is the antipodal binary information from the k-th user,

s_k is the received signal from the k-th user, here represented as a member of a Hilbert space $\mathcal{H}$.

n is the projection onto $\mathcal{H}$ of a white gaussian stochastic process with zero mean and variance equal to σ.

We are interested in the first user's information. Due to multipath, at the receiver is available only an estimate of the first user's signal: say $\tilde{s}_1$.

The linear MUD receiver is composed by two components, the estimate of the desired user's signal $\tilde{s}_1$ and an orthogonal component x_1 dedicated to the suppression of the interferers.

For a binary, antipodal equiprobable signaling the estimated information is given by a threshold detector which follows the decorrelator as shown in fig. 1.

If a reliable estimate of the first user's signal is available, i.e. $\tilde{s}_1 = s_1$, then x_1 may be adaptively modified in order to reduce the so called *Mean Output Energy* defined as

$$MOE[x_1] = E[< y, \tilde{s}_1 + x_1 >^2] \qquad (2)$$

where $E[\bullet]$ is the expectation value and the operator $< \alpha, \beta >$ is the scalar product defined over the Hilbert space chosen to represent the signal considered in the trasmission.

When the transmitted waveform from user 1 is not exactly known, the adaptation rule of x_1 needs to be modified in order to prevent the cancellation of the desired signal. The blind adaptation rule as described in [3] is modified with an additional costraint on the so called *surplus energy*, defined as

$$\chi = \|c_1\|^2 - 1 \qquad c_1 = \tilde{s}_1 + x_1 \qquad (3)$$

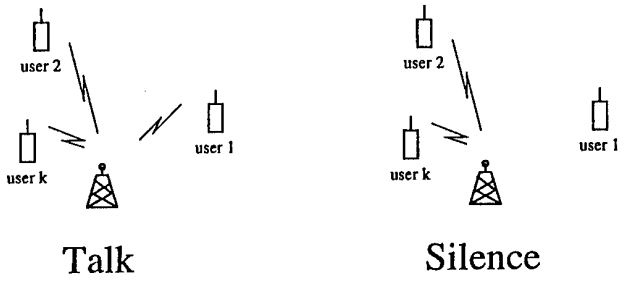


Figure 2: The two possible states of user 1.

By imposing a value of χ in the blind adaptation rule, is achieved a lower degradation of the *signal to interference ratio (SIR)* even in the presence of an estimation error of the desired user's signal.

In order to achieve a good level of SIR in the high SNR region, the silence period listening is introduced to move the desired signal estimate towards a signal less sensible to the interference.

Two possible states are considered with respect to the first user as shown in fig 2.

TALK corresponding to a received signal as

$$\mathbf{y} = \sum_{k=1}^{K} A_k b_k s_k + n$$

SILENCE with a received signal as

$$\mathbf{y}_s = \sum_{k=2}^{K} A_k b'_k s_k + n$$

We look at the quantity called *Mean Silence Output Energy* defined as

$$MSiE[\tilde{s}_1] = E[(< y_s, \tilde{s}_1 > + < y, x_1 >)^2] \qquad (4)$$

As shown in appendix A in the low noise region ($\sigma \to 0$) the mean silence output energy is equal to

$$\underset{\sigma \to 0}{MSiE[\tilde{s}_1]} = \sum_{k=2}^{K} A_k^2 < \tilde{s}_1, s_k >^2 \qquad (5)$$

The local minimization of such a quantity, along with the constraint of a unitary norm of $\tilde{s}_1$ ($\|\tilde{s}_1\| = 1$), yelds a correction on the estimate of $\tilde{s}_1$ towards a signal less sensible to the interferent components of the received signal.

In the practical realization of the proposed receiver, the above mentioned correction in the estimate of $\tilde{s}_1$ is

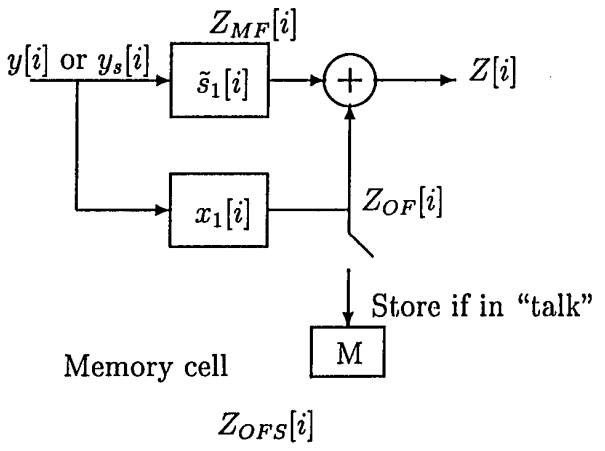


Figure 3: The proposed receiver.

performed when the user of interest is known to be in the *silent state*. Instead, during the *talk state*, the adaptivity mechanism of $\tilde{s}_1$ is "frozen" while it is running the adaptivity branch of the canceller x_1.

It should be also noticed that the minimization of MSiE does not lead to a signal closer to s_1;instead it tends to minimize the components of $\tilde{s}_1$ which lie along one or more interfering signals. The resulting signal, chosen among those unitary normed signals ortogonal to x_1, will have intuitively a relatively stronger component along s_1 (which bears the seeked information) and also its component along the interferers will be reduced.

It is important to point out that the minimization of MSiE is a local one, with its initial value of $\tilde{s}_1$ being the uncorrupted signal $\hat{s}_1$, assigned to the user of interest.

If for any reason tha adaptivity algorithm falls in the attractor region of another signal equally orthogonal to the interference signal but without any component along s_1, the receiver will no longer be able to recover the information. To prevent this problem is under study a technique to limit the amount of corrections performed during silence periods only when an attractor boundary is reached.

Part III
The receiver ...

The structure of the proposed receiver is shown in fig 3.

The stochastic gradient descent algorithm for adaptivity of $x_1[i]$ is fully described in [3] so it is here reported only the final formulation.

For the adaptation rule of $\tilde{s}_1$ it is recalled the expression of $MSiE[\tilde{s}_1]$

$$MSiE[\tilde{s}_1] = E[(< y_s, \tilde{s}_1 > + < y, x_1 >)^2] = \\ = E[< y_s, \tilde{s}_1 >^2] + E[< y, x_1 >^2] + \\ + 2E[< y_s, \tilde{s}_1 >< y, x_1 >] \quad (6)$$

Its uncostrained stochastic gradient ($\tilde{s}_1$ being the variable vector) is

$$\nabla MSiE[\tilde{s}_1] = 2 < y_s, \tilde{s}_1 > y_s + 2 < y, x_1 > y_s = \\ = 2\,(< y_s, \tilde{s}_1 > + < y, x_1 >)\, y_s \quad (7)$$

The component of (7) orthogonal to x_1 is

$$2\,(< y_s, \tilde{s}_1 > + < y, x_1 >)\,(y_s - < y_s, x_1 > x_1) \quad (8)$$

so the adaptation rule for $\tilde{s}_1$ is

$$\tilde{s}_1[i] = \tilde{s}_1[i-1] - \\ - \mu\,(< y_s[i], \tilde{s}_1[i-1] > + < y[i], x_1[i-1] >) \\ (y_s[i] - < y_s[i], x_1[i-1] > x_1[i-1]) \quad (9)$$

From now onwards we will call

$$< y[i], \tilde{s}_1[i-1] > \;=\; Z_{MF}[i] \quad (10)$$
$$< y[i], x_1[i-1] > \;=\; Z_{OF}[i] \quad (11)$$
$$< y_s[i], x_1[i-1] > \;=\; Z_{OFS}[i] \quad (12)$$

It should be noticed that the quantity $< y, x_1 >$ it is not available at the same time of $< y_s, x_1 >$. For that reason, during the "talk" period, the quantity $< y, x_1 >$ is continuously stored in a memory cell called $Z_{OFS}[i]$. When a state transition occurs at i^*, the last value of $Z_{OFS}[i^*]$ is taken as an estimate of $Z_{OFS}[i]$ for $i > i^*$.

The overall adaptation rules for each state of the receiver are:

TALK

$$x_1[i] = x_1[i-1](1 - \mu_x \nu_x) \\ - \mu_x Z[i]\,(y[i] - Z_{MF}[i]\tilde{s}_1[i-1]) \quad (13)$$

$$\tilde{s}_1[i] = \tilde{s}_1[i-1] \quad \text{(holding ...)} \quad (14)$$
$$Z_{OFS}[i] = Z_{OF}[i] \quad \text{(storing ...)} \quad (15)$$

SILENCE

$$\tilde{s}_1[i] = \tilde{s}_1[i-1] - \mu_s \left(Z_{MF}[i] + Z_{OFS}[i]\right) \left(y_s[i] - Z_{OF}[i]x_1[i-1]\right) \quad (16)$$

$$x_1[i] = x_1[i-1] \quad \text{(holding ...)} \quad (17)$$

$$Z_{OFS}[i] = Z_{OFS}[i-1] \quad \text{(holding ...)} \quad (18)$$

where

μ_x, μ_y are the adaptation steps of the two adaptivity branches. Under stationary conditions of the stochastic processes involved in the reception, the algorithm is conducted to the solution [2] if $\mu_{s,x} = \frac{1}{i}$. In order to follow the channel variations a lower bound is introduced.

ν is the Lagrange multiplier responsible for the upper bound to the *surplus energy* as described in [3].

Part IV
... its testbed ...

The performance of the receiver and the gain obtained by silence listening is evaluated by computer simulations. The simulated CDMA trasmission system is characterized by:

1. DS-CDMA 31-chips Gold Sequence [1] for the wanted user, unitary energy of the desired user's spreading sequence. If $\hat{s}_1 \quad t \in [0, T_b]$ is the spreading sequence assigned to the first user, the received signal corrupted by multipath is :

$$s_1(t) = \frac{\hat{s}_1 + a\hat{s}_1(t - T_b/2)}{\|\hat{s}_1 + a\hat{s}_1(t - T_b/2)\|}$$

with a being the *multipath index*. The received signal from user 1 is thus

$$b_1(i)s_1(t - iT_b) \qquad t \in [iT_b, (i+1)T_b]$$

The self interference effect is modeled as an additional interfering user which uses a shifted version of the signal assigned to it :

$$A_{K+1}b_{K+1}(i)s_{K+1}(t - iT_b) = ab_{K+1}(i)\hat{s}_1(t + T_b/2 - iT_b) \quad t \in [iT_b, (i+1)T_b] \quad (19)$$

where a is again the *multipath index*.

Number of "real" interfering users (K)	7
$A_k \quad k = 2,,K$	$\sqrt{10}$
Talk/silence periods ratio	1 : 1
Multipath index (a)	1.0

Table 1: Simulation parameters.

2. K interferent users, each with a different 31-chip Gold sequence and an amplitude $A_k \quad (k = 2,,K)$. Each received signal waveform from the interfering users is supposed to have unitary energy. The received signal from the k-th interferer is thus:

$$A_k b_k(i)s_k(t - iT_b) \quad t \in [iT_b, (i+1)T_b], \quad k = 2,,K \quad (20)$$

3. Additive white gaussian zero-mean noise process with variance σ^2.

The performances are computed in terms of *Signal to Interference Ratio* defined as the ratio (in dB) between the power of the information-bearing signal and the power of the interfering signal which passes through the linear receiver as shown in the following formula

$$SIR = \frac{< s_1, c_1 >^2}{\|c_1\|\sigma^2 + \sum_{k=2}^{K+1} A_k^2 < s_k, c_1 >^2} \quad (21)$$

In the simulations neither time-variant multipath nor loss of synchronization between the transmitters and the receiver have been considered. At the time of this writing, the ability of the receiver to follow deep fades, it is tested and the results will be presented as soon as possible.

Part V
... and the simulation results.

In fig 4 are shown the values of the SIR of the proposed receiver *versus* the SNR of the first user's signal.

The transitions from talk to silence is performed every 10 symbols of the first user. The simulation parameters are shown in table 1.

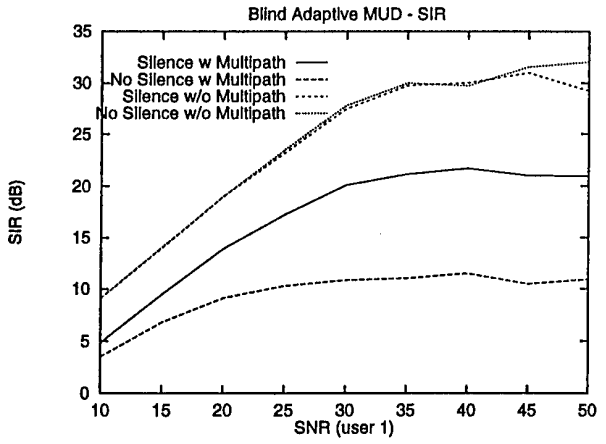


Figure 4: SIR vs SNR for user 1.

The two curves labeled "Silence w Multipath" and "No Silence w Multipath" represent, respectively, the performance of the receiver with and without the silence adaptation algorithm.

The two curves labeled "Silence w/o Multipath" and "No Silence w/o Multipath" represent the performance in case of a reception not corrupted by any multipath effects, that is with $a = 0$.

Part VI
Appendix

Here is derived here the expression for the *Mean Output Silence Energy*. Let's start from the definition of MSiE as defined in (4) :

$$E[(< y_s, \tilde{s}_1 > + < y, x_1 >)^2] =$$
$$= E[< y_s, \tilde{s}_1 >^2 + < y, x_1 >^2 + 2 < y_s, \tilde{s}_1 >< y, x_1 >] =$$
$$= E[\left(\sum_{k=2}^{K} A_k b'_k < s_k, \tilde{s}_1 >\right)^2 + \left(\sum_{k=1}^{K} A_k b_k < s_k, x_1 >\right)^2] +$$
$$+ 2\sum_{k=2}^{K}\sum_{h=1}^{K} A_k A_h b'_k b_h < s_k, \tilde{s}_1 >< s_h, x_1 >] =$$

considering only the terms that will survive after the $E[\bullet]$ operator,

$$= E[\sum_{k=2}^{K} A_k^2 < s_k, \tilde{s}_1 >^2 + \cdots + \sum_{k=1}^{K} A_k^2 < s_k, x_1 >^2 + \cdots] =$$
$$= A_1^2 < s_1, x_1 >^2 + \sum_{k=2}^{K} A_k^2 \left(< s_k, \tilde{s}_1 >^2 + < s_k, x_1 >^2\right) =$$

and considering the parts which vary with $\tilde{s}_1$,

$$= \sum_{k=2}^{K} A_k^2 < s_k, \tilde{s}_1 >^2. \qquad (22)$$

1. REFERENCES

[1] F. D. Garber and M. B. Pursley. "Optimal phases of maximal sequences for asynchronous spread-spectrum multiplexing". *IEE Electronics Letters*, 16(19):756–757, sep 1980.

[2] László Györfi. "Adaptive Linear Procedures Under General Conditions". *IEEE Transactions on Information Theory*, 30(2):262–267, mar 1984.

[3] Upamanyu Madhow Michael Honig and Sergio Verdú. "Blind Adaptive Multiuser Detection". *IEEE Transactions on Information Theory*, 41:944–960, July 1995.

[4] Ruxandra Lupas and Sergio Verdú. "Linear Multiuser Detectors for Synchronous Code-Division Multiple-Access Channels". *IEEE Transactions on Information Theory*, 35(1):123–136, jan 1989.

[5] Sergio Verdú. "Minimum Probability of Error for Asynchronous Gaussian Multiple-Access Channels". *IEEE Transactions on Information Theory*, 32(1):85–96, jan 1986.

[6] Upamanyu Madhow and Michael L. Honig. "MMSE Interference Suppression for Direct-Sequence Spread-Spectrum CDMA". *IEEE Transactions on Communications*, 42(12):3178–3188, dec 1994.

LIMITED LINEAR CANCELLATION OF MULTIUSER INTERFERENCE IN DS/CDMA ASYNCHRONOUS SYSTEMS

Ángel M. Bravo

Dpto de Teoría de la Señal y Comunicaciones e I. T.
E.T.S.I. Telecomunicación - Universidad de Valladolid
C/Real de Burgos s/n
47011 VALLADOLID - SPAIN
Tel: +3483 423260; fax +3483 423261
e-mail: abravo@tel.uva.es

ABSTRACT

This paper presents a linear cancellation detector for the Gaussian channel, operating over limited time intervals of the received signal, in a similar way to a multi-input multi-output FIR filter. The parameters defining the detector become time invariant, and the conditions to be met by the signatures are stated. A bound for the multiuser interference due to the limited time correlation, and upper and lower bounds for the error probability are obtained. The theoretical bounds and numerical results show that the detector is adequate for systems intended for many users whose amplitudes can be restricted to a given range.

1. INTRODUCTION

In a seminal paper, Verdú [1] demonstrated that the optimal multiuser coherent detector can be implemented using the Viterbi algorithm, which is exponential related to the number of users. For this reason the optimal detector is not suitable for real systems and less complex suboptimal alternatives [2] have been studied. The decorrelating detector, being part of the suboptimal linear family, is not exponential, is near-far resistant and can be implemented as a K-input, K-output, linear time invariant (K is the number of users) filter [3]. Several strategies have been adopted to maximize the efficiency in the implementation of this linear detector, such as the sliding window [4], or the isolation bit insertion [5], where the tridiagonal structure of correlation matrix is used. If the system design allows for a slight level of interference, then it is possible to realize very efficient detectors with limited linear cancellation, as it will be shown in the paper.

This work was partially supported by CICYT, under grant N0 TIC95-0320

2. SYSTEM MODEL

In a CDMA system several users share the medium at the same time. Each user employs a different code -i. e. a different waveform or signature- to carry their own information. Following the notation of Verdú [1], the multiuser signal is given by

$$S(\mathbf{b},t)=\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} b_k(i)\ s_k(t-iT-\tau_k) \qquad (1)$$

where $b_k(i)$ is the bit transmitted by the user k in the ith period. The bit is taken from the binary alphabet: $b_k(i)\in\{-1,1\}$. Moreover, $s_k(t)$ is the signature associated to the user k, T is the bit period, k is the delay associated to the user k and is supposed to be lower than the bit period, K is the number of users and N is the number of bits of each of them.

By designing as $u_{k,i}(t)$ the signal $s_k(t-iT-\tau_k)$ scaled to unit norm, the multiuser signal must be in the subspace generated by the basis $\{u_{k,i}(t)\}$, assuming linearly independence for these signals. The coefficient corresponding to the signal $u_{k,i}(t)$ indicates the bit of the user k in the interval i, weighted by the attenuation suffered along the path. The same information can be obtained by the receiver if it uses the reciprocal basis $\{v_{k,i}(t)\}$, whose generic element is defined by

$$\langle v_{g,l}(t), u_{k,i}(t)\rangle=\begin{cases}1 & \text{for } g=k \text{ and } l=i\\ 0 & \text{otherwise}\end{cases} \qquad (2)$$

The elements of the reciprocal basis can be expressed as linear sums of those of the original basis $\{u_{k,i}(t)\}$, and to this end the following matrices and vectors are defined

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_0 & \mathbf{R}_{-1} & 0 & \cdots & 0 \\ \mathbf{R}_1 & \mathbf{R}_0 & \mathbf{R}_{-1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \mathbf{R}_1 & \mathbf{R}_0 & \mathbf{R}_{-1} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \mathbf{R}_1 & \mathbf{R}_0 & \mathbf{R}_{-1} \\ 0 & \cdots & 0 & \mathbf{R}_1 & \mathbf{R}_0 \end{bmatrix}$$

$$\mathbf{a}^{k,i} = \begin{bmatrix} \mathbf{a}_0^{k,i} \\ \mathbf{a}_1^{k,i} \\ \cdots \\ \mathbf{a}_i^{k,i} \\ \cdots \\ \mathbf{a}_{N-2}^{k,i} \\ \mathbf{a}_{N-1}^{k,i} \end{bmatrix} \quad \mathbf{e}^i = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ \mathbf{e}_k \\ \cdots \\ 0 \\ 0 \end{bmatrix} \quad (\text{pos. } i^{th}) \tag{3}$$

Where $\mathbf{a}_l^{k,i}$ is the coefficients vector of the reciprocal basis element $v_{k,i}(t)$ corresponding to the bit l, and $\mathbf{R}_0$, $\mathbf{R}_1$ and $\mathbf{R}_{-1}$ are the correlation matrices for lags 0, 1, and -1, respectively. From 2 and the above definitions,the following matrix equation is obtained

$$\mathbf{R}\,\mathbf{a}^{k,i} = \mathbf{e}^i \tag{4}$$

To obtain the solution to 4 the following matrices are defined recursively:

$$\begin{aligned} \mathbf{A}_i &= \mathbf{R}_0 - \mathbf{R}_1\,\mathbf{A}_{i-1}^{-1}\,\mathbf{R}_1^H \\ \mathbf{B}_i &= \mathbf{R}_0 - \mathbf{R}_1^H\,\mathbf{B}_{i-1}^{-1}\,\mathbf{R}_1 \end{aligned} \tag{5}$$

Making substitutions from the first equation in (8) to the i^{th}, and from the last to the i^{th} an easier to solve system is obtained, namely

$$\begin{gathered} \mathbf{a}_{i-1}^{k,i} = -\mathbf{A}_i^{-1}\,\mathbf{R}_{-1}\,\mathbf{a}_i^{k,i} \\ \mathbf{R}_1\,\mathbf{a}_{i-1}^{k,i} + \mathbf{R}_0\,\mathbf{a}_i^{k,i} + \mathbf{R}_{-1}\,\mathbf{a}_{i+1}^{k,i} = \mathbf{e}_k \\ \mathbf{a}_{i+1}^{k,i} = -\mathbf{B}_{N-i-1}^{-1}\,\mathbf{R}_1\,\mathbf{a}_i^{k,i} \end{gathered} \tag{6}$$

2.1. CONVERGENCE OF THE RECURSIVE MATRIX FUNCTIONS

Assuming that the recursion 5 have limits $\mathbf{A} = \lim_{n\to\infty} \mathbf{A}_n$ and $\mathbf{B} = \lim_{n\to\infty} \mathbf{B}_n$, respectively, then if the number of bits tends to infinity, the coefficient vector $\mathbf{a}^{k,i}$ does not depend on the interval i. Therefore the reciprocal element $v_{k,i}(t)$ is invariant against index i shifts and thus, it is not necessary to calculate it for each interval.

The signatures must have correlation matrices $\mathbf{R}_0$ and $\mathbf{R}_1$ which make the recursion 5 to converge in order to guarantee that $v_{k,i}(t)$ is invariant, and arranging in a vector $\mathbf{v}_\mathbf{A}^i$ the elements of $\mathbf{A}_i$ really intervening in the iteration, the referred iteration results: $\mathbf{v}_\mathbf{A}^{i+1} = \mathbf{f}_\mathbf{A}(\mathbf{v}_\mathbf{A}^i)$, where $\mathbf{f}(\cdot)$ is a nonlinear matrix function.

The convergence is determined by the behavior of the function near the stationary point [6] , [7] $\mathbf{v}_E = \mathbf{f}_\mathbf{A}(\mathbf{v}_E)$. In a neighborhood of the stationary point the function $\mathbf{f}_\mathbf{A}(\cdot)$can be substituted by their linear approximation if it has continuous first derivatives, in this case:

$$\mathbf{f}_\mathbf{A}(\mathbf{x}) = \mathbf{f}_\mathbf{A}(\mathbf{v}_E) + \mathbf{Df}_\mathbf{A}(\mathbf{v}_E)(\mathbf{x} - \mathbf{v}_E)$$

where $\mathbf{Df}_\mathbf{A}(\mathbf{v}_E)$ is the Jacobian of de vector function $\mathbf{f}_\mathbf{A}(\cdot)$ evaluated at the stationary point $\mathbf{v}_E$, and $\mathbf{x}$ is any point in the neighborhood. It can be considered that $\mathbf{x}$ is obtained by summing a perturbation $\boldsymbol{\xi}$ to $\mathbf{v}_E$. The perturbation in the interval $i+1$ is given by $\boldsymbol{\xi}_{i+1} = \mathbf{Df}_\mathbf{A}(\mathbf{v}_E)\,\boldsymbol{\xi}_i$, then $\boldsymbol{\xi}_{i+1} = (\mathbf{Df}_\mathbf{A}(\mathbf{v}_E))^n \boldsymbol{\xi}_0$, and the succession $\boldsymbol{\xi}_0, \boldsymbol{\xi}_1 \cdots$ converges to $\mathbf{0}$ if and only if $lim_{n\to\infty} \|\boldsymbol{\xi}_n\| \to 0$, where the norms referred, here and in the rest, are supposed 2-norms. The limit tends to 0 if and only if [6] the spectral radius of the Jacobian of $\mathbf{f}_\mathbf{A}(\cdot)$: $\rho_{\mathbf{Df}_\mathbf{A}(\mathbf{v}_E)}$, evaluated in the stationary point, is lower than one:

$$\rho_{\mathbf{Df}_\mathbf{A}(\mathbf{v}_E)} < 1 \tag{7}$$

At the same time, the spectral radius is a direct indication of the convergence rate and allows to compare different set of signatures.

3. REALIZATION OF THE DETECTOR

The process of detection of the ith bit of th kth user is made correlating the received signal r(t) with the element $v_{k,i}(t)$, which is considered invariant:

$$\hat{b}_k(i) = \mathrm{sgn}(\langle \mathrm{r}(t), \mathrm{v}_{k,i}(t) \rangle) \tag{8}$$

From 4 is clear that the number of vector coefficients defining $v_{k,i}(t)$ is equal to the number of bits, and this can be large. It is desirable to deal with a finite number of coefficients in each bit interval. The solution of the system 4, 6, in case of convergence of the matrix recursion, is:

$$\begin{aligned} \mathbf{a}_{i-L-1}^{k,i} &= (-1)^{L+1} \left(\mathbf{A}^{-1}\mathbf{R}_{-1}^H\right)^{L+1} \mathbf{a}_i^{k,i} \\ \mathbf{a}_{i+L+1}^{k,i} &= (-1)^{L+1} \left(\mathbf{B}^{-1}\mathbf{R}_{-1}\right)^{L+1} \mathbf{a}_i^{k,i} \end{aligned} \tag{9}$$

and it shows that the coefficient vectors that are $L+1$ intervals apart from vector i vanish progressively if the spectral radius of the matrices $\mathbf{A}^{-1}\mathbf{R}_1^H$ and $\mathbf{B}^{-1}\mathbf{R}_1$ is lower than 1, the norm of the coefficients $\mathbf{a}_{i-L-1}^{k,i}$ and $\mathbf{a}_{i+L+1}^{k,i}$ tends progressively to 0 when L is large enough [6], by which, if a little error is allow in forming $v_{k,i}(t)$, then only a finite set of coefficients, $\left\{\mathbf{a}_l^{k,i}\right\}$,

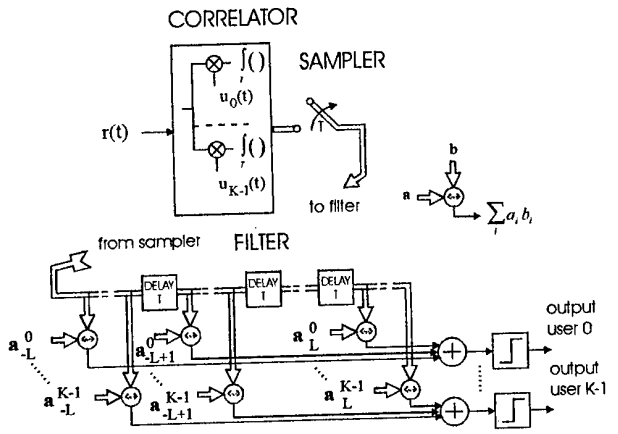


Figure 1: Structure of the detector

$l = i - L \cdots i + L$is needed to define the reciprocal basis for any interval i.

In the receiver the correlation of the received signal $r(t)$ with the K signals $v_{k,i}(t)$ is done, where the last ones are obtained from $\{u_{k,i}(t)\}$ by means of a finite and invariant set of coefficients. The resulting scheme, shown in 1 is made up by a correlator followed by a invariant multi-input, multi-output filter, very similar to a FIR filter, in which the taps are the coefficients $\mathbf{a}_{i-l}^{k,i}$, with delays of one bit period.

4. MULTIUSER INTERFERENCE BOUND

It is supposed that the iteration is near enough to its limit for the error by this being negligible. The bits transmitted by the different users in the j interval can be arranged in a vector:$\mathbf{b}_j$, and the square root of the energies associated to them put in in the diagonal of a matrix: $\mathbf{W}$. The row of transmitted bits, weighted by their corresponding attenuation, is given by

$$(\mathbf{bW}) = \left[\mathbf{b}_0^T\mathbf{W} \;\; \mathbf{b}_1^T\mathbf{W} \;\cdots\; \mathbf{b}_{N-1}^T\mathbf{W}\right] \tag{10}$$

To obtain a bound for the multiuser interference the vector $\mathbf{a}\prime^{k,i}$ is defined. It is made up by the $2L+1$ coefficient vectors, centered around the vector of subscript i, enlarged with $\mathbf{0}$ vectors to complete N components. This vector, $\mathbf{a}\prime^{k,i}$, can be expressed as the sum of the exact and complete coefficients vector $\mathbf{a}^{k,i}$, plus an error vector

$$\mathbf{a}\prime^{k,i} = \mathbf{a}^{k,i} + \mathbf{er}^{k,i} \tag{11}$$

The bit estimated by the detector can be expressed as

$$\begin{aligned}\hat{b}\prime_k(i) &= \mathrm{sgn}\left(\langle r(t), v\prime_{k,i}(t)\rangle\right) = \\ &= \mathrm{sgn}\left(S\prime^{k,i} + N\prime^{k,i}\right)\end{aligned} \tag{12}$$

where $S\prime^{k,i}$ is the contribution of the multiuser signal to the detected bit and $N\prime^{k,i}$ groups the effect of the noise.

Using 11 and 12 it is possible to express explicitly the effect of an imperfect interference cancellation due to not having into account the N intervals in forming the reciprocal basis. Thus

$$\begin{aligned}S\prime^{k,i} &= (\mathbf{bW})\;\mathbf{R}\,\mathbf{a}^{k,i} + (\mathbf{bW})\;\mathbf{R}\;\mathbf{er}^{k,i} = \\ &= \sqrt{w_k}\,b_k(i) + I_k(i)\end{aligned} \tag{13}$$

so that, the contribution of the multiuser signal to the detected bit can be decomposed in two parts: one due to a perfect interference cancellation that in absence of noise would lead to a perfect detection, and other, $I_k(i)$, that represents the multiuser interference to the user k in the interval i due to the limited cancellation.

In the computation of the coefficients, the number of iterations in the recursions 5 is considered large enough for the errors to be negligible because the price paid for it only increases the number of iterations done, but not the circuitry. The error vector is then

$$\begin{aligned}\mathbf{er}^{k,i} = \;& [-\mathbf{a}_0^{k,i} \cdots -\mathbf{a}_{i-L-1}^{k,i}\;\mathbf{0}\cdots \\ & \cdots\mathbf{0}\;-\mathbf{a}_{i+L+1}^{k,i}\cdots -\mathbf{a}_{N-1}^{k,i}]^T\end{aligned} \tag{14}$$

From this, it is clear that the multiuser interference can be decomposed in two terms, one of them affected by coefficients of the past, backward interference, $I_{k,i}^-$ and the other affected by those of the future, forward interference, $I_{k,i}^+$: $I_k(i) = I_{k,i}^- + I_{k,i}^+$.

Letting the number of bits N tend to infinity, the terms of backward and forward interference, $I_{k,i}^-$ and $I_{k,i}^+$ respectively, can be expressed as

$$\begin{aligned}I_{k,i}^- &= \sum_{j=L+1}^{\infty}(\mathbf{b}_{i-j-1}^T\mathbf{W}\,\mathbf{R}_1^H + \mathbf{b}_{i-j}^T\mathbf{W}\,\mathbf{R}_0 + \\ &\qquad +\mathbf{b}_{i-j+1}^T\mathbf{W}\,\mathbf{R}_1\mathbf{a}_{i-j}^{k,i}) \\ I_{k,i}^+ &= \sum_{j=L+1}^{\infty}(\mathbf{b}_{i+j-1}^T\mathbf{W}\,\mathbf{R}_1^H + \mathbf{b}_{i+j}^T\mathbf{W}\,\mathbf{R}_0 + \\ &\qquad +\mathbf{b}_{i+j+1}^T\mathbf{W}\,\mathbf{R}_1\mathbf{a}_{i+j}^{k,i})\end{aligned} \tag{15}$$

From this, a bound of the multiuser interference can be obtained easily with expression 9 substituting for the coefficients in 15. In the numerical calculations the matrices $\mathbf{A}^{-1}\mathbf{R}_1^H$ and $\mathbf{B}^{-1}\mathbf{R}_1$ where simple, and

this is the more likely case. Using this fact, a bound for the multiuser interference is obtained from 15

$$\begin{aligned} I_{msr} &= \max_{k,i} \left(\left| I_{k,i}^{-} \right| + \left| I_{k,i}^{+} \right| \right) \leq \\ &\leq \left\| \mathtt{diag}\left(\mathbf{W}\right) \right\| \left(\left\| \mathbf{R}_1^H \right\| + \left\| \mathbf{R}_1 \right\| + \left\| \mathbf{R}_0 \right\| \right) \times \\ &\times \left\| \left(\mathbf{R}_0 - \mathbf{R}_1 \mathbf{A}^{-1} \mathbf{R}_1^H - \mathbf{R}_1^H \mathbf{B}^{-1} \mathbf{R}_1 \right)^{-1} \right\| \times \\ &\times \left(v\left(\mathbf{V_A}\right) \rho_{\mathbf{A}}^{L+1} + v\left(\mathbf{V_B}\right) \rho_{\mathbf{B}}^{L+1} \right) \end{aligned} \tag{16}$$

where $v\left(\mathbf{V_A}\right)$ and $\rho_{\mathbf{A}}$ are the condition number and the spectral radius of the matrix $\mathbf{A}^{-1}\mathbf{R}_1^H$. $v\left(\mathbf{V_B}\right)$ and $\rho_{\mathbf{B}}$ have similar meaning for the matrix $\mathbf{B}^{-1}\mathbf{R}_1$. From this equation, the following bounds for the error probability for the user k, P_k are obtained in a straightforward way, assuming equally likely bits,

$$Q\left(\frac{\sqrt{w_k} + I_{msr}}{\sigma \left\| v\prime_{k,i}\left(t\right) \right\|} \right) < P_k < Q\left(\frac{\sqrt{w_k} - I_{msr}}{\sigma \left\| v\prime_{k,i}\left(t\right) \right\|} \right) \tag{17}$$

$$\text{where } Q\left(x\right) = \int_x^{\infty} \left(1/\sqrt{2\pi} \right) e^{-t^2/2}\, dt$$

with σ^2 the spectral density of the noise.

5. NUMERICAL RESULTS

Several numerical experiments have been carried out in order to validate the theoretical results and to increase the knowledge about the behavior of the detector. The experiments where carried out with 8, 16, 32 and 64 users using different sets of signatures, binary and non binary. In Fig.2 the results corresponding to 16 non binary signatures are shown. From the results is clear that the simulated maximum multiuser interference follows the same exponential trend that its theoretical bound, but the theoretical bound is always larger.

It is also clear, from the experiments and the theoretical bound, that only a small number of coefficients is needed to maintain the interference below a value fixed in advance.

6. CONCLUSIONS

A method for the limited linear cancellation of multiuser interference has been presented and also a bound for the multiuser interference has been given The theoretical analysis and the numerical results shown that the method of limited linear cancellation of the multiuser interference, here presented, is a good alternative for the implementation of DS/CDMA systems over Gaussian channels.

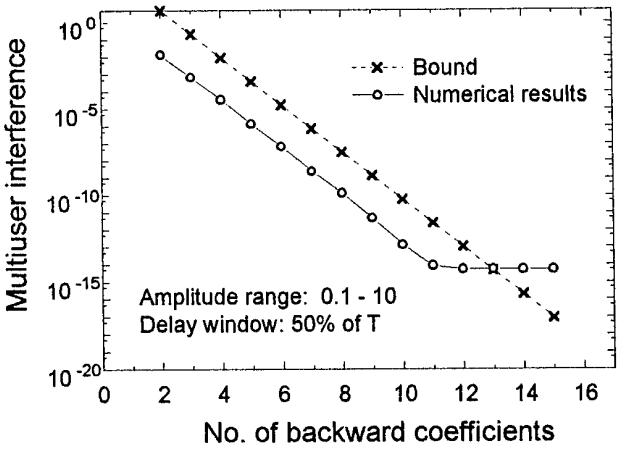


Figure 2: Maximum multiuser interference for 16 Gaussian signatures

7. REFERENCES

[1] S. Verdú , "Minimum probability of error for asynchronous Gaussian multiple-access channels," IEEE Trans. Inform. Theory, vol. IT-32, pp. 85-96, Jan. 1986

[2] A. Duel-Hallen, J. Holtxman and Z. Zvonar, "Multiuser detection for CDMA systems," IEEE Personal Communications, pp. 46-58, Apr. 1995.

[3] R. Lupas and S. Verdú , "Near-far resistance of multiuser detectors in asynchronous channels," IEEE Trans. Commun., vol. COM-38, pp. 496-508, Apr. 1990.

[4] S. S. H. Wijayasuriya, G. H. Norton and J. P. McGeehan, "Sliding window decorrelating algorithm for DS-CDMA receivers," Electronic Letters, vol. 28, no. 17, pp. 1596-1598, Aug. 1992.

[5] F. Zheng and S. K. Barton, "Near-far resistant detection of CDMA signals via isolation bit insertion," IEEE Trans. Commun., vol. COM-43, pp. 1313-1317, Feb/Mar/Apr 1995.

[6] D. M. Young, "Iterative solution of large linear systems," Academic Press, New York, 1971.

[7] L. Perko, "Differential equations and dynamical systems," Springer - Verlag, New York, 1991

BLIND SEPARATION OF SOURCES: STABILITY RESULTS AND COMPARISONS

S. Cruces, R. Martín, J.I. Acha

TESEYCO group, ESI-Telecom, Universidad de Sevilla
Av. Reina Mercedes s/n, 41012-Sevilla, Spain
Ph:(+34) 5-4556872; E-mail: sergio@viento.us.es , acha@obelix.cica.es

ABSTRACT

In this paper we study the blind separation problem of non-Gaussian independent sources. We show how to generalize the separation problem in convolutive mixtures for any number of sources, and we present a new algorithm which solves it. The algorithm is a minimization of the functional quadratic sum of a special set of cross-cumulants between output signals, and can be interpreted as a quasi-Newton search for the zeros of the gradient of this functional. Local asymptotic stability of the algorithm is proved and its relation with alternative cancelation criterion is showed.

We also study another approach to the blind separation problem of two sources for instantaneous mixtures based on robust cumulant estimation criterion.

1. INTRODUCTION

Blind separation of sources can be defined as the problem of identifying and estimating multiple source signals from an array of sensors without knowing the characteristic of the transmission channels. The typical assumptions allowed to resort to are the linearity of the transmission channels and the independence of the sources. Two basic architectures are possible in the separation process, the feed-forward and the feed-backward, shown in Figures 1 and 2, respectively.

In the recent literature of the separation problem, cumulants based methods have played a central role when there are two mixed sources, but less attention was given to a generalized to any number of sources mixture problem.

We will first analyze the separation problem of two sources in instantaneous mixtures, while in section 3, we show how to deal with multiple sources in convolutive and instantaneous mixtures. In sections 4 and 5 we will propose and derive the minimization and cancelation algorithms. And finally, section 6 shows a separation example that corroborates the theoretical results.

2. SEPARATION OF TWO SOURCES

Consider a simplified signal mixing model where by means of two sensors, one observes two instantaneous linear mixtures, $y_1(n)$ and $y_2(n)$, of the two zero-mean sources $x_1(n)$ and $x_2(n)$. The assumptions are that the sources are non-Gaussian and statistically independent. If the channels are noiseless, the sensors outputs are given by

$$\begin{bmatrix} y_1(n) \\ y_2(n) \end{bmatrix} = \begin{bmatrix} 1 & a_{12} \\ a_{21} & 1 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} \quad (1)$$

where a_{12} and a_{21} are the instantaneous mixing coefficients. If we assume that the channel is lossy, and without echoes, the above mixing coefficients satisfy $|a_{ij}| < 1$, for $i, j = 1, 2$, $i \neq j$.

The objective is to obtain the source separation by estimating a 2×2 matrix **B** (see Fig. 3) such that

$$\begin{bmatrix} s_1(n) \\ s_2(n) \end{bmatrix} = \begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix} \begin{bmatrix} y_1(n) \\ y_2(n) \end{bmatrix} = \begin{bmatrix} 1 + a_{21}b_{12} & a_{12} + b_{12} \\ a_{21} + b_{21} & 1 + a_{12}b_{21} \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} \quad (2)$$

The separation is achieved if the coefficients b_{12} and b_{21} are such that (i) $b_{ij} = -a_{ij}$, or (ii) $b_{ij} = -1/a_{ji}$, $i, j = 1, 2$, $i \neq j$. In this paper, only the first solution is treated. For this case, the separated signals, $s_1(n)$ and $s_2(n)$, are related to the sources by

$$s_i(n) = (1 - b_{12}b_{21})x_i(n), \quad i = 1, 2 \quad (3)$$

Note that the estimation of all fourth-order cumulants of $s_1(n)$ and $s_2(n)$ needs the previous estimation of the matrix **B** in a recursive manner, which may be very time-consuming. In addition it implies, in fact, an adaptive environment. We propose an alternative where the cross-cumulants of $s_1(n)$ and $s_2(n)$ are given in term of the cross-cumulant estimation of the sensor outputs $y_1(n)$ and $y_2(n)$.

2.1. FOURTH-ORDER CROSS-CUMULANT ESTIMATION

If the sources $x_1(n)$ and $x_2(n)$ are assumed to be a zero-mean stationary fourth-order non-Gaussian white noise and statistically independent, it can be shown that cross-cumulants of the outputs $s_1(n)$ and $s_2(n)$ can be expressed by [1]

$$\gamma_{13}(s_1,s_2) = h_{11}h_{21}^3\gamma_{x_1} + h_{12}h_{22}^3\gamma_{x_2} \quad (4)$$

$$\gamma_{31}(s_1,s_2) = h_{11}^3h_{21}\gamma_{x_1} + h_{12}^3h_{22}\gamma_{x_2} \quad (5)$$

$$\gamma_{22}(s_1,s_2) = h_{11}^2h_{21}^2\gamma_{x_1} + h_{12}^2h_{22}^2\gamma_{x_2} \quad (6)$$

where, from relation (2)

$$h_{ii} = 1 + a_{ji}b_{ij} \quad (7)$$

$$h_{ij} = a_{ij} + b_{ij}, \quad i,j = 1,2 \quad (8)$$

and γ_{x_i} represents the kurtosis of $x_i(n)$, $i = 1,2$. We note that all expressions in (4)-(6) are given in terms of γ_{x_i}, i.e. the kurtosis of unknown sources. It may be more interesting to put out (4)-(6) in terms of the sensor output cross-cumulants since these quantities are directly measurable. For this purpose, first we obtain expressions for the kurtosis and cross-cumulants of the sensor outputs, which are given by

$$\gamma_{y_1} = \gamma_{x_1} + a_{12}^4\gamma_{x_2} \quad (9)$$

$$\gamma_{y_2} = a_{21}^4\gamma_{x_1} + \gamma_{x_2} \quad (10)$$

$$\gamma_{13}(y_1,y_2) = a_{21}^3\gamma_{x_1} + a_{12}\gamma_{x_2} \quad (11)$$

$$\gamma_{31}(y_1,y_2) = a_{21}\gamma_{x_1} + a_{12}^3\gamma_{x_2} \quad (12)$$

$$\gamma_{22}(y_1,y_2) = a_{21}^2\gamma_{x_1} + a_{12}^2\gamma_{x_2} \quad (13)$$

Next, substituting (7)-(8) in (4)-(6), and using (9)-(13), we have (after some algebra)

$$\gamma_{31}(s_1,s_2) = (3b_{12}^2 + b_{12}^3b_{21})\gamma_{13}(y_1,y_2) + (1 + 3b_{12}b_{21})\gamma_{31}(y_1,y_2) + 3(b_{12} + b_{12}^2b_{21})\gamma_{22}(y_1,y_2) + b_{21}\gamma_{y_1} + b_{12}^3\gamma_{y_2}$$

$$\gamma_{13}(s_1,s_2) = (1 + 3b_{21}b_{12})\gamma_{13}(y_1,y_2) + (3b_{21}^2 + b_{21}^3b_{12})\gamma_{31}(y_1,y_2) + 3(b_{21} + b_{21}^2b_{12})\gamma_{22}(y_1,y_2) + b_{21}^3\gamma_{y_1} + b_{12}\gamma_{y_2}$$

$$\gamma_{13}(s_1,s_2) = 2(b_{12} + b_{12}^2b_{21})\gamma_{13}(y_1,y_2) + 2(b_{21} + b_{21}^2b_{12})\gamma_{31}(y_1,y_2) + (1 + 4b_{12}b_{21} + b_{12}^2b_{21}^2)\gamma_{22}(y_1,y_2) + b_{21}^2\gamma_{y_1} + b_{12}^2\gamma_{y_2}$$

A cumulant cancelation algorithm [2] may now be used to solve the separation problem. Assuming that the signs of the signal kurtosis are the same, the problem is to solve simultaneously $\gamma_{31}(s_1,s_2) = 0$ and $\gamma_{13}(s_1,s_2) = 0$, with spurious solutions removed by cancelling $\gamma_{22}(s_1,s_2) = 0$. For this purpose a Newton- Raphson method or quasi-Newton (global) method are well suited to solve it.

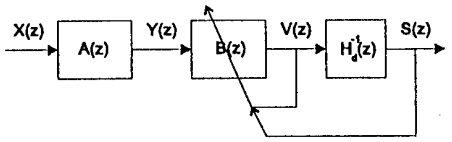


Figure 1: Feed-forward structure.

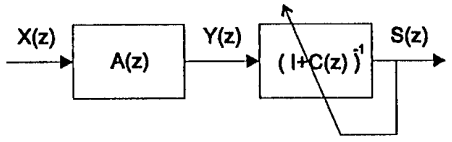


Figure 2: Feed-Backward structure.

3. SEPARATION OF MULTIPLE SOURCES

Suppose we have P independent non-Gaussian sources $x_1[n],\ldots,x_P[n]$ which are mixed through the causal FIR filters matrix $A(z)$ to give $y_1[n],\ldots,y_P[n]$, the mixed signals. Let the transfer function $H(z)$ be the product of the separation and mixing filter matrices.

$$\mathrm{H(z)} = \begin{pmatrix} 1 & B_{12}(z) & \cdots & B_{1P}(z) \\ B_{21}(z) & \ddots & & \vdots \\ \vdots & & \ddots & B_{P-1,P} \\ B_{P1}(z) & \cdots & B_{P,P-1}(z) & 1 \end{pmatrix} \begin{pmatrix} 1 & A_{12}(z) & \cdots & A_{1P}(z) \\ A_{21}(z) & \ddots & & \vdots \\ \vdots & & \ddots & A_{P-1,P} \\ A_{P1}(z) & \cdots & A_{P,P-1}(z) & 1 \end{pmatrix}$$

The filter matrices $A(z)$ and $B(z)$ are supposed to have unit diagonal elements, while $C(z)$ zero diagonal elements, in order to avoid output indetermination in the order of the sources.

If we only want to separate the source signals up to a shaping factor, depending on the structure chosen, it is only necessary to seek for a diagonal transfer function $H(z) = B(z)\cdot A(z)$ or $H(z) = (I + C(z))^{-1}\cdot A(z)$. When $H(z)$ is diagonal, separation is achieved, and we will call this transfer function matrix $H_d(z)$.

If we want to recover the original sources and we use the feed-forward structure, we have to remove own signal distortions introduced in the separation process, so it is need to post-filter $V(z)$ signals through $H_d^{-1}(z)$ to obtain the estimated source $S(z)$; in the case of the feed-backward structure this post-processing is not needed, since $I + C(z)$ can be set to the inverse of the mixing matrix $A(z)$.

Once on the solution the next relation hold:

$$B(z) = H_d(z)\cdot A^{-1}(z) \quad (14)$$

Equating diagonal terms we can obtain the elements of $H_d(z)$ as:

$$H_{ii}(z) = \frac{Det|B(z)|}{Adj_{ii}(B(z))} = \frac{Det|A(z)|}{Adj_{ii}(A(z))} \quad (15)$$

This is an approximate expression that relates $H_{ii}(z)$ and $B(z)$ when we are close to the separation solution. Substituting $H_d(z)$ in equation (14) we obtain for each element of of $B(z)$ the next expression

$$B_{ij}(z) = \frac{Adj_{ij}(A^t(z))}{Adj_{ii}(A(z))} , \qquad (16)$$

that can help us to determine the number of taps necessary for the filter in order to make possible the solution to the separation problem.

When there are only two sources, the expression (16) simplifies to $B_{ij}(z) = -A_{ij}(z)$, so it is only needed for $B(z)$ to have the same number of coefficients of $A(z)$. On the other hand, when the sources are more than two $B_{ij}(z)$ need to be, in general, a rational expression, so it will need an infinity long series of impulse response coefficients. As long as we use FIR filters in $B(z)$ we could not find the true solution to the problem, but we could get as close to it as our degrees of freedom allow. An exception to this argument is the case of instantaneous mixtures, where the solution can be achieved with only one coefficient for each filter. The feed-backward case do not have this kind of problems because the solution is provided by $C(z) = A(z) - I$, and we only need for $C(z)$ the same number of coefficients of $A(z)$ filters to get it.

4. THE MINIMIZATION ALGORITHM

The separation problem could be solved canceling all cross-cumulants of the signals, but this is not a feasible approach. So we will try to cancel only a set of cross-cumulants, on the hope that, in general, if the set is not too small, this will lead to a unique solution which coincides with the separation. In order to cancel the cross-cumulants we use a minimization approach based on a functional ϕ, which is a weighed quadratic sum of some set of cross-cumulants Ω chosen between all pair combinations of output signals under the form $(v_i[n], v_j[n-k])|_{i\neq j}$. Then

$$\phi(B) = \sum_{(\alpha,\beta)\in\Omega} w_{\alpha\beta} \sum_{i=1}^{P} \sum_{\substack{j=1 \\ j\neq i}}^{P} \sum_{k=0}^{L-1} \gamma^2_{\alpha\beta}(v_i[n], v_j[n-k]) \qquad (17)$$

where $w_{\alpha\beta}$ are the weights, $\gamma_{\alpha\beta}$ are the cross-cumulants, and B is the vector of all filters coefficients $\{b_{ij}[k]; k = 0, \cdots, L-1;\ i, j|_{i\neq j} = 1, \cdots, P\}$.

Minimization is done through a quasi-Newton algorithm that searches for the zeros of the gradient of ϕ. The iteration for the algorithm can be set as

$$B|_{n+1} = B|_n - \mu\, \Delta \qquad (18)$$

$$H_B(\phi) \cdot \Delta = \nabla_B \phi , \qquad (19)$$

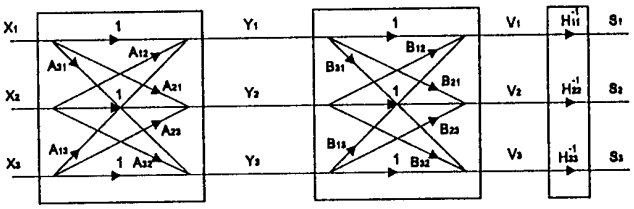


Figure 3: Mixing and Separation filters, in the z domain, for three independent non-Gaussian sources.

where $H_B(\phi)$ is the Hessian matrix, $\nabla_B\phi$ is the gradient vector and μ is a diagonal matrix of adaptation steps. Hessian and gradient can be both approximated for their dominant terms around the separation solution.

4.1. CHOOSING THE CUMULANTS SET

Minimization through the previous algorithm is based on the possibility of having a good estimate for the gradient and the Hessian of ϕ around the solution to the separation problem. This condition does not hold for any kind of cumulants; it can be shown that cross-cumulants with both α and β greater than or equal to two lead to a null second order approximation for the gradient of ϕ, so they can't be used with this method. When α and β are chosen to be equal to 1 we have the decorrelation criterion; this problem was studied in [3] but estimation for the gradient can be shown to be less robust than with other set of cumulants, and the asymptotic stability of the algorithm can be only guaranteed on simplified cases of study. Then, there only remains two possible class of sets: cumulants of the form $\gamma_{1\beta}$ and cumulants of the form $\gamma_{\alpha 1}$. We will choose the first set ($\gamma_{1\beta}$), since it will be shown that with it can be derived conditions that ensure asymptotic stability for the algorithm, while the same does not hold for the other set.

It is interesting to note that for instantaneous mixtures both sets will lead to the same solution, but with our set, Hessian matrix will became diagonal and gradients are well assigned to each coefficient variable, while for the other set, Hessian matrix takes the form of a permutation matrix which corrects the assignment between gradients and variables.

4.2. ASYMPTOTIC STABILITY

We derive the algorithm assuming instantaneous mixtures and non-Gaussian independent sources or, convolutive mixtures and non-Gaussian independent $(1 + \beta_{max})$-order white sources. We also consider FIR causal filters for the sake of simplicity, but the algorithm can

be easily extended to non-causal FIR filters with a few additional considerations. Under this suppositions the output cross-cumulant between two signals $v_i[n]$ and $v_j[n-k]$ has the following relation with the cumulants of the input signals

$$\gamma_{\alpha\beta}(v_i[n], v_j[n-k]) = \sum_{p=1}^{P}\sum_{l=0}^{\infty} h_{ip}^{\alpha}[l+k]\; h_{jp}^{\beta}[l]\; \gamma_{\alpha+\beta}(x_p[n-k-l]) \quad (20)$$

Substituting (20) in equation (17) and taking dominant terms of the gradient vector and Hessian matrix around the solution, it is obtained

$$\frac{\partial \phi}{\partial b_{ij[k]}} \approx 2\textstyle\sum_{(1,\beta)\in\Omega} w_{1\beta}\; \gamma_{1+\beta}(x_j[n-k]) \cdot \sum_{l=0}^{k} h_{jj}^{\beta}[k-l]\; \gamma_{1\beta}(v_i[n], v_j[n-l]) \quad (21)$$

$$\frac{\partial^2 \phi}{\partial b_{ij[k]} \partial b_{ij[m]}} \approx 2\textstyle\sum_{(1,\beta)\in\Omega} w_{1\beta} \cdot \gamma_{1+\beta}(x_j[n-k])\; \gamma_{1+\beta}(x_j[n-m]) \cdot \sum_{l=0}^{\min(k,m)} h_{jj}^{\beta}[k-l]\; h_{jj}^{\beta}[m-l] \quad (22)$$

while the remainder set of derivatives can be approached by null terms. Then, from relation (22), it can be seen that Hessian have a upper triangular structure, and asymptotic stability for the algorithm can be ensured with the following condition:

$$0 < \mu_{ijk} \sum_{(1,\beta)\in\Omega} w_{1\beta}\; \gamma_{1+\beta}^2(x_j[n-k]) \sum_{l=0}^{k} h_{jj}^{2\beta}[k-l] < 1$$

The fastest convergence rate is reached when the adaptation step is

$$\mu_{ijk}^{opt} = \frac{1}{2 \sum\limits_{(1,\beta)\in\Omega} w_{1\beta} \cdot \gamma_{1+\beta}^2(x_j[n-k]) \sum\limits_{l=0}^{k} h_{jj}^{2\beta}[k-l]} \quad (23)$$

but, since we could make errors in the estimates of the terms $\sum h_{jj}^{2\beta}[k-l]$ and $\gamma_{1+\beta}^2(x_j[n-k])$, it is preferable to use $\mu_{ijk} = \mu_0 \cdot \mu_{ijk}^{opt}$, with $0 < \mu_0 < 2$, chosen in a way that always ensures the asymptotic stability of the algorithm at the worst case.

4.3. SIMPLIFICATIONS

When we have instantaneous mixtures, Hessian matrix is diagonal, so it results trivial to invert and the method is easy stated. With convolutional mixtures we can do a simplification in order to avoid solving the equation (19). Assuming dominance of the diagonal elements in the Hessian matrix (this holds for normal mixtures since diagonals terms are always greater than the rest $\sum_{l=0}^{k} h_{jj}^{2\beta}[k-l] \gg \sum_{l=0}^{min(k,m)} h_{jj}^{\beta}[k-l] h_{jj}^{\beta}[m-l]$), then we can approach Hessian as a diagonal matrix. The $h_{jj}[k-l]$ terms can be estimated through equation (15) from the vector of coefficients B, while input cumulants $\gamma_{1+\beta}(x_j[n-k])$ can be replaced by the output cumulants $\gamma_{1+\beta}(s_j[n-k])$. Then the adaptation algorithm simplifies to

$$b_{ij[k]} = b_{ij[k]} - \mu_0 \cdot \frac{\sum\limits_{(1,\beta)\in\Omega} w_{1\beta}\; \gamma_{1+\beta}(s_j[n-k]) \sum\limits_{l=0}^{k} h_{jj}^{\beta}[k-l]\; \gamma_{1\beta}(v_i[n], v_j[n-l])}{\sum\limits_{(1,\beta)\in\Omega} w_{1\beta}\; \gamma_{1+\beta}^2(s_j[n-k]) \sum\limits_{l=0}^{k} h_{jj}^{2\beta}[k-l]} \quad (24)$$

for the convolutive case, and with instantaneous mixtures the method can be simplified further:

$$b_{ij} = b_{ij} - \mu_0\; h_{jj}[0] \cdot \frac{\sum_{(1,\beta)\in\Omega} w_{1\beta}\; \gamma_{1+\beta}(v_j[n])\; \gamma_{1\beta}(v_i[n], v_j[n])}{\sum_{(1,\beta)\in\Omega} w_{1\beta}\; \gamma_{1+\beta}^2(v_j[n])}. \quad (25)$$

5. THE CANCELATION ALGORITHM

A cumulants cancelation algorithm was proposed in [2] to solve the separation problem. We extend this algorithm to the case of multiple sources as

$$b_{ij}[k]|_{n+1} = b_{ij}[k]|_n - \mu_{ijk} \cdot \gamma_{1\beta}(v_i[n], v_j[n-k]) \quad (26)$$
$$(i\neq j)=1,\ldots,P;\; k=0,\ldots,L-1.$$

From the same arguments used in subsection (4.1), we arrive to the conclusion that our best choice for the set of cumulants to cancel must be of the form $\gamma_{1\beta}$, since this lead to a well structured Hessian and then, to a correct assignment between variables and gradients terms. As long as our proposed set of cumulants will be used , Hessian matrix will be upper triangular, and asymptotic stability of the algorithm requires only the following condition

$$0 < \mu_{ijk} \cdot h_{jj}^{\beta}[0]\gamma_{\beta}(x_j[n-k]) < 2 \quad (27)$$
$$(i\neq j)=1,\ldots,P;\; k=0,\ldots,L-1;$$

which ensures a contractive iteration toward the fixed point that constitutes the solution. This condition can be satisfied adjusting the adaptation steps values μ_{ijk} as

$$\mu_{ijk} = \frac{\mu_0}{h_{jj}^{\beta}[0]\gamma_{\beta}(x_j[n-k])} \quad (28)$$

where $0 < \mu_0 < 2$. The most precise are our estimates, closer to 1 can be set μ_0. At $\mu_0 = 1$ we reach the

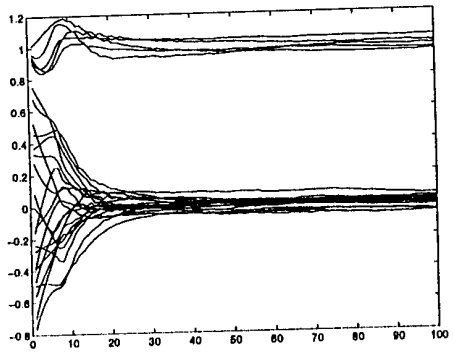


Figure 4: Coefficients of the global transfer function vs. iterations.

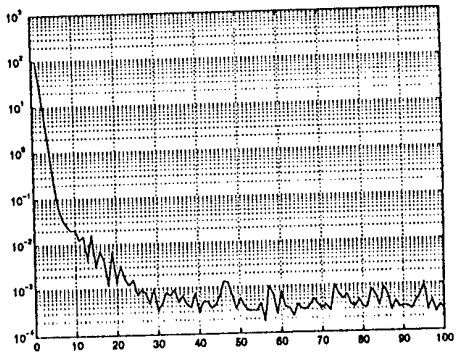


Figure 5: Minimization of ϕ vs. iterations.

maximum speed of convergence. Approaching input sources cumulants $\gamma_\beta(x_j[n-k])$ by the output ones $\gamma_\beta(s_j[n-k])$, we arrive to the cancelation algorithm

$$b_{ij}[k]|_{n+1} = b_{ij}[k]|_n - \mu_0 \frac{\gamma_{1\beta}(v_i[n], v_j[n-k])}{h^\beta_{jj}[0]\gamma_\beta(s_j[n-k])}. \quad (29)$$

It can be seen that the cancelation algorithm we set, is very similar to a minimization algorithm of a quadratic functional with one chosen cumulant, and they coincide in the case of instantaneous mixtures.

6. SEPARATION EXAMPLE

In this section we present an example that corroborate the performance of the proposed minimization and cancelation algorithms. In simulations we use the set of cumulants γ_{13} and γ_{15} since are easy to derive.

We implement for instantaneous mixtures and stationary signals an algorithm that makes a robust estimate of the cumulants through 500 outputs in each iteration. Five source signals of 600 samples each one are chosen to be i.i.d. white uniform noise. The stabilization step size is $\mu_0 = 0.8$, and weighing coefficients are $w_{13} = 1$, $w_{15} = 0.25$. The mixing matrix is:

$$\mathrm{A} = \begin{pmatrix} 1 & -0.42 & -0.57 & 0.36 & -0.27 \\ 0.35 & 1 & -0.87 & -0.18 & -0.30 \\ -0.51 & 0.47 & 1 & -0.16 & 0.75 \\ -0.51 & 0.35 & -0.64 & 1 & 0.03 \\ -0.82 & 0.63 & -0.23 & 0.82 & 1 \end{pmatrix}$$

Separation is reached near the 30 iteration as can be seen in Figure 4 where the value of the coefficients of the global transfer function are shown through iterations. The final transfer function after 100 iterations is:

$$\mathrm{H_t} = \begin{pmatrix} 1.04 & -0.06 & -0.05 & -0.05 & 0.00 \\ 0.03 & 1.01 & 0.02 & 0.01 & -0.01 \\ 0.02 & -0.01 & 0.95 & -0.02 & -0.01 \\ -0.05 & 0.05 & -0.01 & 0.96 & 0.03 \\ -0.01 & 0.00 & 0.02 & 0.00 & 0.98 \end{pmatrix}$$

Minimization of the cumulants functional ϕ is shown in Figure 5. The algorithm is able to lower the value of ϕ in more than five orders of magnitude. The descent stopped when accuracy on the cumulants estimation becomes insufficient, in our example it happens near the 30 iteration.

We also have implemented a stochastic algorithm with lower computational complexity that also works fine with non-stationary signals such as voice.

7. SUMMARY

In this paper we have studied the problem of blind separation of independent non-Gaussian sources. We have presented a method based on a robust cross-cumulant estimation criterion for the case of instantaneous mixtures of two sources. We have derived two algorithms for solving the generalized problem of multiple sources in convolutive mixtures: a minimization algorithm for the quadratic sum of a determined set of cross-cumulants, and the cancelation algorithm. We found conditions for asymptotic stability of both algorithms and gave examples of convergence that corroborate their good performance. Future research will be oriented to the global stability study of the proposed algorithms.

8. REFERENCES

[1] A. Mansour and C. Jutten, "Fourth-Order Criteria for Blind Sources Separation," *IEEE Trans. on Signal Processing*, vol. 43, pp. 2022-2025, Aug. 1995.

[2] H.-L. Nguyen Thi and C. Jutten, "Blind Source Separation of Convolutive Mixtures," *Signal Processing*, vol. 45, pp. 209-229, 1995.

[3] Van Gerven and Van Compernolle, "Signal Separation by Symmetric Adaptive Decorrelation: Stability, Convergence, and Uniqueness," *IEEE Trans. on Signal Processing*, vol. 43, pp. 1602-1612, July 1995.

ADAPTIVE-CORRELATOR RECEIVER FOR DS-CDMA MOBILE RADIO SYSTEM

Rafal Krenz and Krzysztof Wesołowski

Poznań University of Technology
Institute of Electronics and Telecommunication
Piotrowo 3a, PL 60-965 Poznań, Poland
tel. +48 61 782290, fax +48 61 782572
e-mail: Rafal.Krenz@et.put.poznan.pl

ABSTRACT

In this paper an adaptive correlator receiver for a DS-CDMA system is presented. It is well suited for a mobile radio propagation environment which is characterised by a long delay spread of the channel. A decision feedback can be easily implemented in such receiver which can increase the performance of high-speed transmission systems.

1. INTRODUCTION

In the past few years some mobile radio systems exploiting Code Division Multiple Access technique have been proposed in US (IS-95) and Europe (CODIT). They offer high capacity, soft handover capability, immunity to fadings in a channel and many other features important in a cellular environment. In case of a spread-spectrum communication system a well-known RAKE receiver can be applied which distinguishes different propagation paths and practically realizes a time diversity reception [1]. If the receiver is made adaptive it can follow the channel variations and combine the energy of the transmitted signal using a limited number of arms. In a DS-CDMA mobile radio system a dedicated pilot channel can be easily applied in a down-link which can be used for estimation of the channel parameters: delays of the strongest paths as well as the phase shift and the amplitude of each of the paths. Tracking the phase shift and amplitude of the received signal is relatively simple and can be done in each arm independently using a very simple estimator [2]. However, estimation of the path delays must be done in a separate block which finds a given number of the strongest paths in a time window of a fixed length [3].

In this paper we investigate an adaptive correlator receiver, the structure which is equivalent to the RAKE receiver and subsequently we study its new modification resulting from introduction of a decision feedback. The structure of adaptive correlator is well suited for the mobile radio propagation environment where the delay spread can be as long as $30 \div 50 \mu s$ and there are many propagation paths.

The paper is structured as follows. First the mobile radio channel is briefly characterised. Next the structure of the adaptive correlator receiver is presented along with its decision feedback version. Finally, the performance of the receivers is verified by a computer simulation and the results in terms of BER versus E_b/N_0 are shown.

2. THE MOBILE RADIO PROPAGATION ENVIRONMENT

In a mobile radio propagation environment a transmitted signal is severely corrupted by a Doppler shift, shadowing effects and a multipath phenomenon. For each path the amplitude, phase rotation and delay of the received signal are time-varying. The multipath phenomenon results from reflections and diffractions which are quite severe in urban and hilly enviroments. As an effect the

received signal is corrupted by an intersymbol interference limiting the effective bit rate of the transmission system which does not exploit decision feedback or Viterbi detector in the receiver. The multipath phenomenon causes a frequency selective fading as well. The fades can be 20-30 dB deep and result in the burst errors. They can be combatted by implementing FEC coding (block and/or convolutional) and interleaving. Due to the movement of the mobile station the received signal exhibits a Doppler spread. For the system working at 900 MHz and the mobile speed 250 km/h the Doppler shift can be as big as 200 Hz.

3. THE ADAPTIVE CORRELATOR RECEIVER

Due to the implementation constraints the adaptive RAKE receiver contains no more than 3-4 branches. As a result it tracks only a few strongest paths and part of the transmitted signal energy carried by the other paths is neglected. The adaptive correlator inherently exploits all the paths with delays falling within the range specified by the time span of the channel estimator used in the receiver. The basic idea of the receiver is as follows (see Fig. 1). The estimator calculates the instantaneous channel impulse response estimate $\underline{\hat{h}}$. Its coefficients are low-pass filtered to remove the effect of noise. The resulting coefficients $\underline{\tilde{h}}$ are then applied for synthesis of a reference signal in a FIR filter using the despreading sequence u_k^*, generated locally in the receiver. The reference signal which is an estimate of the received pilot signal is correlated with the received data signal r_k and the output of the correlator is sampled once every trasmitted bit. The receiver operates according to the following formula:

$$\hat{a}_m = \mathrm{dec}\left\{\mathrm{Re}\left[\frac{1}{N_c}\cdot\sum_{j=0}^{N_c-1} r_{mN_c+j}\cdot v^*_{mN_c+j}\right]\right\}$$

$$= \mathrm{dec}\left\{\mathrm{Re}\left[\frac{1}{N_c}\cdot\sum_{j=0}^{N_c-1} r_{mN_c+j}\sum_{i=0}^{L-1}\hat{h}_i^*\cdot u^*_{mN_c+j-i}\right]\right\}$$

where N_c is a number of chips per bit (or processing gain). Tracking of the channel is relatively simple since in the investigated system based on the Qualcomm proposal [4] the pilot signal is transmitted in parallel with the user signals and channel estimation can be done using an adaptive or correlative channel estimator.

The RAKE receiver (as well as the adaptive correlator receiver) has been derived under the assumption that the delay spread of the channel is much smaller than the bit duration and thus the intersymbol interference due to multipath may be neglected [1]. This is not fully true in a real mobile radio channels. For instance in the DS-CDMA mobile radio system proposed by Qualcomm [4] the transmitted bit at the basic chip rate 1.228Mcps is $128\cdot 1/1228.8kcps \approx 100\mu s$ long. If the delay spread is equal to $20\mu s$ (which is typical for hilly terrain) the intersymbol interference spans 20% of the bit duration. However, at the higher chip rate 5 Mcps the transmitted bit is only $\approx 20\mu s$ long which is of the order of the delay spread. Thus, the intersymbol interference resulting from the previous transmitted bit must be taken into account. While this is not obvious in the RAKE receiver, in our adaptive correlator the one-bit decision feedback can be implemented in a relatively simple way (see Fig. 2). A feedback signal FIR filter (FSF) must be added which reconstructs the interference using the previously decided bit, the channel impulse response estimate

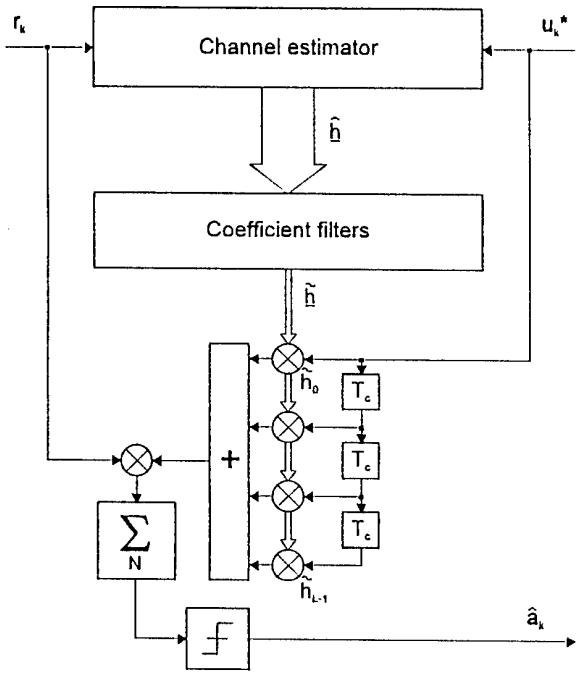


Fig. 1 The block diagram of the adaptive correlator receiver

and the appropriate part of the despreading sequence. The reconstructed interference is next subtracted from the received signal carrying information about the subsequent bit. The receiver works according to the following algorithm:

1. the TDL of the feedback signal filter is filled with zeros
2. the TDL of the despreading sequence filter is fed with the despreading sequence chips and the resulting reference signal is correlated with the received signal samples
3. after N_c chips the decision is made; the decided bit is taken into account during the next correlation period
4. the contents of the TDL of the DSF is copied into the TDL of the FSF and the TDL of the DSF is filled with zeros
5. the reconstructed interference signal is subtracted from the following samples of the received signal (belonging to the next bit) which are next correlated with the reference signal while the TDL of the FSF is fed with zeros
6. goto step 3

The algorithm is managed by the control unit (see Fig. 2).

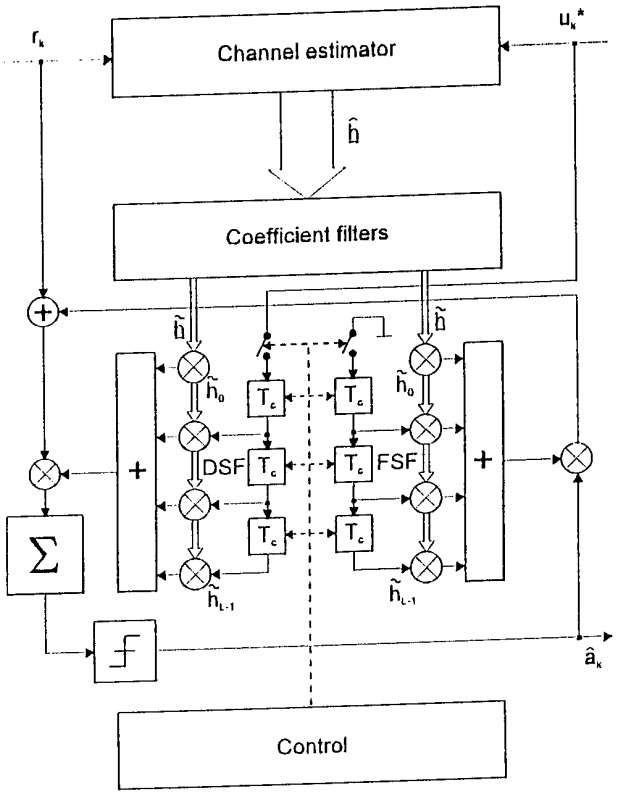


Fig. 2 The block diagram of the decision feedback adaptive correlator receiver

4. SIMULATION RESULTS

In order to verify the performance of the proposed receivers computer simulations have been performed. The simulated system was based on the Qualcomm proposal [4] which employs two stage despreading using Walsh functions depending on a channel number and pseudo-noise sequence characterising a given base station. As a result the despreading sequence u_k^* used in the adaptive correlator receiver is a product of the two sequences. The channel model adopted was based on a GSM recommendation [5]. The channel estimator implemented in the receiver used an LMS algorithm.

The obtained results show that the adaptive correlator receiver with 30 taps estimator covering $\approx 24\mu s$ delay spread performs in a HT100 channel at the basic chip rate 1.228 Mcps identically as the 4-arms adaptive RAKE receiver with the delays adjusted in a time window of the same length (Fig. 3). The gain in comparison to a generic 4-arms RAKE receiver with constant delays between the arms is equal to 3.5-4 dB @ BER=3e-3. The decision feedback version of the adaptive correlator gives basically the same results since the intersymbol interference spans only 17% of the bit duration.

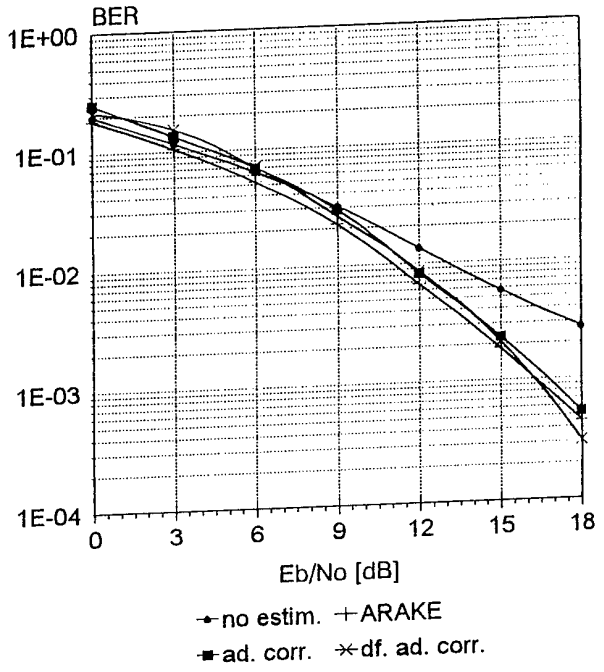


Fig. 3 The performance of the adaptive correlator receiver at 1Mcps chip rate

The advantages of the decision feedback adaptive correlator receiver can be seen at the higher chip rate 4.912 Mcps, where the intersymbol interference spans 70% of the bit duration in the same propagation conditions (Fig. 4). The gain is equal to 2.5 dB @ BER=3e-3 in comparison to the receiver without the decision feedback. The only disadvantage in this case is the length of the estimator which must have 100 taps to cover the whole delay spread of the channel.

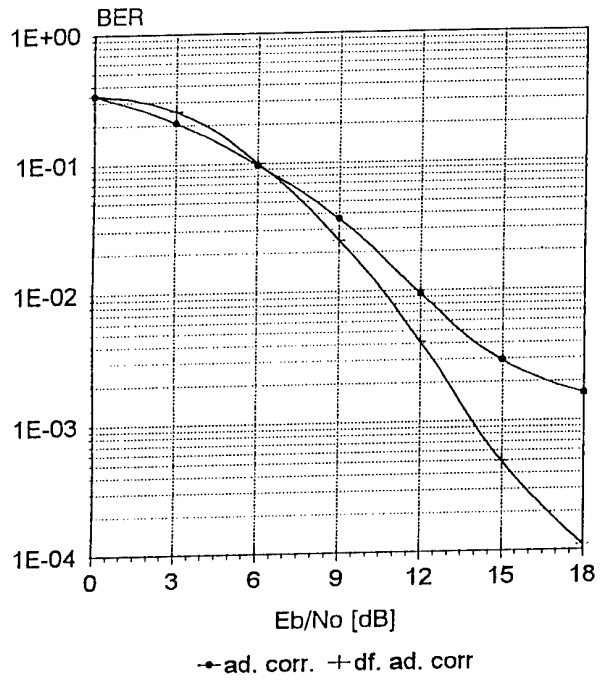


Fig. 4 The performance of the adaptive correlator receiver at 5Mcps chip rate

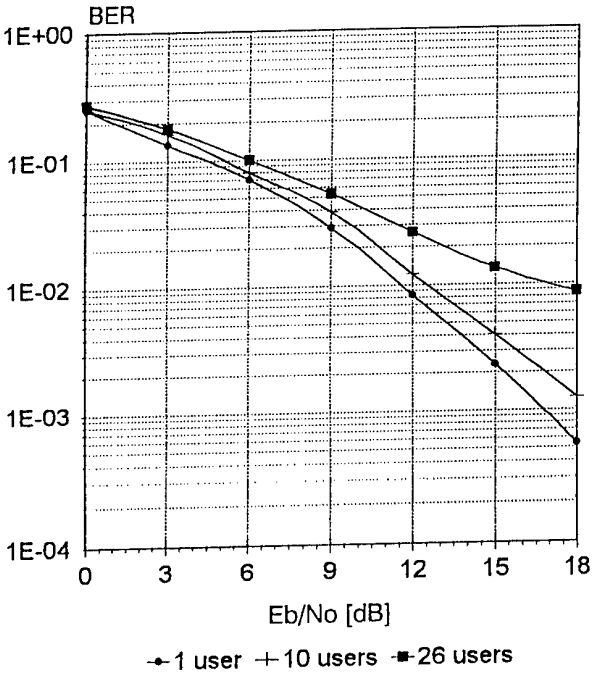


Fig. 5 The performance of the adaptive correlator receiver in the presence of intracell interference

The performance of the adaptive correlator in the presence of the intracell interference is presented on Fig. 5. The degradation due to the increasing number of users in the same cell is comparable to the degradation exhibited by the RAKE receiver.

5. CONCLUSIONS

The presented adaptive correlator receiver may be applied in the transmission systems characterised by a long delay spread of the channel. Although it is more complicated than the adaptive RAKE receiver with comparable performance, it has one important feature. Namely, a decision feedback can be easily implemeted which is indispensable in a high-speed transmission systems, e.g. DS-CDMA mobile radio systems with a chip rate of the order of 5 Mcps.

REFERENCES

[1] J. G. Proakis, "Digital Communications", McGraw-Hill, 1983

[2] R. Krenz, F. Muratore, G. Romano, "Channel Estimation for a DS-CDMA Mobile Radio System with a Coherent Reception", Proc. IEEE VTC '94, vol. 2, pp. 724-728, 1994

[3] R. Krenz, "Adaptive Receivers For DS-CDMA Mobile Radio Systems", Proc. XIth International Microwave Conference MIKON '96, Warszawa 1996

[4] A. Salmasi, K. S. Gilhousen, "On the System Design Aspects of Code Division Multiple Access Applied to Digital Cellular and Personal Communications Networks", Proc. IEEE ICC '91, pp. 57-62, 1991

[5] ETSI-GSM Technical Specification, series 05.05, "Transmission and Reception", June 1991

ON THE ROBUST SPR CONDITION IN ADAPTIVE RECURSIVE SCHEMES

Carlos Mosquera and Fernando Pérez

Dept. Tecnologías de las Comunicaciones.
ETSI Telecomunicación, Universidad de Vigo,
36200 Vigo, SPAIN

ABSTRACT

The problem of finding and LTI filter making a set of plants Strictly Positive Real is presented, given its fundamental role in the convergence of an important class of adaptive recursive algorithms, based on hyperstability concepts. The problem is solved in some important cases with applications in many different contexts.

1. INTRODUCTION

Adaptive infinite-impulse response (IIR) filters are desirable in many situations as an alternative for adaptive finite-impulse response (FIR) filters, for their reduced complexity and improved performance. Important applications include adaptive noise canceling, channel equalization, adaptive differential pulse code modulation, etc. Adaptive techniques for IIR filters have been under investigation during the last years, taking results from the system identification field in many cases, due to the similarities between both areas. Convergence of the algorithms has been the main issue throughout this process; error surfaces are in most cases multimodal, and the analysis of convergence of gradient-based techniques becomes quite hard [1], with convergence to the global minimum not guaranteed in many cases. In addition, most of those procedures need a stability monitoring: otherwise the algorithm may diverge during the adaptation stage. Spurred by the convergence problems, other investigators have borrowed from the control field some tools based on hyperstability, which allow the design of algorithms with proven convergence, provided that a Strictly Positive Real (SPR) condition is satisfied. Such condition involves the poles of the system under study, either unknown or only partially known. Although various suggestions have been made in order to relax the SPR condition, none of them is completely satisfactory given their suboptimality with output disturbance in some cases [2],[3], or conditions imposed on the input in other cases [4].

This paper deals with the robust SPR problem: trying to design a compensator to make the whole set of possible denominators of the system SPR for those cases at which the uncertainty is known, without altering the performance of the algorithm nor imposing conditions on the input. Some numerical results will show how convergence can be achieved when the appropriate compensator $C(z)$, obtained with our synthesis procedure, is used in the adaptive algorithm.

2. ADAPTIVE IIR ALGORITHMS

Many adaptive IIR filtering problems may be addressed in a system identification framework, in which a reference model is hypothesized. The unknown transfer function is assumed rational, and the objective is to construct a rational approximation to the transfer function, based on the input-output measurements, usually noise-corrupted. Figure 1 shows the adaptive filter in a system identification configuration, where θ is the unknown parameter vector. The goal is the minimization of a performance criterion of the error $e(n)$. Traditionally there have been two main approaches to the adaptive IIR filtering problem which correspond to different formulations of the error: equation error and output error methods. We will not consider the first family of methods here; equation error methods have well understood properties, given their similarity to adaptive FIR methods. Their main drawback is the bias in the estimate when a disturbance $v(n)$ is present in the output.

The output error adaptive IIR filter is characterized by the following recursive equation:

$$\hat{y}(n) = \sum_{k=1}^{N} a_k(n)\hat{y}(n-k) + \sum_{k=0}^{M} b_k(n)u(n-k) \quad (1)$$

with $a_k(n)$ and $b_k(n)$ the adaptive parameters of the filter.

The output feedback is the reason why this algorithm is more complex than its equation error counterpart, in which feedback is done with the output of the

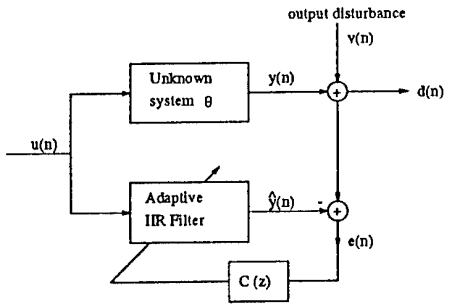


Figure 1: System Identification Configuration

unknown transfer function.

Two main approaches can be made in the output error problem [5]: minimization (gradient descent) viewpoint and stability theory viewpoint. The minimization approach leads to a gradient descent formulation. One of its drawbacks is the sometimes multimodality of the error surface being descended, although in some cases it can be shown to be unimodal [6]. However, the main concern in this approach is the need for on-line stability monitoring of the time-varying AR filter, which filters the regressor vector, and is recomputed at each stage as the estimation of the denominator of the transfer function.

An alternate approach for adapting the parameters of the IIR filter is based on the theory of hyperstability [7], a concept that was developed for the stability analysis of time-varying nonlinear feedback. The adaptive IIR process can be viewed as a linear system having time-varying nonlinear feedback, and is chosen on the basis of assuring that the resulting closed-loop configuration is hyperstable, and hence convergent. This type of algorithms has a filtered version of the output error by $C(z)$, an optional compensator to try to satisfy the SPR condition exposed below. No need of stability checking is required. However, to ensure global convergence, the following SPR condition must be satisfied:

$$Re\left\{\frac{C(z)}{A(z)} - \gamma\right\} > 0, \text{for all } |z| = 1 \qquad (2)$$

where $Re(\cdot)$ denotes real part, and $A(z) = 1 - \sum_{k=1}^{k=N} a_k$ is the denominator of the unknown plant. The scalar γ depends on the specifics of the adaptive algorithm. For the Hyperstable Adaptive Recursive Filter (HARF) and its simplified version (SHARF) [8], two of the main adaptive IIR algorithms based on hyperstability concepts, it suffices with $\gamma = 0$. Other more complex algorithms, such as the Pseudolinear Regression algorithm(PLR), require $\gamma = \frac{1}{2}$, the same as in the original algorithm based on hyperstability ideas, developed by Landau [9] for identification purposes, and adapted for adaptive IIR filtering by Johnson [7].

The main drawback of this type of algorithms is that the satisfaction of the SPR condition is critical for proper algorithm behavior, although convergence can be achieved in some cases even though such a condition is not satisfied [10]. Yet, there is no general method to eliminate the condition entirely, despite the efforts made in that direction [2],[3],[4].

The use of the compensator $C(z)$ requires an a priori estimate of $A(z)$ to satisfy the SPR condition. A first approach would be to start with a different identification scheme, such as least squares, to obtain an approximation of $A(z)$ [11]. In some other cases, an a priori confidence set can be established for either the coefficients [12] or the roots of $A(z)$ [13]. In both cases, satisfaction of the SPR condition for the whole uncertainty set reduces to the study of a finite number of transfer functions. In the following section it is shown how a $C(z)$ can be found to make the whole uncertainty set SPR in some interesting cases, including those cases with nonparametric uncertainty.

3. THE ROBUST SPR CONDITION

The robust SPR problem was raised by Dasgupta and Bhagwat [14] and then by Anderson *et al.* in [12]. In the latter paper they provided a necessary and sufficient condition for the existence of a compensator $C(z)$ making simultaneously several polynomials $A_i(z)$ SPR, namely

$$\max_{\omega\in[0,2\pi)} |arg(A_i(e^{j\omega})) - arg(A_j(e^{j\omega}))| < \pi, \ \forall i, j \quad (3)$$

and they depicted a method to get a minimum-phase FIR $C(z)$ making several plants SPR, but with a numerical procedure which does not provide an a priori bound for the degree of the design.

We will take here a different approach, trying to obtain practical procedures for the synthesis of the compensator $C(z)$. We will present first the case of the simultaneous "SPRization" of a two-member family. There are two main reasons for considering this type of sets. First, it provides much insight to other simultaneous SPRization problems, in the same way as the simultaneous stabilization problem. Second, there are quite a few sets for which finding $C(z)$ making all the members SPR is equivalent to finding $C(z)$ making two specific members SPR, namely, straight line segments

in parameter space, disks in root space, horizontal line segments in root space and vertical line segments in root space [13].

Let us consider two minimum-phase polynomials $A_1(z)$ and $A_2(z)$, with the following definitions:

$$Q(z) = \frac{C(z)A_1(z^{-1}) + C(z^{-1})A_1(z)}{2} \quad (4)$$

$$R(z) = \frac{C(z)A_2(z^{-1}) + C(z^{-1})A_2(z)}{2} \quad (5)$$

$$D(z) = \frac{A_1(z)A_2(z^{-1}) - A_1(z^{-1})A_2(z)}{2} \quad (6)$$

$$E(z) = \frac{A_1(z)A_2(z^{-1}) + A_1(z^{-1})A_2(z)}{2} \quad (7)$$

We have designed an algebraic procedure to obtain $C(z)$ for the two plants case, bounding the degree of the solution in terms of the number of roots of $D(z)$ in (6) on the unit circle. The condition for the existence of $C(z)$ is as follows:

Theorem 1 *A causal and minimum-phase $C(z)$ making $A_1(z)$ and $A_2(z)$ SPR exists if and only if the polynomial $E(z)$ is real and positive at the roots of $D(z)$ on the unit circle* [15] .

The construction of such a $C(z)$ can be done by finding two symmetric positive functions $Q(z)$ and $R(z)$, that as can be noted from (4) and (5), share their sign with the real part of $C(z)/A_1(z)$ and $C(z)/A_2(z)$ respectively. With the definitions above we can obtain $C(z)$ as

$$C(z) = \frac{R(z)A_1(z) - Q(z)A_2(z)}{D(z)} \quad (8)$$

with $R(z)$ and $Q(z)$ symmetric positive functions, satisfying the following interpolation conditions in order to cancel the roots of the denominator:

$$R(\alpha_i)A_1(\alpha_i) - Q(\alpha_i)A_2(\alpha_i) = 0 \quad (9)$$

with $\{\alpha_i\}$ roots of $D(z)$. The resulting filter $C(z)$ could be non minimum-phase (and non causal), depending on the degree of the interpolating functions $R(z)$ and $Q(z)$. It can be made causal and minimum-phase by dividing it by a symmetric positive function containing the roots outside the unit circle and their reciprocals.

To build $R(z)$ and $Q(z)$ we have derived a recursive algorithm, based on the Youla-Saito interpolation algorithm of SPR functions in the Laplace transform domain.

The following corollary indicates those cases at which a minimum-phase FIR $C(z)$ is obtained with our design algorithm:

Corollary 1 *If the number of roots of $D(z)$ on the unit circle is less or equal than four, then an FIR $C(z)$ can be computed.*

For instance, in the case of disks in root space mentioned above, $D(z)$ is a polynomial with only two roots on the unit circle, so an FIR compensator is obtained. It is worth saying that the previous corollary gives a sufficient condition, but not necessary, for the absence of poles of $C(z)$ in $\mathcal{C}\backslash\{0\}$.

The extension of these results to the three plants case involves the use of an additional symmetric positive polynomial $S(z)$ defined for $A_3(z)$:

$$S(z) = \frac{C(z)A_3(z^{-1}) + C(z^{-1})A_3(z)}{2} \quad (10)$$

If $C(z)$ is obtained in (8), then $S(z)$ can be expressed as

$$S(z) = \frac{R(z)D_{13}(z) - Q(z)D_{23}(z)}{D_{12}(z)} \quad (11)$$

with the new polynomials D_{ij} defined as $D_{ij}(z) = (A_i(z)A_j(z^{-1}) - A_i(z^{-1})A_j(z))/2$.

Now, in addition to the interpolation conditions on $R(z)$ and $Q(z)$ shown above, some additional constraints are imposed by the need of having $S(z)$ also SPR:

$$\frac{Q(z)}{R(z)} = \frac{A_1(z)}{A_2(z)} \quad (12)$$

when $D_{12}(z) = 0$ on the unit circle and

$$\frac{Q(z)}{R(z)} \neq \frac{D_{13}(z)}{D_{23}(z)} \quad (13)$$

at the rest of the unit circle. Note that $\frac{D_{13}(z)}{D_{23}(z)} = \frac{A_1(z)}{A_2(z)}$ at the roots of $D_{12}(z)$ on the unit circle (except 1 and -1).

Thus, the general form of the filter which makes simultaneously SPR three plants is of the same form as that for two plants. But some additional conditions (avoidance conditions) must be imposed when obtaining $Q(z)$ and $R(z)$ in (8).

The problem becomes more difficult when dealing with four or more plants: new avoidance conditions

arise with each new plant, while keeping the interpolation conditions. This is a multiple avoidance problem, whose solution would provide a compensator $C(z)$ simultaneously making all the plants SPR.

The extension of this results to the strengthened case does not seem an easy task. In [16] the strengthened robust SPR problem was defined as the search for a monic minimum-phase $C(z)$ such that $C(z)/A(z) - \gamma$ is SPR for the whole uncertainty set. The problem was solved for the continuous-time case, with some comments about the difficulty of its discrete-time counterpart. It was proved that the necessary and sufficient condition of the robust SPR problem ($\gamma = 0$) expressed in (3) is also necessary and sufficient for the general case ($0 < \gamma < 1$). In the two plants discrete-time case, a monic compensator for the case $0 < \gamma < 1$ can be found based again on interpolation conditions. The open problem remains how to get a minimum-phase compensator. In [17] a necessary condition is provided for the solvability of the discrete-time robust SPR problem showing that the results of [16] cannot be extended to the discrete domain.

However, for adaptive algorithms, it seems more interesting the derivation of a compensator $C(z)$ with some constraints on its norm, either $\|C\|_2$ or $\|C\|_\infty$, for disturbance rejection purposes. This is also an open research problem, given the difficulties of combining those constraints with the simultaneous positivity .

Nonparametric uncertainty can also be solved in some cases. If we define the set to be made SPR as $U(A(z), K(z))$, with

$$A(z) = A_0(z) + \Delta A(z), |\Delta A(e^{j\omega})| \leq |K(e^{j\omega})| \quad (14)$$

$\forall\, \omega \in [0, 2\pi)$, then the following lemma can be stated:

Lemma 1 *If the set $U(A(z), K(z))$ is stable, then the transfer function $C(z) = A(z)$ makes the whole set SPR.*

The lemma is illustrated in Fig. 2, where $L(z)$ denotes the bound for the additive term $\frac{-\Delta A(z)}{A(z)+\Delta A(z)}$. Stability of the set $U(A(z), K(z))$, provided that $A(z)$ is stable, is equivalent to the so-called zero exclusion condition [18], if no degree reduction exists. In that case, $(A(z) + \Delta A(z))/C(z)$ will be different than zero for all ω, so SPRness of that term and its inverse is obvious.

The following section will illustrate with a practical example how the previous results can be applied.

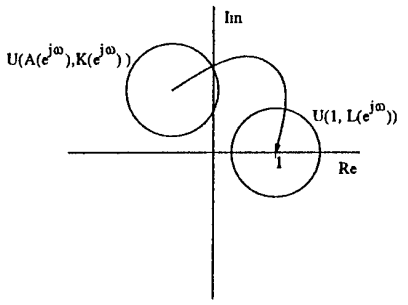


Figure 2: Making SPR a set with nonparametric uncertainty

4. EXPERIMENTAL RESULTS

Figures 3 and 4 show the parameter trajectories when the HARF algorithm is used to identify a second-order plant whose poles are known to lie in a circle centered at 0.6 and with radius 0.35. Using the design algorithm presented in section 3 the compensator $C(z) = 1 - 1.0013z^{-1} + 0.0619z^{-1}$ was obtained. Convergence is achieved for a Signal to Noise Ratio of 20 dB and a step size $\mu = 0.01$ if $C(z)$ is used (Figure 3). Otherwise, the parameters will not converge (Figure 4) to the correct values.

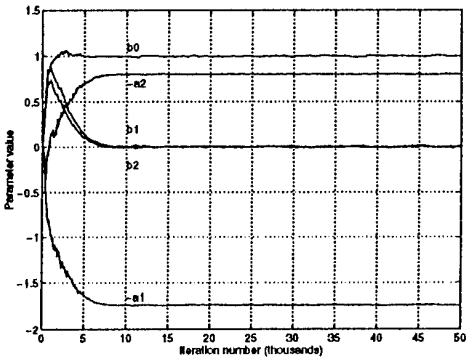


Figure 3: Parameter trajectories when SPRness is satisfied

5. CONCLUSIONS

This paper has presented the links between the robust SPR condition and an important family of adaptive recursive schemes, namely, those based on hyperstability concepts, which require the satisfaction of the SPR property in order to ensure global convergence. An algorithm was provided to design a linear time-invariant filter such that the robust SPR problem is solved for two plants, with some insights for the exten-

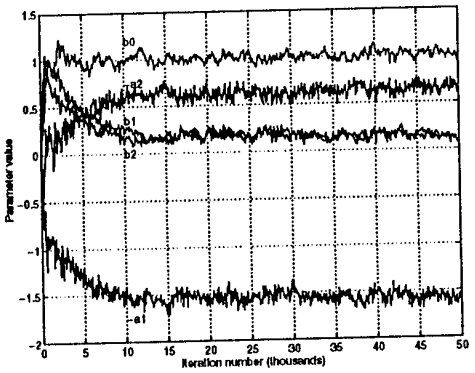


Figure 4: Parameter trajectories when SPRness is not satisfied

sion of that solution to more complex cases, such as a higher number of plants and the strengthened robust SPR problem. It was also solved the robust problem with unstructured uncertainty, for which the number of possible plants is infinite.

6. REFERENCES

[1] S. D. Stearns. Error surfaces of recursive adaptive filters. *IEEE Transactions on Circuits and Systems*, CAS-28:603–606, June 1981.

[2] I.D.Landau. Elimination of the real positivity condition in the design of parallel MRAS. *IEEE Transactions on Automatic Control*, 23:1015–1020, December 1978.

[3] A. Betser and E. Zeheb. Modified output error identification - elimination of the SPR condition. *IEEE Transactions on Automatic Control*, 40:190–193, January 1995.

[4] M. Tomizuka. Parallel MRAS without compensation block. *IEEE Transactions on Automatic Control*, 27:505–506, April 1982.

[5] C.R. Johnson Jr. Adaptive IIR filtering: Current results and open issues. *IEEE Transactions on Information Theory*, 30:237–250, March 1984.

[6] M. Nayeri. A weaker sufficient condition for the unimodality of error surfaces associated with exactly matching adaptive IIR filters. In *22nd Asilomar Conference on Signals, Systems and Computers*, pages 35–38, November 1988.

[7] C.R.Johnson Jr. A convergence proof for a hyperstable adaptive recursive filter. *IEEE Trans. on Information Theory*, 25:745–759, November 1979.

[8] M.G.Larimore, J.R. Treichler, and Jr. C.R.Johnson. SHARF: An algorithm for adapting IIR digital filters. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 28:428–440, August 1980.

[9] I.D.Landau. Unbiased recursive identification using model reference adaptive techniques. *IEEE Trans. on Automatic Control*, 21:194–202, April 1976.

[10] P. A. Regalia. *Adaptive IIR Filtering in Signal Processing and Control.* Marcel Dekker, 1995.

[11] L. Ljung. On positive real transfer functions and the convergence of some recursive schemes. *IEEE Trans. on Automatic Control*, AC-22:539–551, August 1977.

[12] B.D.O. Anderson, S. Dasgupta, P. Khargonekar, F.J. Kraus, and M.Mansour. Robust strict positive realness: Characterization and construction. *IEEE Trans. on Circuits and Systems*, 37:869–876, July 1990.

[13] F. Perez and C. Abdallah. Phase-convex arcs in root space and its application to robust SPR problems. In *Proc. 33rd IEEE Conf. Dec. and Control, Orlando, FL*, pages 3729–3730, 1994.

[14] S. Dasgupta and A. Bhagwat. Conditions for designing strictly positive real transfer functions for adaptive output error identification. *IEEE Trans. on Circuits and Systems*, 34:731–736, July 1987.

[15] F. Perez and C. Mosquera. Algebraic LTI filter synthesis for simultaneously making a convex combination of discrete-time plants SPR. In *Proc. 34th IEEE Conf. Dec. and Control, New Orleans, LA*, pages 780–781, 1995.

[16] B.D.O. Anderson and I.D. Landau. Least squares identification and the robust strict positive real property. *IEEE Transactions on Circuits and Systems I*, 41:601–607, September 1994.

[17] C. Mosquera and F. Perez. A necessary condition for the strengthened robust SPR problem. Technical Report DTC/CMN/300696, Universidad de Vigo, May 1996.

[18] B. R. Barmish. *New tools for robustness of linear systems.* Macmillan, 1994.

AN EM APPROACH TO CHANNEL EQUALIZATION WITH MODULAR NETWORKS

Jesús Cid-Sueiro, Johny Ghattas

ETSI Telecomunicación. Univ. Valladolid. Spain

ABSTRACT

In this paper we discuss the application of the Expectation-Maximization (EM) algorithm to the equalization of digital communication channels with modular neural networks, extending the learning approach discussed by Jacobs and Jordan in [Jor91]. We present a novel algorithm which shows a faster convergence than stochastic gradient rules at a moderate cost, and which can be applied to learning in both supervised and blind mode. Finally, we disscuss the elimination of the hidden variables in the algorithm by means of some previous symbols of the training sequence. Simulation results support the final conclusions.

1. INTRODUCTION

It is a well known property in the communications field that the equalization of digital communication channels is a non-linear problem, even if the channel distortion is linear. Several authors have shown that adaptive non-linear systems based on neural networks can reduce the bit error rate of conventional detectors ([Cid95], [Gel93], [Kec94] and [Mul96] are just some examples). However, it is not easy to keep a moderate computational cost and a reduced training time.

In this paper we go further in the application of modular networks to the symbol detection problem, extending some previous work [Gel93][Cid94][Cid95]. This kind of networks explore the idea of partitioning the input space in subregions so as to assign different modules or experts to each region; working in this way, the learning problem is divided in simpler tasks. The Hierarchical Mixture of Experts (HME) architecture is based on a soft space partition, and it has demonstrated a better performance than backpropagation networks in equalization and other mapping problems [Cid95][Jor91]. In [Cid95], several cost functions for stochastic gradient learning were compared, concluding that the logarithmic cost is the more adequate in an equalization application. The stochastic learning rules are simple, but they usually present a slow convergence speed. Jacobs and Jordan [Jor91] have shown that the Expectation-Maximization (EM) algorithm of Dempster et al.[Dem77] can be applied to this structures reducing the training time, but at a very high computational load. They also propose an on-line algorithm, based on the application of Recursive Least Squares (RLS) algorithms to every module of the HME, that can be applied to regression problems; unfortunately, it shows a poor performance for classification tasks.

The EM algorithm can be simplified if the global maximization step is replaced by a local maximization; in this paper we show that the convergence speed of the resulting algorithm is faster than that of the stochastic minimization of the logarithmic cost function. Moreover, the proposed method can be extended to non-supervised learning, resulting soft-decision directed rules, similar to that proposed by Nowlan in [Now93], which has shown better convergence properties than decision-directed methods in blind equalization.

When the EM algorithm is applied to train modular classifiers, the current symbol is used as a reference for computing the error measurement. However, the references for the gating nets should be constructed from a statistical data model. This paper concludes showing that previous symbols of the channel can be used as references for the gating nets. This allows the application of RLS algorithms to symbol detection.

2. THE HME NETWORK

Fig. 1 shows the structure of a 3-level HME network. Its output is given by

$$y = \sum_{i=0}^{1} \sum_{j=0}^{1} g_i g_{j/i} y_{ij} \qquad (1)$$

where y_{ij}, $g_{1|i}$ and g_1 are the expert outputs, the first level gating nets and the top level gating net, respectively, $g_{0|i} = 1 - g_{1|i}$ and $g_0 = 1 - g_1$. We assume that both expert and gating nets are linear filters with a

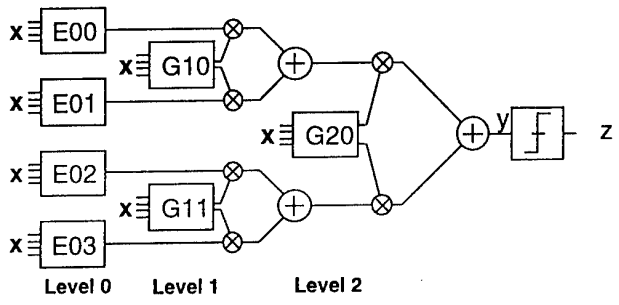


Figure 1: A Hierarchical Mixture of Experts (HME) with three levels

sigmoidal activation unit; thus

$$y_{ij} = \text{sigm}\left(\mathbf{w}_{ij}^T \mathbf{x}\right) \quad (2)$$

$$g_{1/i} = \text{sigm}\left(\mathbf{v}_i^T \mathbf{x}\right) \quad (3)$$

$$g_1 = \text{sigm}\left(\mathbf{v}^T \mathbf{x}\right) \quad (4)$$

where $\mathbf{w}_{ij}$, $\mathbf{v}_i$ and $\mathbf{v}$ are the weights of the different modules.

Although a 3-level network is assummed in the following,, the final results can be generalized to a higher number of levels.

3. THE EM APPROACH.

Let us assume that desired output d (the transmitted symbol in an equalizaton application) is generated according to one of 4 possible Bernoulli distributions with parameters μ_{ij}, $(i = 0, 1,\ j = 0, 1)$. Following [Jor91], we define indicator variable z_{ij} such that $z_{ij} = 0$ for every (i, j) except for the selected model (k, l). The Bernoulli distribution of the selected model can be expressed as follows,

$$P\left(d \mid \mathbf{x}, z_{ij} = \delta_{i-k}\delta_{j-l}\right) = \mu_{kl}^d \left(1 - \mu_{kl}\right)^{1-d} \quad (5)$$

where μ_{ij} are the (i, j)-Bernoulli distribution parameters. Therefore, we can write,

$$P\left(d, z_{ij} \mid \mathbf{x}\right) = \prod_i \prod_j \left(P\left(z_{ij} \mid x\right) P\left(d \mid \mathbf{w}_{ij}, \mathbf{x}\right)\right)^{z_{ij}} \quad (6)$$

Assume that output y_{ij} of the (i, j)-expert of an HME architecture is an estimate of μ_{ij}, and the gating net outputs are used to estimate conditional probabilities $P\left(z_{ij} \mid x\right)$ in such a way that $g_{ij} = g_i g_{j/i} = P\left(z_{ij} \mid x\right)$. It is easy to see that $P\left(d \mid \mathbf{x}\right) = y^d \left(1 - y\right)^{1-d}$, where y is just the output of the HME network. In such a case, it is straightforward to see that the HME becomes the optimal bayesian detector.

To compute the network weights $\mathbf{w}_{ij}$, $\mathbf{v}_i$ and $\mathbf{v}$ leading to such estimates, we take the logarithm of probability model in Eq. (6),

$$L\left(d, \mathbf{x}\right) = \sum_i \sum_j z_{ij} \left(\ln g_i + \ln g_{j/i} + \ln P\left(d \mid \mathbf{w}_{ij}, \mathbf{x}\right)\right) \quad (7)$$

When the indicator variables are known, the HME weights could be estimated maximizing the log-likelihood function in Eq. (7). Unfortunately, this is not the case. The EM algorithm of Dempster et al. [Dem77] solves this problem by taking the expectation of $L\left(d, \mathbf{x}\right)$ (step E) with respect to the so-called *hidden variables* z_{ij} before its maximization (step M). It can be shown [Jor91] that

$$E\left\{L \mid \mathbf{x}, \mathbf{w}_{ij}, \mathbf{v}_i, \mathbf{v}, d\right\} = \sum_i \sum_j h_{ij} \ln g_i g_{j/i} P\left(d \mid \mathbf{w}_{ij}, \mathbf{x}\right) \quad (8)$$

$$h_{ij} = \frac{g_{ij}\left(1 - d - y_{ij}\right)}{1 - d - y} \quad (9)$$

The expected log-likelihood is a non-linear function of the weights. In order to avoid the computational load required by its complete maximization (M step), we replace the global maximization step by a local one computing a single iteration of a stochastic gradient search; in [Jor91], this is done assuming that h_{ij} in Eq.(8) is *independent* of the network parameters; the resulting algorithm is equivalent to a stochastic gradient minimization of a logarithmic cost function. However, it is clear from Eq. (8) that h_{ij} depends on the network weights. The learning rules resulting when this dependence is considered are derived in Appendix A, and shown below

$$\Delta \mathbf{w}_{ij} = \rho f_{ij}\left(d - y_{ij}\right)\mathbf{x} \quad (10)$$

$$\Delta \mathbf{v}_i = \rho\left(f_{i1} - f_i g_{1/i}\right)\mathbf{x}$$

$$\Delta \mathbf{v} = \rho\left(f_1 - g_1\right)\mathbf{x}$$

where ρ is the adaptation step and

$$f_{ij} = h_{ij}\left(1 + \epsilon_{ij}\right) \quad (11)$$

$$f_i = \sum_j f_{ij}$$

Variables ϵ_{ij} are defined as

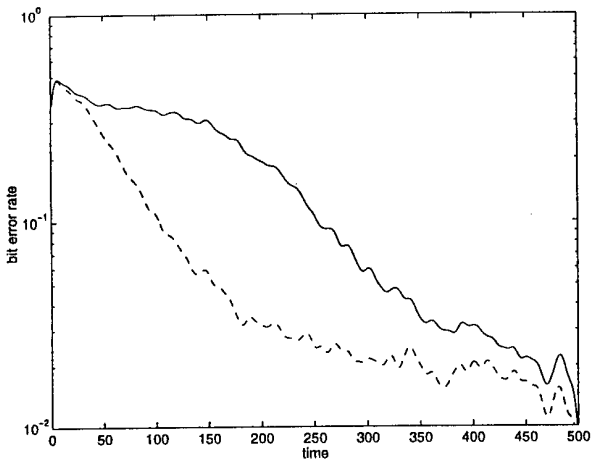


Figure 2: Evolution of the BER vs time, using a logarithmic cost (continuous) and with the proposed method (dashed).

$$\varepsilon_{ij} = \ln h_{ij} - \sum_i \sum_j h_{ij} \ln h_{ij} \qquad (12)$$

and they make the difference between the previous algorithm and the minimization of a logarithmic cost, which can be obtained by assumming $\varepsilon_{ij} = 0$.

As an example, Fig. 2 compares the convergence vs time of both algorithms when training a 4-level HME architecture to equalize the linear non-minimum-phase channel with transfer function $H(z) = 0.5 + z^{-1}$, for $15dB$ SNR The inputs to the network are the last 2 received samples. The adaptation step was optimized independently for each algorithm.The resulting EM algorithm has a significantly faster convergence.

Although it is not observed in the figure, we found that the final bit error rate of the proposed scheme is higher than that of the resulting from using a logarithmic cost. This suggest the possibility of using the EM-based LMS algorithm for a fast startup and, after that, switching to the stochastic minimization of the log cost. As we have seen, this can be done simply making $\varepsilon_{ij} = 0$ in Eq. (11).

3.1. BLIND EQUALIZATION USING EM APPROACH.

Although this line is not explored in this paper, we note that a non-supervised EM algorithm that can be applied to blind equalization can be derived form Eq. (7) by assuming that the desired response, d, is also a hidden variable. Applying expectations over it, we derive the algorithm for an N-level HME Network in App. B. The resulting algorithm is directed by soft-decisions, in a way similar to that of Nowlan and Hinton [Now93], which has shown a better performance than decision directed methods in tracking abrupt channel variations.

4. LIGHTING THE HIDDEN VARIABLES

Each level of the HME network makes consecutive soft partitions of the input space. Note that, in the EM approach, the hidden variables z_{ij} determine which expert network should make the symbol decision. The EM algorithm solves the lack of knowledge about these hidden variables by computing their expected values. However, if they were known, they could be used as output references for the network modules, and each module could be trained separately in a more efficient way. In an equalization application, a posible way to do so is to use the past training symbols as references, as we describe below.

Let us consider, for example, a 3 level HME network. We train the gating net in the top level (level 2) using symbol x_{k-2} as their output reference; level-1 modules with reference x_{k-1}, and level-0 modules (i.e. the expert networks) using x_k. In each level only one expert is trained at a time depending on the value of the past training symbols, which leads to lower numerical cost. Using the same 3 levels HME network example, in level 2, the unique expert is always trained, while in level 1 the first expert is trained only if $x_{k-2} = 0$ and the other is trained only if $x_{k-2} = 1$; in level 0, expert 1 is trained only if $x_{k-2}x_{k-1} = 00$, expert 2 is trained only if $x_{k-2}x_{k-1} = 01$, expert 3 only if $x_{k-2}x_{k-1} = 10$ and expert 4 only if $x_{k-2}x_{k-1} = 11$.

Working in this way, each module (a sigmoidal perceptron) can be trained separately; moreover, as only one filter at each level is updated at every training step, the computational load is reduced,: it increases linearly with the number of levels (i.e. with the logarithm of the number of modules). Also, making a hard partition of the input space (replacing the soft decision devices by hard slicers), each module becomes a linear equalizer which can be trained using standard LMS or RLS algorithms.

Fig. (3) shows that the use of the previous symbols of the training sequence to update the gating nets offers significant faster convergence and lower bit error rate than those based on hidden variables. Linear channel $H(z) = 0.5 + z^{-1}$ with 15dB of signal to noise ratio.

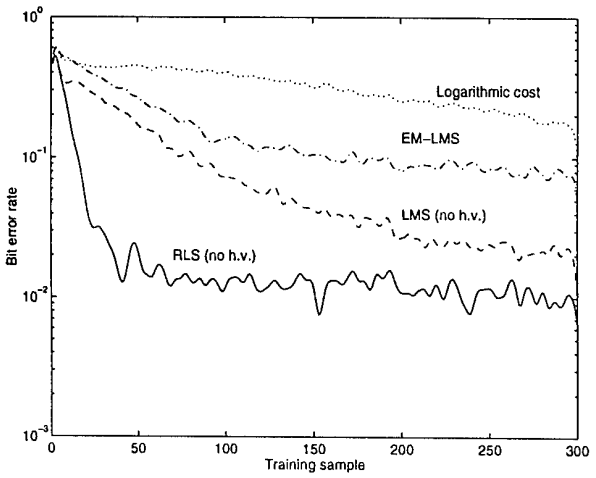


Figure 3: Evolution of the bit error probability during training for the RLS and LMS with diferent references for each level and no hidden variables (h.v.) vs the EM-LMS and the standard stochastic gradient search with a logarithmic cost. The curves are the average of 400 simulations, after smoothing with a low pass filter.

5. CONCLUSIONS AND FURTHER WORK

The main topic of this paper is the problem of finding references for training the gating nets in HME networks. We found that the EM algorithm computes references based on a Bernoulli distribution for the data. If the global M-step is replaced by a local M step, the resulting algorithm has a faster convergence than stochastic gradient search rules. Blind learning rules are also derived, which generalize the soft-decision equalizer proposed by Nowlan [Now93]. On the other hand, we found that, in an equalization application, the previous symbols of the channel can be used as references for the gating nets. Working in this way, the network is not globally optimized, but learning is much faster and standard linear estimation methods can be applied.

The work carried out on this paper opens several lines for future research. While the EM approach can be applied to any classification problem, the use of previous symbols for training gating networks takes advantage of the fact that every sample at the receiver input depends on more that one transmitted symbol. Other applications have this property, and will be explored in the future.

Finally, it is possible to use previous symbols as references for starting up the HME network, tracking the channel variations using the blind EM algorithm. The advantages of this approach over other decision-directed methods are being studied.

6. APPENDIX A: SUPERVISED EM-BASED ALGORITHM.

During the M step of the EM algorithm, function $Q = E\{L \mid \mathbf{x}, \mathbf{w}_{ij}, \mathbf{v}_i, \mathbf{v}, d\}$ given in Eq. (8) has to be maximized with respect to the network parameters. Differenciating Q with respect to P_{ij}, $g_{j|i}$ and g_i respectively we get

$$\begin{aligned}
\nabla_{w_{ij}} Q &= \frac{\partial Q}{\partial P_{ij}} \cdot \nabla_{w_{ij}} P_{ij} = \frac{h_{ij}}{P_{ij}} (1 + \varepsilon_{ij}) \nabla_{w_{ij}} P_{ij} \\
\nabla_{v_i} Q &= \sum_k \frac{\partial Q}{\partial g_{j|i}} \nabla_{v_i} g_{k|i} = \\
&= \sum_k \frac{h_{ik}}{g_{k|i}} (1 + \varepsilon_{ik}) \nabla_{v_i} g_{k|i} \\
\nabla_v Q &= \sum_k \frac{\partial Q}{\partial g_k} \nabla_v g_k = \\
&= \sum_k \frac{1}{g_{k|i}} \sum_n h_{kn} (1 + \varepsilon_{ik}) \nabla_v g_k
\end{aligned}$$

where ε_{ij} has been defined in Eq. (12). Considering a Bernoulli model for the output of each expert, we can write

$$\nabla_{w_{ij}} P_{ij} = (2d - 1) y_{ij} (1 - y_{ij}) x$$

$$\begin{aligned}
\nabla_{v_i} g_{1|i} &= g_{1|i} \left(1 - g_{1|i}\right) x \\
\nabla_{v_i} g_{0|i} &= -g_{1|i} \left(1 - g_{1|i}\right) x
\end{aligned}$$

$$\begin{aligned}
\nabla_v g_1 &= g_1 \left(1 - g_1\right) x \\
\nabla_v g_0 &= -g_1 \left(1 - g_1\right) x
\end{aligned}$$

so the resulting learning rules (10) for a 3-level HME network result.

The generalization of these formulas to an N-level HME network is straightforward, considering that each of the variables g_{ij}, P_{ij}, g_i in such cases will have more indices.

7. APENDIX B: BLIND EM-BASED ALGORITHM.

A blind equalizer results when we assume that d is also a hidden variable. In such case, the expectation of

Eq.(7) with respect to d has to be computed before the M step, arriving to

$$E_d\{Q\} = P(d=0)\,Q(d=0) + P(d=1)\,Q(d=1)$$

Considering that the HME network is an optimum bayesian estimator, the output of the network is $y = P(d=1)$

$$E_d\{Q\} = (1-y)Q(d=0) + yQ(d=1)$$

So, the resulting learning rules are as follows,

$$\begin{aligned} \triangle w_{ij} &= \rho\left(y f_{ij}(d=1) - \overline{f_{ij}} y_{ij}\right) x \\ \triangle v_i &= \rho\left(\overline{f_{ij}} - \overline{f_i} g_{j|i}\right) x \\ \triangle v &= \rho\left(\overline{f_i} - g_i\right) x \end{aligned}$$

where

$$\begin{aligned} \overline{f_{ij}} &= (1-y) f_{ij}(d=0) + y f_{ij}(d=1) \\ \overline{f_i} &= (1-y) f_i(d=0) + y f_i(d=1) = \sum_k \overline{f_{ik}} \end{aligned}$$

8. REFERENCES

[Cid94] J. Cid-Sueiro, A.R. Figueiras-Vidal: The Role of Objective Functions in Modular Classification (with an Equalization Application); *Proc. of the 1st. Int. Conf. on Neural, Parallel and Scientific Computations*, pp. 110-115; Atlanta, GA, May 1995.

[Cid95] J. Cid-Sueiro, A.R. Figueiras-Vidal: Digital Equalization using Modular Neural Networks: an Overview; *Proc. of the 7th. Int. Thyrrhenian Workshop on Dig. Comm.*, pp. 337-345; Viareggio, Italy, Sep. 1995.

[Dem77] A.P. Dempster, N.M. Laird, D.B. Rubin: Maximum Likelihood from Incomplete Data via the EM Algorithm. *J.R. Statistic.* Soc. B, No. 19, pp. 1-38, 1977.

[Gel93] S.B. Gelfand, C.S. Ravishankar, E.J. Delp: Tree-Structured Piecewise Linear Adaptive Equalization; *IEEE Transactions on Communications*, Vol. 41, No. 1, pp. 70-82, Jan. 1993.

[Jor91] M.I. Jordan, R.A. Jacobs: Hierarchical Mixtures of Experts and the EM Algorithm; *Neural Computation*; Vol. 6, pp. 181-214, 1991.

[Kec94] G. Kechriotis, E. Zervas, E.S. Manolakos: Using Recurrent Neural Networks for Adaptive Communication Channel Equalization; *IEEE Trans. on Neural Networks*, Vol. 5, No. 2, pp. 267-278, Mar. 1994.

[Mul96] B. Mulgrew: Applying Radial Basis Functions, *IEEE Signal Processing Magazine*, Vol. 13, No. 2, Mar. 1996.

[Now93] S.J. Nowlan, G.E. Hinton: A Soft Decision-Directed LMS Algorithm for Blind Equalization; *IEEE Transactions on Communications*, Vol. 41, No. 2, pp. 275-279, Feb. 1993.

NONLINEAR RECURSIVE ALGORITHMS FOR DATA TRANSMISSION-EQUALIZATION.

E. Soria-Olivas[(1)], J.Calpe-Maravilla[(1)], A.R. Figueiras-Vidal[(2)].

(1) G.P.D.S Dpto de Informática y Electrónica, Facultad de Físicas.
C/Dr Moliner, 50, 46100 Burjassot (Valencia). Spain
e-mail: emilio.soria@uv.es
(2) D.I., E.P.S. Telecomunicación, Universidad Carlos III
C/Butarque 15, 28911 Leganés (Madrid). Spain
e-mail:anibal@gtts.ssr.upm.es

ABSTRACT

In this paper some recursive adaptive algorithms to be applied to classification problems are studied. These algorithms result from including a non-linearity at the output of the adaptive filters. The aimed application is binary classification which fixes the type of the non-linearity used. The developed algorithms are compared in a typical channel equalization problem proving their good performance.

1. INTRODUCTION.

At the proposed application in this paper, channel equalization, the adaptive system must classify samples into two classes (binary classification). Figure 1 shows a diagram of a transversal adaptive equalizer.

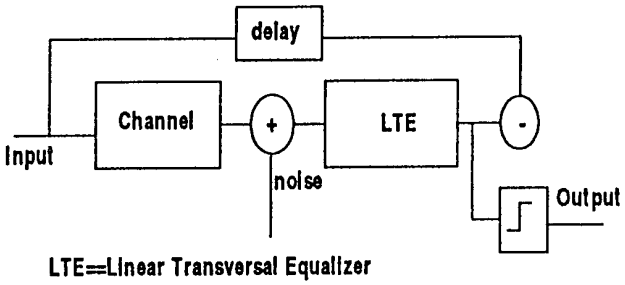


Figure1: Transversal adaptive equalizer diagram

The hard threshold applied to the LTE output classifies the input into one of the two considered classes. By observing Figure 1 we may introduce the first modification to the proposed system. As the goal of our system consists on a classification into two classes, we will apply a nonlinearity before the hard threshold (Figure 2).

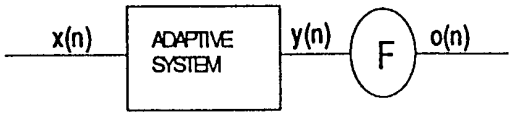


Figure 2: Proposed system representation.

In this paper classes are assumed to be ±1, thus, the proposed function is:

$$F(y)=\begin{cases} y<-1 \Rightarrow F(y)=-1 \\ |y|<1 \Rightarrow F(y)=y \\ y>1 \Rightarrow F(y)=1 \end{cases}$$

In case of analyzed sequence were formed by 0 and 1, the previously defined function would take the value 0 for inputs lower than -1.
The next step for determining the new algorithms is the election of the filter. As the aim lays on applying recursive algorithms (RLS) that reveal much faster than usual LMS, we will use FIR filters [1].
The next section proposes and develops several algorithms applied to the structure given in Figure 2.

2. DEVELOPMENT OF THE ALGORITHMS

2.1 Nonlinear Recursive Least Squares I (NRLSI).

If the adaptive system given in Figure 2 is very near to convergence error and output signal are orthogonal [1]. Hence, if we use an error criterion based on a decreasing weighting, this leads to:

$$\sum_{i=1}^{n}\lambda^{n-i}o(i)d(i)=\sum_{i=1}^{n}\lambda^{n-i}o(i)o(i)$$

Then, we have three different possibilities:

a) If y(i)>1 then system's output o(i) is equal to 1 and the previous sums result as:

$$\sum_{i=1}^{n}\lambda^{n-i}d(i) \quad \text{and} \quad \sum_{i=1}^{n}\lambda^{n-i}.$$

Thus, we may distinguish two sub-cases depending on the first term:

a.1) If they are equal it does not affect the system because it works correctly and it need not any modification.

a.2) If they are different we must modify the system, but we assume it to be near convergence so no modification is performed

b) If y(i) <-1 we have a situation analogous to a)

c) If |y(i)|<1, applying the definition of F, the summing factors result:

$$\sum_{i=1}^{n} \lambda^{n-i} d(i) w^t(i) x(i)$$

and

$$\sum_{i=1}^{n} \lambda^{n-i} d(i) w^t(i) x(i) w^t(i) x(i)$$

and we have to adapt the system. As usual, the starting point for adapting is the minimization of the weighted errors sum, which leads to apply the normal equations, but considering that weights modification occurs in the linear zone exclusively.
In a standard RLS problem, the autocorrelation matrix of the input signal and the cross-correlation matrix between the desired signal and the input are [1]:

$$R(n+1) = \lambda R(n) + X(n+1) X^t(n+1)$$

and

$$g(n+1) = \lambda g(n) + d(n+1) X(n+1)$$

with:

$$X(n) = [x(n).....x(n-L+1)]^t$$

with L equal to the adaptive filter length.
So as to obtain the new algorithm we suppose that from the iteration n to the iteration n+s, no changes have happened i.e, inside that interval, the output of the system remains out of the linear zone, and applying the previous equations iteratively we get to

$$R(n+s) = \lambda^S R(n) + -+ \lambda^{s-1} X(n+1) X^t(n+1).$$

$$g(n+s) = \lambda^S g(n) + -+ \lambda^{s-1} d(n+1) X(n+1).$$

When adapting the filter coefficients we could consider taking all the terms obtained from the previous equations. This procedure would lead us to a "block" version of RLS. The proposed modification consists in neglecting all the terms from the sum except the first two terms. This way, the algorithm discards all the cases out of the linear zone. With this approximation, equations are analogue to those presented in [1] for the RLS but replacing the λ parameter with λ^S, where s gives the number of iterations without changes. A benefit to be remarked is the computational savings: there are no computations when the adaptive filter output comes out of the linear zone.

2.2 Nonlinear Recursive Least Squares II (NRLSII).

The second algorithm is obtained by "smoothing" the first. This results from applying:

$$F(y) = \frac{e^y - 1}{e^y + 1}$$

to the output of the filter. By this way, the output always remains between ±1 in a smoother way (Figure 3).

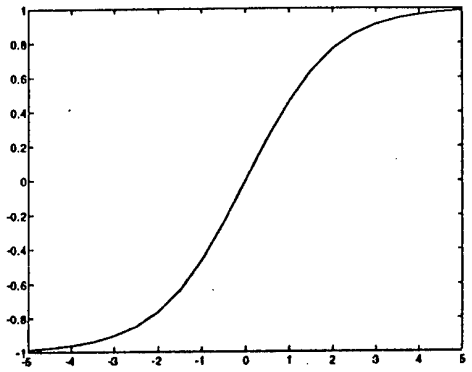


Figure 3: Function used in NRLS II

The next step consists of avoiding the change of the filter coefficients when the output is near the extreme values. This is obtained multiplying the gradient term (error times input) by the factor:

$$1 - o^2$$

With this product we would get the gradient term if we consider the non-linearity action but we have to underline that an RLS algorithm is carried out and after that, adding that factor avoids changes in the filter coefficients. Hence, we have the equations given by [1] for an RLS but including a term when updating weights.
Karanyiannis and Venetsanopoulos [2] propose a similar algorithm, ELEANNE 3. However, there are some differences with the one proposed here because ELEANNE 3 considers the proposed factor for calculating the autocorrelation inverse matrix which is not made here.

2.3 Nonlinear Recursive Least Squares III (NRLSIII).

Following the considerations made in [3], the outputs of the system may be considered as the probability corresponding to the d_j element from the desired output to have a value 0 or 1. Considering mutual independence among samples, we have [3]:

$$P(d|x) = \prod_{j=1}^{n} o_j^{d_j} (1 - o_j)^{(1-d_j)}$$

The logarithm of the previous expression is the cross-entropy between the output and the desired output. Our

system gives to these signals the values ±1. If we consider these new variables we get the expression:

$$P(d|x)=\prod_{j=1}^{n}\left(\frac{1+o_j}{2}\right)^{\left(\frac{1+d_j}{2}\right)}\left(\frac{1-o_j}{2}\right)^{\left(\frac{1-d_j}{2}\right)}$$

This equation is equivalent to:

$$P(d|x)=e^{\sum_{j=1}^{n}\left(\frac{1+d_j}{2}\right)\ln\left(\frac{1+o_j}{2}\right)+\left(\frac{1-d_j}{2}\right)\ln\left(\frac{1-o_j}{2}\right)}$$

handling this expression we have:

$$P(d|x)=e^{\sum_{j=1}^{n}\left(\frac{d_j}{2}\right)\ln\left(\frac{1+o_j}{1-o_j}\right)+\left(\frac{1}{2}\right)\ln\left(\frac{1-o_j^2}{4}\right)}$$

and defining:

$$\ln\left(\frac{1+o_j}{1-o_j}\right)=y_j$$

where y_j is the adaptive filter output. Thus:

$$o_j=\frac{e^{y_j}-1}{e^{y_j}+1}$$

and we have the way to get o_j. Substituting this result in the expression for P(d|x) we get:

$$P(d|x)=e^{\sum_{j=1}^{n}\left(\frac{d_j}{2}\right)y_j+\left(\frac{1}{2}\right)\left(y_j-2\ln\left(1+e^{y_j}\right)\right)}$$

Now we want to maximize the previous function (maximum likelihood) respect to the filter coefficients, W. A possible iterative method for solving this problem is the Newton-Raphson method [4]:

$$W(n+1)=W(n)-\left(\frac{\partial^2 \ln P(d|x)}{\partial W\partial W^t}\right)^{-1}_{(n)}\frac{\partial \ln P(d|x)}{\partial W}_{(n)}$$

Calculating the derivatives:

$$\frac{\partial \ln P(d|x)}{\partial W}_{(n)}=\sum_{j=1}^{n}\left(\frac{d_j-o_j}{2}\right)x_j$$

and:

$$\left(\frac{\partial^2 \ln P(d|x)}{\partial W\partial W^t}\right)_{(n)}=-\frac{1}{4}\sum_{j=1}^{n}(1-o_j^2)x_jx_j^t$$

If we define the variable:

$$u_j=\sqrt{(1-o_j^2)}x_j$$

and

$$R(n)=\sum_{j=1}^{n}(1-o_j^2)x_jx_j^t$$

the second derivative may be obtained in a recursive way because:

$$R(n+1)=R(n)+u_ju_j^t$$

and the matrix-inversion lemma [1] may be applied to this expression so, we have a recursive method to solve classification problems in a similar way to those given in [2]. But here, differently from [2], with the new algorithm, we have obtained a non-linear function for applying the recursive algorithms; besides, no approximations have been made.

We must outline one drawback to this method: when we are very near the optimal system, the gradient term is not zero because all the terms since the beginning of iterations are considered. To solve the problem, only the last term of the sum which gives the gradient is taken into account.

3. EXPERIMENTAL RESULTS

The proposed algorithms have been tested in a channel equalization problem such as the one shown in Figure 1. The simulations assume the transmission channel to have the following transfer function:

$$H_1(z)=1+\frac{1}{2}z^{-1}$$

We study the different convergence speed for the algorithms proposed. In each iteration it has been calculated the number of errors caused by the filter when classifying a fixed number of samples (10^4).

During the comparison, algorithms NRLSI and NRLSII have been considered jointly because both have the same common point and the difference consists of the kind of non-linearity (hard or smooth) applied to the output. These algorithms have been compared to the standard RLS. To be able to distinguish the different speed it has been taken a 5 dB SNR. The three algorithms have the same parameter value (0.99). Furthermore, the results are obtained as an average over 10 experiments .

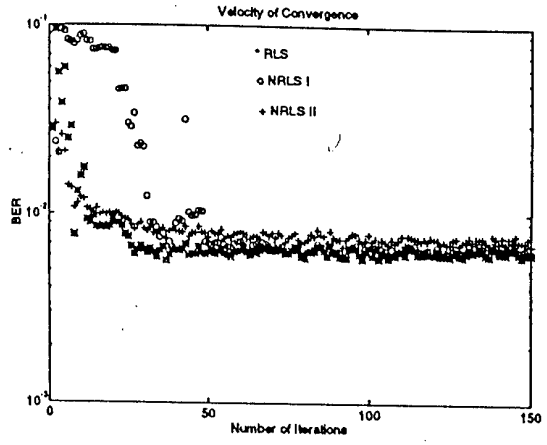


Figure 4: Speed of convergence for RLS, NRLSI and NRLSII.

Figure 4 shows that NRLSI converges in a non-uniform way, while the NRLSII smoothes all these jumps. The three algorithms converge to the same final state, varying their performance only in the first iterations, being the NRLSII faster that the saturated version. The next comparative, concerning the convergence speed, is established among the three algorithms proposed, keeping the same SNR (Figure 5)

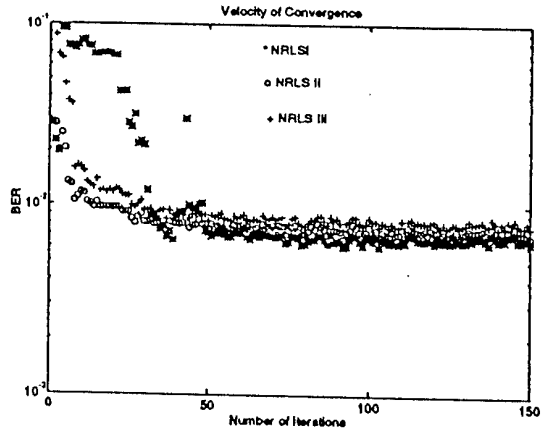


Figure 5: Speed of convergence for NRLSI, NRLSII and NRLSIII.

The comparison among the three algorithms lead us to the same conclusions previously stated: differences are found in the initial iterations, while the final error is the same, in all three cases. The NRLSIII shows a performance which is halfway between the other two.

Once the converge speed has been compared, the probability of error after convergence is studied. Figure 6 shows the probability for the proposed channel and different SNR. Again the same parameter value (0.99) is maintained. The same BER after the convergence is observed in all three variants.

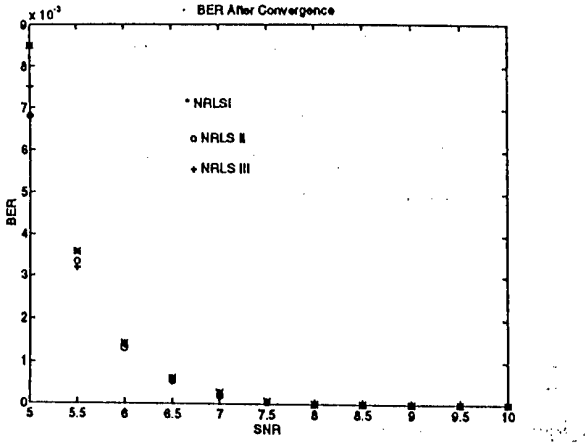


Figure 6: BER after convergence for different SNR

To conclude, Table I below shows the computational savings in the NRLSI due to the adaptation of the filter coefficients in the segment between ±1 solely. It is given the number of iterations in which the weights are updated for a total of 2500 and taking an error threshold under which no adaptation is made. The table has been obtained taking an average on 10 experiments.

SNR (dB)	5	10	15	20
RLS	2467'2	2477'5	2477'1	2477'3
NRLSI	1659'7	1584	1602'7	1605'3

Table I: Number of iterations for different SNR

4. FUTURE WORK

Extensions of this work will focus on the generalizing of the proposed algorithms in order to be applied to multilayer perceptrons. This generalization pretends to apply fast learning algorithms such as RLS to solve the drawback of the low learning speed of these systems.

5. CONCLUSIONS

In this paper three non-linear recursive algorithms for channel equalization have been proposed. Their performance has been verified through simulations. Results show the ability of these systems to carry out the task of equalizing communication channels which gives a first step to extent their use to neural networks.

6. REFERENCES

[1] S. Haykin . "Adaptive Filter Theory". Prentice-Hall 1991.

[2] N.B. Karayiannis, N. Venetsanopoulos. "Efficient Learning Algorithms for Neural Networks (ELEANNE)". IEEE Transc. On Systems, Man, and Cybernetics, Vol 23, NO 5, September/October 1993.

[3] Y. Chauvin, D.E. Rumelhart (Ed): "Backpropagation: Theory, Architectures and Applications". Hillsdale, NJ Lawrence Erlbaum, 1995.

[4] S.M. Kay. "Fundamentals of Statistical Signal Processing". Prentice-Hall, 1993.

AN 'SVD + VITERBI' ALGORITHM FOR ADAPTIVE BLIND EQUALIZATION OF MOBILE RADIO CHANNELS

Piet Vandaele and Marc Moonen

ESAT - Katholieke Universiteit Leuven
K. Mercierlaan 94,
3001 Heverlee -Belgium
tel 32/16/32 18 00
fax 32/16/32 19 86
piet.vandaele@esat.kuleuven.ac.be
marc.moonen@esat.kuleuven.ac.be

ABSTRACT

A fully adaptive algorithm for blind channel equalization is presented. It is based on an adaptive matrix singular value decomposition (SVD) for a (virtual) channel identification type operation, together with the Viterbi algorithm for subsequent symbol detection. True channel modeling, however, is avoided, as will be explained. Simulation results for a GSM type setup are presented.

1. INTRODUCTION

The problem of blind channel identification/equalization using second-order statistics or equivalent deterministic properties of the oversampled channel output has drawn considerable attention recently. Most of the algorithms developed up till now, however, are based on block processing and have a high computational complexity which is an impediment for real-time implementation.

Here we present a new algorithm which is fully adaptive. It has a low computational complexity and furthermore it allows to track very fast varying channels. It will be shown that the performance of previously developed blind equalization algorithms is closely approximated (despite the reduced computational complexity).

Piet Vandaele is a Research Assistant supported by the I.W.T. and M. Moonen is a Research Associate with the Belgian National Fund for Scientific Research (N.F.W.O.). This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven, partly in the framework of a Concerted Action Project of the Flemish Community, entitled *Model-based Information Processing Systems* and partly in the framework of the IT-program of the I.W.T., *Integrating Signal Processing Systems* (ITA/GBO/T23). The scientific responsibility is assumed by its authors.

2. DATA MODEL

The received signal for linear digital modulation over a linear channel with additive noise, is

$$y(t) = \sum_k h(t - kT) \cdot x[k] + n(t)$$

where the $x[\cdot]$ are the transmitted symbols, T is the symbol period, $h(t)$ is the composite channel impulse response (it includes transmitter, channel and receiver filters), and $n(t)$ is additive noise. The channel is assumed to be FIR with duration of approximately LT. With an oversampling factor M, the sampling instants for the received signal are $t_o + (k + \frac{i-1}{M}) \cdot T$ for integer k and $i = 1, 2, \ldots, M$. It is common to use a so-called polyphase description

$$\begin{aligned} y_i[k] &= y(t_o + (k + \frac{i-1}{M}) \cdot T) \\ n_i[k] &= n(t_o + (k + \frac{i-1}{M}) \cdot T) \\ h_i[k] &= h(t_o + (k + \frac{i-1}{M}) \cdot T), \qquad i = 1, 2, \ldots, M \end{aligned}$$

and to view the oversampled received signal as an M-channel output signal at the symbol rate [9, 10]. Define the output vector $\mathbf{y}[k] = [y_1[k] \ldots y_M[k]]^T$, the input vector $\mathbf{x}[k] = [x[k-L] \ldots x[k]]^T$ and the noise vector $\mathbf{n}[k] = [n_1[k] \ldots n_M[k]]^T$ then

$$\mathbf{y}[k] = \underbrace{\begin{bmatrix} h_1[L] & \ldots & h_1[1] & h_1[0] \\ \vdots & & & \vdots \\ h_M[L] & \ldots & h_M[1] & h_M[0] \end{bmatrix}}_{H} \mathbf{x}[k] + \mathbf{n}[k]$$

Spatial oversampling, i.e. using multiple antennas at the receiver, fits into the same framework by considering $y_1[k],\ldots,y_M[k]$ as the outputs of M receiving antennas. From here on we may therefore consider M to be the product of the spatial and temporal oversampling factor.

For the sake of short notation, it is assumed that $\mathbf{n}[k] = 0$. The computational scheme to be presented, involves an SVD which is assumed to be robust against such additive noise [3].

With the above input/output-formula, a data model can be put up as follows (which has been the starting point for many algorithmic developments already). Define

$$Y^{(j)}_{k|k+i-1} = \begin{bmatrix} \mathbf{y}[k] & \mathbf{y}[k+1] & \ldots & \mathbf{y}[k+j-1] \\ \mathbf{y}[k+1] & \mathbf{y}[k+2] & \ldots & \mathbf{y}[k+j] \\ \mathbf{y}[k+2] & \mathbf{y}[k+3] & \ldots & \mathbf{y}[k+j+1] \\ \vdots & \vdots & & \vdots \\ \mathbf{y}[k+i-1] & \mathbf{y}[k+i] & \ldots & \mathbf{y}[k+i+j-2] \end{bmatrix}$$

(the superscript refers to the number of columns, the subscript refers to the time indices in the first column), and with a similar notation

$$X^{(j)}_{k-L|k+i-1} = \begin{bmatrix} x[k-L] & \ldots & x[k-L+j-1] \\ x[k-L+1] & \ldots & x[k-L+j] \\ \vdots & & \vdots \\ x[k+i-1] & \ldots & x[k+i+j-2] \end{bmatrix}$$

then,

$$Y^{(j)}_{k|k+i-1} = \underbrace{\begin{bmatrix} \boxed{H} & 0 & 0 & \ldots \\ 0 & \boxed{H} & 0 & \ldots \\ 0 & 0 & \boxed{H} & \ldots \\ & & \vdots & \\ 0 & 0 & \ldots & \boxed{H} \end{bmatrix}}_{\mathcal{H}} X^{(j)}_{k-L|k+i-1}$$

Here, $Y^{(j)}_{k|k+i-1}$ is a known matrix. The aim is to compute the symbol sequence $X^{(i+j+L-1)}_{k-L|k-L}$ from $Y^{(j)}_{k|k+i-1}$, with or without computing $\mathcal{H}$ explicitly.

3. BLIND EQUALIZATION

Block Processing Algorithm [5]

The algorithm of Liu & Xu [5] is based on a singular value decomposition of $Y^{(j)}_{k|k+i-1}$ (which is a 'short-fat' matrix, *i.e.* with many more columns than rows)

$$Y^{(j)}_{k|k+i-1} = U \cdot \Sigma \cdot V^H$$

Then a symbol sequence is sought that best matches the row space of $Y^{(j)}_{k|k+i-1}$, *i.e.*

$$min_{X^{(j)}_{k-L|k+i-1}} \| X^{(j)}_{k-L|k+i-1} \cdot V^{\perp} \|$$

where $V^{\perp}$ is the (approximate) null space of $Y^{(j)}_{k|k+i-1}$ ($Y^{(j)}_{k|k+i-1} \cdot V^{\perp} \approx 0$), which is extracted from the V-matrix of the SVD. This is equivalent to

$$min_{X^{(i+j+L-1)}_{k-L|k-L}} \| X^{(i+j+L-1)}_{k-L|k-L} \cdot \begin{bmatrix} V^{\perp} & 0 & \ldots & 0 \\ 0 & V^{\perp} & \ldots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & V^{\perp} \end{bmatrix} \|$$

(subject to some constraint to avoid the trivial solution) which may be solved by means of standard least squares techniques.

Adaptive Algorithm

Here, a modification to the above scheme is presented, which is fully adaptive and furthermore employs the Viterbi algorithm in an efficient manner.

A crucial observation is that the $V^{\perp}$ in the block processing algorithm may be viewed as a (virtual) FIR channel (different from the physical channel, obviously) that produces a zero-output when fed with a specific segment of the symbol sequence.

A new SVD is computed in each time step (*i.e.* in each symbol period),

$$Y^{(j)}_{k|k+i-1} = U_k \cdot \Sigma_k \cdot V_k^H$$

This may be viewed as identifying a time-varying FIR channel $V_k^{\perp}$ for which:

$$X^{(j)}_{k-L|k+i-1} \cdot V_k^{\perp} = 0$$

or equivalently,

$$X^{(j)}_{k|k} \cdot \underbrace{[V^{\perp}_{k-i+1} \cdots V^{\perp}_{k} \cdots V^{\perp}_{k+L}]}_{\mathbf{V}^{\perp}_{\mathbf{k}}} = 0$$

This last equation can then be applied in a Viterbi algorithm in a straightforward manner, in order to obtain the original symbol sequence from the knowledge of the 'virtual' FIR channel $\mathbf{V}^{\perp}_{\mathbf{k}}$ and its zero output. State transitions in the Viterbi trellis are then governed by a cost function $\| X^{(j)}_{k|k} \cdot \mathbf{V}^{\perp}_{\mathbf{k}} \|^2_F$ where $\| \cdot \|_F$ denotes the Frobenius norm. Increasing j provides a better noise averaging, but the number of states in the Viterbi algorithm equals 2^{j-1} for a BPSK symbol constellation.

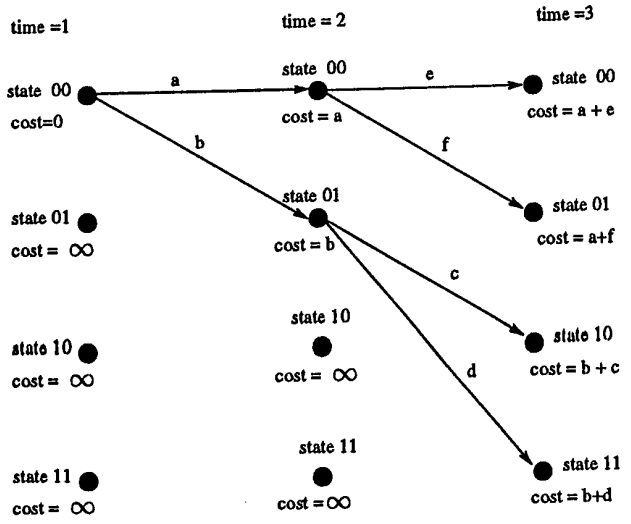


Figure 1: *Viterbi trellis at iteration step 3*

Because of this exponential complexity of the Viterbi algorithm, j should be kept as small as possible. On the other hand, for highly time-varying channels, averaging over long data sequences may not be meaningful and so the usage of smaller matrices may be imposed by practical considerations anyway.

We further reduce the complexity of the Viterbi algorithm in the following way. At iteration step 0, only the first symbol has been received and hence only the smallest digit of the state number is important. We can arbitrarily choose this digit since we can only determine the transmitted symbol sequence up to an unknown constant. As a consequence we can freely choose a starting state in the trellis. This state gets a zero initial cost while all other states get an initial cost ∞. Instead of expanding all paths in the trellis at each time instant, we only extend the 'leave node' having the lowest cost at that iteration step. A 'leave node' is the endpoint of a partially developed path.

Consider figure 1 in which a four-state Viterbi trellis is depicted at iteration step 3. At iteration step 0, we chose state 00 as a reference state and gave it zero initial cost, while all other states got initial cost ∞. Next, in iteration step 1 we extended the paths through the coordinates {(0,0),1} with costs a and b. In the 2nd iteration a was higher than b and so we extended { (0,1),2}, and in step 3 the costs $b+c$ and $b+d$ were both higher than cost a so we extended the path through { (0,0),2}. If we apply this method systematically, we are guaranteed to find the minimum cost path, and this at a lower computational cost (we do not expand all paths any longer). A disadvantage is that the computation time is no longer constant, the upper limit however stays the same as with the full Viterbi algorithm.

For the computation of the $V_k^{\perp}$'s, an efficient SVD-updating algorithm from [7] is used, which performs true recursive (sample-by-sample) processing (instead of block processing). This algorithm has a reduced computational complexity, and is amenable to application-specific hardware implementation [6].

4. SIMULATION MODEL

The measurement model assumes a large number of rays impinging on a uniform linear array. If, in a Rayleigh flat fading channel, the rays have a Gaussian angular distribution [1] around the nominal direction of arrival, then it can be shown [12] that the $M \times 1$ channel impulse response vector h can approximately be modeled as a Gaussian distribution $\mathcal{N}(0, R_{hh})$. The $\{x, y\}$ element of the covariance matrix R_{hh} (denoting the correlation between the signals received on the antennas x and y) equals:

$$R_{hh}(x,y) \approx e^{-2(\pi\sigma_\theta(x-y)d\cos(\theta_0)/\lambda)^2} e^{j2\pi(x-y)d\sin(\theta_0)/\lambda}$$

with σ_θ the spread on the angular distribution, d the inter-antenna-element distance, λ the wavelength and θ_0 the nominal direction of arrival of the signal. With the above expression, simulation samples of h can be constructed as:

$$h = R_{hh}^{1/2} S$$

where S is a $M \times 1$ complex zero mean Gaussian vector with the variance of real and imaginary parts equal to $1/2$.

This model can be extended to frequency selective fading by including several time-clusters of scatterers. The r^{th} cluster is then characterized by its vector h_r and its time delay τ_r. If p denotes the number of clusters then the received signal $y[t]$ can be expressed as:

$$y[t] = \sum_{r=1}^{p} h_r x(t - \tau_r)$$

The modulation scheme at the transmitter is GMSK. Although this is a nonlinear type of modulation, it was shown in [11] that GMSK modulation can be approximated well by a linear channel model, requiring only minor modifications to the receiver algorithms.

5. SIMULATIONS RESULTS

In order to qualify the performance of the proposed algorithm, it is compared against [8] which has proven to be very efficient in combating multi-path effects [2]. The main computational steps of this algorithm are summarized in the appendix.

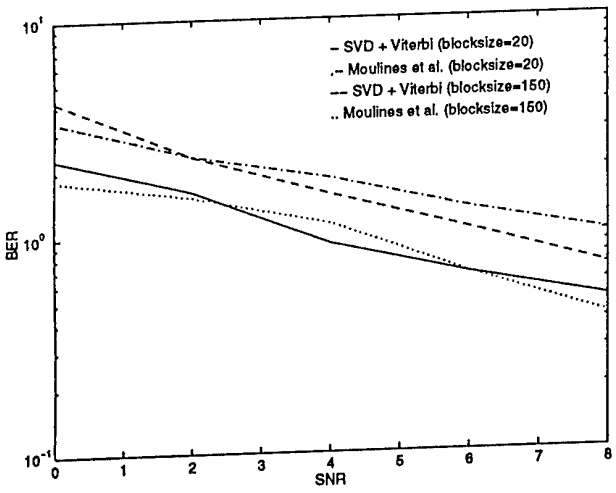


Figure 2: *BER estimation*

In a first setup we took bursts of 150 symbols, generated a time-invariant channel using the above model with the reduced TUx tap specifications [4]. We performed 400 runs and a new channel model was generated in each cycle.
In a second setup we considered the case of a channel for which the time constant is about 20 symbols. In order not to violate the assumptions of the channel burst invariance of algorithm [8] we generated bursts of 20 symbols.
We use 6 antennas which are 2 times oversampled, the antenna-element spacing is $\lambda/2$. The SNR is taken with respect to the input power of the channel, i.e. $SNR = E[x[k]^2]/\sigma^2$. In both setups the parameters i and j of our algorithm were set to respectively 2 and 8. The channel order was estimated from the SVD singular values.
The results are depicted in figure 2. The figure shows that for blocks of 150 symbols, algorithm [8] outperforms the algorithm presented above, mainly due to its ability to average the noise over the full length of the burst. The results for blocks of 20 symbols however, are favorable to the 'SVD+Viterbi' algorithm, indicating the real power of the proposed algorithm.

6. CONCLUSIONS

In this paper we developed a new algorithm for blind channel equalization. The algorithm differs from previous algorithms in that it is a fully adaptive. This approach has two advantages: the computational complexity can be lowered and the model is able to track very fast varying channels. The performance of the algorithm was investigated through experiments in a GSM type setup.

7. REFERENCES

[1] F. Adachi, M.T. Feeney, A.G. Williamson, and J.D. Parsons. Crosscorrelation Between the Envelopes of 900 MHz Signals Received at a Mobile Radio Base Station Site. In *IEE Proceedings*, volume 133, pages 506–512, October 1986.

[2] J. Buisán and E. Biglieri. Algorithms for Blind Identification of Terrestrial Micorwave Radio Channels. In *Proceedings Bayona workshop*, pages 1–2, October 1994.

[3] B. De Moor. The Singular Value Decomposition and Long and Short Spaces of Noisy Matrices. *IEEE Trans. Signal Processing*, pages 2826–2838, September 1993.

[4] European Telecommunications Standards Institute. European digital cellular telecommunications system (phase 2): Radio transmission and reception (GSM 05.05). Technical report, ETSI, Sophia Anitpolis, France, 1994.

[5] H. Liu and G. Xu. A Deterministic Approach to Blind Symbol Estimation. *IEEE Signal Processing Letters*, pages 205–207, December 1994.

[6] M. Moonen, E. Deprettere, I.K. Proudler, and J.G. McWhirter. On the Derivation of Parallel Filter Structures for Adaptive Eigenvalue and Singular Value Decompositions, Detroit (USA). In *Proceedings ICASSP-95*, pages 3247–3250, May 1995.

[7] M. Moonen, P. Van Dooren, and J. Vandewalle. An SVD Updating Algorithm for Subspace Tracking. *SIAM Journal on Matrix Analysis and Applications*, pages 1015–1038, 1992.

[8] E. Moulines, P. Duhamel, J. Cardoso, and S. Mayrargue. Subspace Methods for the Blind Identification of Multichannel FIR Filters. In *Proceedings ICASSP-94*, volume 4, pages 573–576, April 1994.

[9] D.T.M. Slock and C.B. Papadias. Blind Fractionally-Spaced Equalization Based on Cyclostationarity. In *Proc. Vehicular Technology Conf., Stockholm, Sweden*, June 1994.

[10] L. Tong, G. Xu, and T. Kailath. A New Approach to Blind Identification and Equalization of Multipath Channels. In *Proc. of the 25th Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA*, pages 856–860, November 1991.

[11] A.-J. van der Veen and A. Paulraj. Singular Value Analysis of Space-Time Equalization in the GSM Mobile System. In *Proceedings ICASSP*, pages 1073–1076, May 1996.

[12] P. Zetterberg. A Comparison of two Systems for Down Link Communcation with Antenna Arrays at the Base, Report Version. Technical Report IR-S3-SB-9521, Royal Institue of Technology, Sweden, 1995.

Appendix: Moulines et al. [8]

Consider a processing window of length r. Define $\mathbf{y}[k] = [\mathbf{y}_1[k]^H \ldots \mathbf{y}_M[k]^H]^H$ with $\mathbf{y}_i[k] = [y_i[k+r-1] \cdots y_i[k]]^T$ and $\mathcal{H} = [\mathcal{H}_1^H \cdots \mathcal{H}_M^H]^H$ with

$$\mathcal{H}_i = \begin{bmatrix} h_i[0] & \cdots h_i[L] & & \\ & \ddots & & \ddots \\ & & h_i[0] & \cdots \quad h_i[L] \end{bmatrix}_{r \times (r+L)}$$

so that $\mathbf{y}[k] = \mathcal{H}\mathbf{x}[k]$ with $\mathbf{x}[k] = [x[k+r-1] \cdots x[k-L]]^T$. Define $\mathbf{h} = [\mathbf{h}_1^H \cdots \mathbf{h}_M^H]^H$ with $\mathbf{h}_i = [h_i[0] \cdots h_i[L]]^T$. Then the algorithm can be summarized as follows:

1. Compute the output covariance matrix R_y: $R_y = \frac{1}{N-r} \sum_{k=1}^{N-r} \mathbf{y}[k]\mathbf{y}[k]^H$.
2. Compute the eigenvalue decomposition of R_y and determine the noise subspace:
$$U = [U_1 \cdots U_{(rM-L-r)}]$$
3. Exploit the orthogonality of U and $\mathcal{H}$ (i.e. $U^H\mathcal{H} \approx 0$) by minimizing $\mathbf{h}^H Q \mathbf{h}$ under the constraint $\|\mathbf{h}\| = 1$. The matrix Q is defined by $Q = \sum_{i=1}^{rM-r-L} \mathcal{U}_i \mathcal{U}_i^H$ with the $M(L+1) \times (L+r)$ matrix $\mathcal{U}_i$ constructed using the M $(r \times 1)$ segments $U_i^{(l)}$ of the noise vector U_i:
$$\begin{aligned} U_i &= [U_i^{(1)H} \cdots U_i^{(M)H}]^H \\ \mathcal{U}_i &= [\mathcal{T}_{L+1}(U_i^{(1)})^H \cdots \mathcal{T}_{L+1}(U_i^{(M)})^H]^H \end{aligned}$$

We replaced the first two steps in the algorithm by an estimate of the left singular vectors forming the nullspace of Y^H,

$$Y = [\mathbf{y}[1] \cdots \mathbf{y}[N-r]]$$

(because of better numerical properties), and added a fourth step in which the estimated channel impulse response was fed into a Viterbi trellis to determine the original symbol sequence.

Fast Start-up of Linear Constrained Bussgang Blind Equalizers

S.Zazo, J.M. Páez-Borrallo, I.A. Pérez-Álvarez**
Universidad Alfonso X El Sabio, Madrid (Spain)
** ETS Ingenieros de Telecomunicación, Universidad Politécnica de Madrid*
Phone: 341-8109153, FAX: 341-8109101, e-mail: szazo@uax.es

Abstract
Global convexity and fast convergence are both the keypoints of blind, baud rate, equalization techniques. In some previous works we proposed blind schemes based on linear constrained Bussgang equalizers. Easy implementation and globally convex characteristics, under certain hypotheses, have been their main advantages over standard algorithms. However, our proposal shares the slow convergence behaviour with existing equalizers and therefore, our interest is now pointed to improve the convergence speed. Our main goal is the development of a blind updating technique based on more efficient stochastic gradient tools: conjugate gradient techniques as a trade-off between complexity (matrix inversions are not required) and convergence speed.

1. Introduction

Blind equalization techniques intends the retrieval of the input data given only the channel output and some statistical or deterministic information of the channel input. At this moment, there exist two main approaches to blind, baud rate, equalization schemes: on one hand, Bussgang algorithms are based on the estimation of the transmitted symbol by a zero-memory non linearity; simple implementation is achieved but also non convex cost error functions are involved due to the nonlinearity. On the other hand, higher order statistics based algorithms intend the equalization of the transmitted sequence through the identification of the minimum and maximum phase components of non minimum phase channels; it is shown the global convexity of the cost functions but the computation requirements make difficult the practical implementation in many applications; also an important delay is usually needed to properly estimate the cumulants involved [Hay94]. In some previous works [Zaz94, Zaz95] we developed a new scheme for blind equalization labelled *Modified Decision Directed Algorithm* (MDDA) which improves the performance of the standard algorithm providing convex functions under certain conditions. However, the scheme although faster than the CMA or Stop-Go ones, is not comparable in convergence speed to non blind strategies by using an instantaneous gradient updating.

As an alternative to the instantaneous gradient schemes, it is well known that *Conjugate Directions* techniques have been developed for the quadratic problem leading to significant practical advances. Also, an extension for non quadratic problems is possible by introducing line search methods [Lue89]. Additionally, [Bor92] propose the application of conjugate gradient techniques for adaptive filtering as a trade-off between computational complexity and convergence performance: the method proposed is capable of providing convergence comparable to RLS schemes with a computational complexity that is intermediate between the LMS and the RLS methods.

Our main goal in this paper is to incorporate conjugate gradient techniques to the MDDA in order to achieve global convexity with convergence similar to gradient search non blind schemes. In this analysis (CG-MDDA) we have considered binary transmission, first through minimum phase channel to get sensitivity to the problem, and finally we propose a scheme to rapidly start-up a non minimum phase channel. Some simulations over a raised cosine impulse response with controlled amplitude distortion corroborate this desirable performance.

2. Our proposal. Minimum phase channel

Let us start showing a simple transmission scheme assumed perfectly synchronised at baud rate (fig.1). The channel response $h[n]$ includes the transmitter filter, the channel and the receiver filter. First at all we will consider minimum phase channel, and later in other section we will extend our strategy to a generic non minimum phase channel. Recall that he MDDA is based on the anchoring of one equalizer tap (the first one for minimum phase channel and the center tap for non minimum phase channel), in order to preserve the current symbol, avoiding the presence of local minima. Also, another degree of freedom is introduced as an adaptive decision device for gain (and optionally phase) recovery.

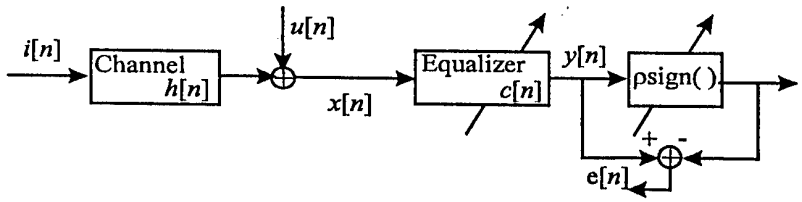


Fig.1. Binary transmission block diagram

The cost error function is very similar to the Decision Directed (DD) scheme, updating the parameter set in the direction of the minus instantaneous gradient:

$$J = E\left\{\left[Y - \rho\text{sign}(Y)\right]^2\right\} \tag{1}$$

In a similar way, we have considered another gradient strategy to update the tap-weight vector: a conjugate gradient technique. In a general case, the whole adaptive parameter set $(\mathbf{c},\rho)$ in the cost function should be considered into the conjugate gradient scheme. However, in this case, the behaviour of the equalizer taps is very different to the gain parameter: quadratic (at least for minimum phase channels) in the coefficient set and not quadratic (but convex) in the gain parameter. Therefore, we propose a combined adaptation rule: instantaneous gradient for the gain parameter and conjugate gradient for the equalizer coefficients (of course including the linear constraint):

$$\begin{aligned} \rho_{k+1} &= \rho_k - \mu\nabla_\rho(J) \\ \mathbf{c}_{k+1} &= \mathbf{c}_k + \alpha_k\mathbf{d}_k \end{aligned} \tag{2}$$

Observe the important differences in both updating rules: in the case of instantaneous gradient, the step size is fixed and the innovation is proportional to the instantaneous gradient. In the other hand, for the conjugate gradient technique, the step size α_k is chosen under a certain optimal criteria; So on, the innovation is proportional to a conjugate direction $\mathbf{d}_k$ [Lue89], both to be determined. The key point of the algorithm should be to design an optimal condition for the line search method minimising the cost error function along the line:

$$\min_{\alpha_k}\left(J(\mathbf{c}_k + \alpha_k\mathbf{d}_k)\right) \tag{3}$$

It can be shown that, for minimum phase channels and assuming a conditioned gaussian model for the ISI [Zaz95 and references therein], that condition (3) is equivalent to minimise the cross correlation between the equalizer output and part of the input vector (the whole vector except the first element in the minimum phase case). This choice seems to be appropriate because we assume that the current data is not present in this part of the input vector, and so on we are minimising the ISI. The expression we have found solving the line search equation given by (3) is:

$$\alpha_k = -\frac{\mathbf{c}_k^T\mathbf{P}'\mathbf{R}\mathbf{P}\mathbf{d}_k}{\mathbf{d}_k^T\mathbf{P}'\mathbf{R}\mathbf{P}\mathbf{d}_k} \tag{4}$$

where superscript ^{T} means transpose, $\mathbf{R}$ is the input correlation matrix (should be estimated recursively), and $\mathbf{P}$ is the projection matrix on the linear constraint.

The problem is then to compute the appropriate set of direction vectors. It is well known that the conjugate gradient method is the conjugate direction method that is obtaining by selecting the successive direction vectors as a conjugate version of the successive gradients obtained as the method progresses:

$$\mathbf{d}_{k+1} = -\mathbf{g}_k + \beta_k\mathbf{d}_k \tag{5}$$

where $\mathbf{g}_k$ means the successive gradients, and β_k are the constants chosen to provide the $\mathbf{R}$-conjugacy for the vector $\mathbf{d}_k$ with respect to the previous direction vectors. To calculate the iterative conjugate directions we have applied an extension to nonquadratic problems known as the Fletcher-Reeves method [Lue89]:

$$\beta_k = \frac{\mathbf{g}_{k+1}^T\mathbf{g}_{k+1}}{\mathbf{g}_k^T\mathbf{g}_k} \tag{6}$$

2. Simulations. Minimum phase channel

A very simple but also intuitive result is shown in this section: let us consider a one pole channel (z=-0.5) and also a two taps equalizer: we have represented the error surface to observe the different trajectories followed for the instantaneous gradient and the conjugate gradient algorithm (See Fig.2): a very interesting topic is the different convergence speed: about 10 samples for the CG-MDDA and almost two hundred for the MDDA (See Fig.3); under this performance, we must point out that the cost function is still convex but the convergence speed could be now comparable to a trained stochastic gradient schemes.

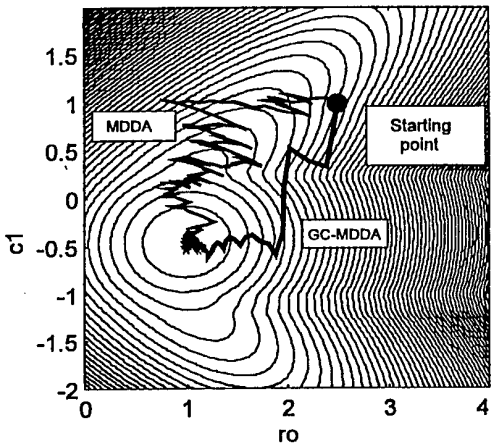


Fig.2. Error surface and trajectories of the instantaneous gradient and conjugate gradient algorithms.

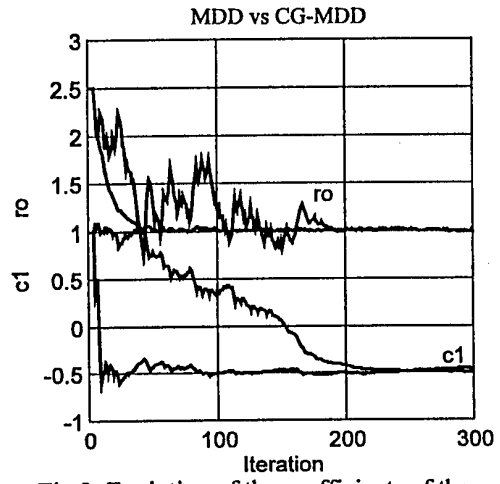


Fig.3. Evolution of the coefficients of the instantaneous gradient and conjugate gradient algorithms.

3. Non minimum phase channel

A minimum phase system could be a good model for most of the radio multipath propagation channels, where a strong direct ray is received among a few reflected rays with varying delays, amplitudes and phases. However, in other applications like data transmission over telephone lines, channels involved have a non minimum phase characteristic.

Although the extension to this new situation could be intuitive: fixing the center tap as the main responsible of the current data, and also leaving parameter ρ for gain adjustment, the analysis of this situation arise in complexity. A closed expression for the optimal stepsize α_k as equation (4) could not be found because the current symbol is present in the post and also previous symbols. Therefore, we have followed a similar development as we developed in [Zaz95] to justify an equivalent condition as one given by (4): we guess that the main part of the current symbol is provided by fixing the central tap; in spite of the fact that this symbol is also present in the remainder input vector, we have observed in several real channels that the linear constraint will preserve the current data avoiding from any possible misconvergence (the cancellation of this symbol will lead to a local minima). Therefore, we have a very simple key for fast start-up a non minimum phase channel implementing the same expression as one given by equation (4) where matrix **P** in this case is the projection matrix on the linear constraint fixing the center tap. Of course we realise that this strategy is not able to assure the right convergence for any channel, but in the next section we verify by simulations that it could be a trade-off criterium for not very high distortion channels

5. Simulations. Non minimum phase channel.

One of the most critical points of blind equalization is that the convergence speed is not comparable to non blind algorithms. Of course, RLS algorithms will be always much faster, but also it is known that the computational requirements (matrix inversions) could be a serious disadvantage in many applications. In this section we want to evaluate the evolution of the CG-MDDA to compare its performance with trained LMS algorithm (the most popular non blind adaptation rule).

The channel we have chosen is given in equation (6) and is taken of [Bor92-Hay91] as a raised cosine impulse response where W controls the amplitude distortion in the channel (eigenvalue spread). We have consider two different situations with a low and high eigenspread to show that in any case the performance is competitive, also showing an interesting property of conjugate gradient techniques: the convergence speed is almost independence of the eigenspread in contrast with LMS achieving a great dependence with this issue.

$$h(i)=\begin{cases}0.5\left[1+\cos\left(2\pi\left(i-2\right)/W\right)\right] & i=1,2,3\\ 0 & \text{otherwise}\end{cases} \qquad (7)$$

In Fig.4: we have considered parameter W=2.9 which implies a low eigenvalue spread of 6.07. Also in Fig.5 we have chosen W=3.5 as the worst

eigenvalue spread of 46.8. In both cases we have consider the same conditions choosing randomly the starting point to justify the convex characteristics of the GC-MDDA error surface. The equalizer length is set as 11 taps (the center tap is kept fixed) and the signal to noise ratio is of 30 dB.

The implementation of a conjugate gradient adaptive filtering is not so simple as in a purely quadratic problem: we have followed the recommendation of [Bor92] using an averaging window to estimate the gradient; in this simulations we have chosen an average of three (An interesting discussion about this issue is adressed in [Bor92] and we have not observed any significant improvement for wider windows).

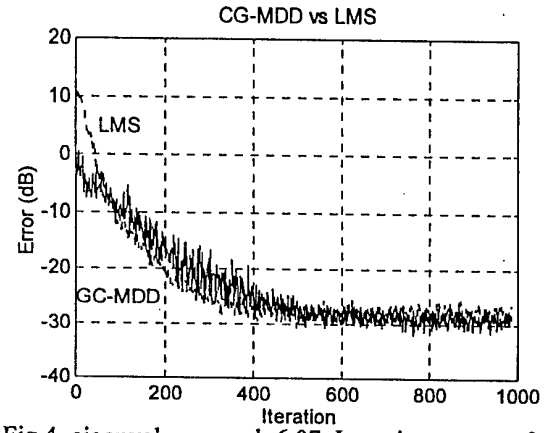


Fig.4. eigenvalue spread: 6.07. Learning curves of the CG-MDDA and LMS.

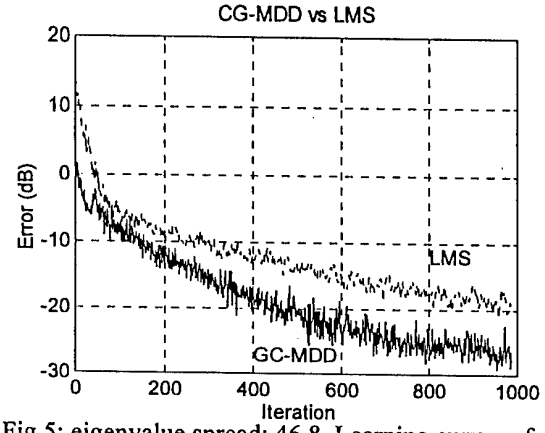


Fig.5: eigenvalue spread: 46.8. Learning curves of the CG-MDDA and LMS.

Finally, to point out that our proposal could be implemented more efficiently in two steps using the CG-MDDA just for fast opening the eye-diagram and a second step (after about 50 samples) driven by the MDDA to reduce the complexity: in this way the computation and storage requirements are comparable to the LMS.

5. Conclusions and lines of further research.

The main conclusion we want to point out is the fact that we have speed-up a globally convex blind equalization to a convergence speed similar to trained strategies, also independently of the eigenvalue spread. We have introduced a different updating rule, considering an instantaneous gradient technique for the adaptation of the gain parameter and a conjugate gradient technique to update the equalizer taps. Although the calculation of the optimal step size is only available for minimum phase channel, an intuitive approach has been developed for non minimum phase channels. Simulations over controlled amplitude distortion channels support the scheme we have proposed.

Also recall that the channel (3) could be considered as a good model for many situations in radio mobile applications where a fast and desirable blind equalization should be required.

The use of non instantaneous gradient techniques to speed up blind algorithms, as conjugate gradient techniques, could be applied to other Bussgang equalizers. This could be a line of further research where much work must be done in the theoretical approach to design the optimum step size for the line search method providing the minimization of the desired cost function.

6. References

[Bor92]: Boray, G.K. and Srinath, M.D.. Conjugate Gradient Techniques for Adaptive Filtering. *IEEE Trans. on Circuits and Systems I.* Vol.39, No1,Jan. 1992.

[Hay91]: Haykin, S. *Adaptive filter Theory.* Prentice Hall 1991.

[Hay94]: Haykin, S. *Blind deconvolution.* Prentice Hall 1994.

[Lue89]: Luenberger, D.G. Linear and non Linear Programming. Addison Wesley 1989.

[Zaz94]: Zazo, S., Páez Borrallo, J.M., Pérez Álvarez, I.A.. Analysis of a Globally Convex Algorithm for Blind Equalization of Non Minimum Phase Channels. *Proc. EUSIPCO 94*, Edinburg, U.K

[Zaz95]: Zazo, S., Páez Borrallo, J.M., Pérez Álvarez, I.A. A linear constrained blind equalization scheme based on Bussgang type algorithms. *Proc. ICASSP 95*, Detroit, USA.

Emergent Techniques

A GENETIC ALGORITHM FOR SYNTHESIZING LIP MOVEMENTS FROM SPEECH

Darnell Moore, Antai Peng, and Monson H. Hayes

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250 USA
Phone: +1 404-894-2958 Fax: +1 404-894-8363
email: djmoore@eedsp.gatech.edu

ABSTRACT

This paper introduces a means of generating synthetic facial image sequences from speech. Using the Facial Action Coding System (FACS), facial expressions related to speech are correlated with phonemes. The Genetic Algorithm is introduced as a means of generating the parameters for manipulating a neutral face image to match the intended target image. Examples generated by a prototype system are also included.

1. INTRODUCTION

Somewhere in our collective imagination, we have all envisioned the day when computer systems can communicate and interact with humans in a natural, instinctive fashion. Motivated by this possibility, there has been tremendous interest in developing signal processing systems that are capable of understanding and emulating human expressions, gestures, and speech. While text-to-speech synthesizers have made great advances since their introduction over 20 years ago, realistic facial animation systems are still at embryonic stages of development [1]. The most common communication modality between humans is speech; likewise, we are interested in systems that can not only generate and recognize realistic human speech but are also capable of synthesizing the corresponding facial animation.

Although the expressive bandwidth of the entire human face is enormous, the region containing the mouth and lips are, arguably, the most involved in communication. We have elected to concentrate on a system that emulates lip motion because of its obvious and natural complement to speech synthesis. There are many different problems associated with the development of a system for generating synthetic lip movements from speech. These include:

- Extracting phonemes and timing information from speech or text.
- Developing an anatomic and muscular model that may be used to generate lip movements from a command string.
- Developing a distortion measure that may be used to measure the performance of a sequence of lip movements.
- Search for the optimum sequence of model parameters that will generate the most natural- looking lip and facial motion.

In this paper, we will focus, primarily, on the latter problem: search for the optimum model parameters.

2. MODELING OF FACIAL EXPRESSION

Reducing the, seemingly infinite, expression space of the human face to a finite parameter-based model is a difficult task. To quantify and represent this infinite space, we are employing the facial expression representation system based on the Facial Action Coding System (FACS), which was developed by two psychologists, P. Ekman and W.V. Friesen in 1977 [6]. FACS is widely used in contemporary facial animation engines and offers comprehensive parameterization of facial expressions. This approach uses *action units* (AUs) to describe the changes in the appearance of the face caused by contraction of an isolated or a contiguous group of facial muscles. FACS furnishes 46 AUs for describing facial expressions and 12 AUs for gaze directions and head movements. However, we are primarily concerned with the four groups of lower face AUs germane to speech. These four groups, classified by their translation or rotation along the primary axes, are denoted as the *vertical, horizontal, orbital,* and *oblique*

Much of this work was conducted for A. Peng's dissertation research.

Action Units	*Description*
Vertical	
*AU*10	Raise Upper Lip
*AU*15	Depress Lip Corner
*AU*25	Part Lip
*AU*26	Drop Jaw
*AU*27	Open Mouth Wide
Horizontal	
*AU*20	Stretch Lip Horizontal
Oblique	
*AU*18	Pucker Lip
*AU*22	Funnel Lip
Orbital	
*AU*12	Stretch Lip Corner
Miscellaneous	
*AU*29	Thrust Jaw
*AU*32	Bite

Table 1: The 11 Speech Related Action Units

AUs. We also use two action units whose unique displacement escaped classification and are referred to as *miscellaneous*. Table 1 contains the 11 AUs that are involved with speech production.

2.1. WIRE-FRAME HEAD MODEL

In order to generate each facial expression, a neutral face image will have to be modified. The person's image is first texture mapped on to a 3D wire-frame head model. To correct for facial asymmetries, each model has been tailored to fit the anatomical structure of the person's head. As AUs manipulate the location of specific nodes on the wire-frame model, the image wrapped around the model is, consequently, distorted. This procedure allows us to render facial expressions.

Because AUs represent displacement vectors at discrete points, e.g., $AU = (dx\ dy\ dz)^T$, they are additive and non-orthogonal. For example, several AUs can be combined to form new action units. By exploiting this property, we can use weighted, linear combinations of these 11 AUs as a basis set for synthesizing all speech related motions:

$$\mathrm{B} = [\mathrm{AU10}, \mathrm{AU12}, \mathrm{AU15}, \ldots, \mathrm{AU32}].$$

For convenience, we let vector $\mathbf{v}_1$ represent AU10, $\mathbf{v}_2$ represent AU12, and so on until we have 11 vector components to represent our *AU-space* such that

$$\mathbf{V} = \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,11} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,11} \\ v_{3,1} & v_{3,2} & \cdots & v_{3,11} \end{pmatrix}.$$

2.2. PHONEMES

Phonemes are the distinctive, discrete sounds that compose speech. Construction of each phoneme is often correlated with a particular contortion of a facial muscle group, making FACS an ideal tool for quantifying expressions. The goal of our system is to map speech phonemes to facial expressions by constructing the expressions from a weighted, linear combination of AUs. Thus, for the synthesis of lip movements, we must identify the optimum eleven-dimensional AU weight vector $\mathbf{w}$ for each frame in an image sequence. An expression $\mathbf{f}$ can be constructed as

$$\mathbf{f} = w_1\mathbf{v}_1 + w_2\mathbf{v}_2 + \ldots + w_{11}\mathbf{v}_{11} = \mathbf{V}\mathbf{w}, \qquad (1)$$

where w_i, $i = 1, 2, 3, \ldots, 11$, are the corresponding expression weight vectors. For example, to model the upper lip raise, $w_1 = 1$ and $w_i = 0$, $i = 2, 3, \ldots, 11$.

3. ITERATIVE FACIAL EXPRESSIONS

In order to determine the expression weight vectors for a particular phoneme, we must train our system with a target image representing this expression. Given a target image $\mathbf{T}$ with an expression that we want to synthesize, $\mathbf{w}$ represents the parameterization of the target expression, e.g., $f_{Parm} : \mathbf{T} \longrightarrow \mathbf{w}$. Let $\mathcal{D}(\mathbf{w}, \tilde{\mathbf{w}})$ be a function that measures the distortion between two parameterized expressions $\mathbf{w}$ and $\tilde{\mathbf{w}}$. A synthetic facial image $\mathbf{S}$ wearing an expression parameterized by $\tilde{\mathbf{w}}$ is generated from an initial, neutral expression image $\mathbf{N}$. e.g., $f_{Synth} : (\mathbf{N}, \tilde{\mathbf{w}}) \longrightarrow \mathbf{S}$. Our goal is to find $\tilde{\mathbf{w}}$ such that

$$min(\mathcal{D}(\mathbf{w}, \tilde{\mathbf{w}})). \qquad (2)$$

If $\mathbf{N}$ contains the same person as $\mathbf{T}$, then

$$\|\mathbf{T} - \mathbf{S}\| \longrightarrow 0. \qquad (3)$$

To solve this optimization problem, we use the genetic algorithm to automatically find $\tilde{\mathbf{w}}$.

3.1. GENETIC ALGORITHM

Genetic Algorithms (GAs) are machine learning techniques based on the principles of genetic variation and natural selection [4]. Developed by Holland (1975) as an attempt to model various natural phenomena, GAs are widely used to model evolution in artificial-life systems. The problem of finding an optimum weight vector $\mathbf{w}$ for a given phonemic facial expression presents a challenge because the 11 weight components of $\mathbf{w}$ must be found simultaneously. We are attracted to the GA's simplistic and elegant nature as well as to

its power to rapidly arrive at good solutions to complex high-dimensional problems [5]. Moreover, GAs can provide an exceptionally powerful search heuristic for large, complex spaces if the space to be searched is not well understood, is relatively unstructured, and can be effectively represented by a GA [5].

The GA method works by evolving one population of chromosomes to a new population, or generation, using selection combined with various operators. In our case, bit strings representing candidate weight vectors play the role of *chromosomes* with individual weights portraying the *genes*. The selection process is driven by a fitness function that evaluates the chromosomes in a population, identifies those that will be allowed to *reproduce*, and decides the number of *offspring* each is likely to have. As in nature, the fittest chromosomes produce more offspring than less fit ones. The two categories of genetic operators that manipulate chromosomes are known as crossover and mutation operators. Crossover operators trade genes between chromosome pairs whereas mutation operators alter genes within a single chromosome.

3.1.1. GENETIC OPERATORS

There are four types of crossover operators utilized in our system: *simple, arithmetic, single,* and *heuristic.* See Table 2 for their definitions. Consider the chromosome pair $\mathbf{w}_1 = (w_{1,1}\ w_{1,2} \ldots w_{1,s} \ldots w_{1,11})^T$ and $\mathbf{w}_2 = (w_{2,1}\ w_{2,2} \ldots w_{2,s} \ldots w_{2,11})^T$, where an integer s in the range [1,11] is randomly selected as the mutation site. After crossover operations, this pair becomes $\tilde{\mathbf{w}}_1$ and $\tilde{\mathbf{w}}_2$, respectively. The heuristic crossover operator

Operator	Chromosomes after the Operator
Simple	$\tilde{\mathbf{w}}_1 = (w_{1,1}\ w_{1,2} \ldots w_{2,s} \ldots w_{2,11})^T$ $\tilde{\mathbf{w}}_2 = (w_{2,1}\ w_{2,2} \ldots w_{1,s} \ldots w_{1,11})^T$
Arithmetic	$\tilde{\mathbf{w}}_1 = (w_{1,1}\ w_{1,2} \ldots \alpha w_{2,s} + (1-\alpha) w_{1,s}$ $\ldots \alpha w_{2,11} + (1-\alpha) w_{1,11})^T$ $\tilde{\mathbf{w}}_2 = (w_{2,1}\ w_{2,2} \ldots \alpha w_{1,s} + (1-\alpha) w_{2,s}$ $\ldots \alpha w_{1,11} + (1-\alpha) w_{2,11})^T$ $0 < \alpha < 1$
Single	$\tilde{\mathbf{w}}_1 = (w_{1,1}\ w_{1,2} \ldots w_{2,s}\ w_{1,s+1} \ldots w_{1,11})^T$ $\tilde{\mathbf{w}}_2 = (w_{2,1}\ w_{2,2} \ldots w_{1,s}\ w_{2,s+1} \ldots w_{2,11})^T$

Table 2: Genetic Crossover Operators, *excluding the Heuristic Crossover*

uses domain specific knowledge to encourage the replication of chromosomes that carry certain gene combinations. More specifically, this operator selectively adjusts certain weight components so that candidate weight vectors reflect the expression. For example, if the face in the target image has a mouth opened widely, weight components applied to AU26 or AU27 will dominate. Heuristic operators are responsible for providing improved performance by considerably reducing the parameter search space.

To perturb individual chromosomes, uniform and non-uniform mutation operators are implemented. Given a chromosome $\mathbf{w} = (w_1\ w_2 \ldots w_s \ldots w_{11})^T$, an integer s in the range [1,11] is selected as the site where mutation will occur. During uniform mutation, the gene at site s, w_s, is replaced with a random number β which is uniformly distributed on the interval [0,1], e.g., $w_s \longrightarrow \beta$. Otherwise for non-uniform mutation, $w_s \longrightarrow w_s + \delta$, where δ is a random number non-uniformly distributed on [0,1]. A special mutation operator has been developed, referred to as the complement mutation, which replaces w_s with its inverse, e.g., $w_s = 1.0 - w_s$.

3.1.2. GENETIC EVOLUTION CYCLE

The initial population has an expression weight vector of all zeros, which corresponds to a neutral facial expression, i.e. mouth closed. The evolution cycle is guided by system parameters that influence selection dynamics. The single, arithmetic, and simple crossover operators each have an associated probability of occurrence, denoted *Psnc*, *Pac*, and *Psic*, respectively. The uniform and non-uniform mutation operators also have probabilities, defined as *Pum* and *Pnm*, respectively. Uniform mutation is primarily used during early iterative stages in order to evaluate a wide range of candidate chromosomes; hence, $Pum \gg Pnm$. Each time a better solution is found, *Pum* is decreased while *Pnm* is inflated. Towards to end of the evolutionary cycle, non-uniform mutation becomes more prevalent so that gene selection can be finely tuned; likewise, $Pnm \gg Pum$. However, if no improvements are made after a certain number of iterations, the complement mutation is applied to stimulate the gene pool with candidates that may have been overlooked. All mutation probabilities are then reset to their initial values.

Both mutation and crossover operators may generate expression weight vectors that do not correspond to any meaningful phonemic expressions. For instance, weights associated with AU27 *(mouth open)* and AU32 *(lip bite)* should not both be dominant since biting one's lip is fairly impossible given an open mouth. Heuristic constraints have been added to eliminate such weight vectors, further shrinking the solution space and increasing algorithm robustness.

After the population of chromosomes has been altered by genetic operators, it is filtered by a fitness function designed to rank its members. Fitness is measured by evaluating the distortion between the synthesized and the target image. We have identified five

control points surrounding the lips: *center of the outer upper lip, center of the inner upper lip, center of the inner bottom lip, center of the outer bottom lip,* and the *corner of the lip.* The distances of all of these points are taken with respect to the center of the eye. By separating the horizontal from the vertical components, we can make quantitative measurements to determine the distance between points; hence, describing the lip's shape.

Since the target image can contain a different person than the synthesized image, distances must be normalized by an expression ratio defined as

$$R = \frac{\textit{distance of a control point on neutral face}}{\textit{distance of same point on target}}.$$

The distortion $\mathcal{D}$ is computed as a weighted sum of the difference between the corresponding ratio components of the target and synthesized expressions. It is given by

$$\mathcal{D} = \sum_{i=1}^{5} w_i (R_{(s,h)i} - R_{(t,h)i} + R_{(s,v)i} - R_{(t,v)i}), \quad (4)$$

where $i = 1 \ldots 5$, $R_{(s,h)i}$ and $R_{(s,v)i}$ are the horizontal and vertical expression ratios, respectively, and $R_{(t,h)i}$ and $R_{(t,v)i}$ are the horizontal and vertical expression ratios, respectively; w_i is a real on [0,1].

3.1.3. GENETIC SELECTION

Generation selection is implemented by a *roulette wheel* method with non-duplicate reproduction and *elitism.* In this approach, fitness scores of the population are first summed, e.g.,

$$Q = \sum_{i=1}^{N} fitness(w_i), \quad (5)$$

where N represents the no. of candidate chromosomes in the generation and $fitness(\cdot)$ denotes the fitness function described in section 3.1.2. A random integer p on the range $[0, Q]$ is generated. Each population member *(candidate weight vector)* is examined sequentially to determine if its fitness score and those of preceding members are collectively equal or greater than p. The first chromosome satisfying this condition is chosen as a member of the new population. Duplicate members are removed and the most fit, or elite, member of the iterative population is always retained in the new generation. This entire process is repeated multiple times until all members of the target population are found. Generally, the fitness level of the population is proportional to the number of evolutionary iterations.

3.1.4. FRAME INTERPOLATION

While language can be disected into phonemic components, stacking phonemes sequentially will not produce natural-sounding speech. Likewise, synthetic images representing phonemes only supply key frame in a natural animation sequence. Moreover, phonemic structures do not provide any timing information, so the natural expressive quality is often lost. We used two interpolation methods between phonemic key frames to simulate fluent animation: *position* and *Expression Weight Vector* interpolation [2]. With position interpolation, vertices on the 3D wire-frame model are adjusted according to the *displacement* between key frames $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$, e.g.,

$$\mathbf{x}_{n,k} = \alpha \mathbf{x}_{i,k} + (1-\alpha)\mathbf{x}_{i+1,k} + \alpha(\mathbf{x}_{i,k} - \mathbf{x}_{i+1,k}), \quad (6)$$

where $0 < \alpha < 1$, $k = 1, 2, \ldots, N$, $n = 1, 2, \ldots, M$, and $\mathbf{x}_{n,k} = (x_{n,k}\ y_{n,k}\ z_{n,k})^T$. N is the no. of vertices and M is the no. of intermediate frames to be generated between $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$. Vector interpolation builds each new expression vector $\mathbf{w}_n$ between $\mathbf{w}_i$ and $\mathbf{w}_{i+1}$ using linear interpolation, e.g.,

$$\mathbf{w}_n = \alpha \mathbf{w}_i + (1-\alpha)\mathbf{w}_{i+1}, \quad (7)$$

where n and α are unchanged from eq. 6.

4. EXPERIMENTS ON PROTOTYPE SYSTEM

To evaluate the system described above, we have implemented a prototype system developed by A. Peng on a Windows NT platform using the OpenGL graphics library and C [3]. Neutral and target face images are displayed, permitting the user to interactively select the lip control locations using the mouse. After specifying the maximum number of iterations in the modeling process, expression weight vectors are passed as parameters to the synthesizer. The synthesizer updates the facial image on the screen in real-time. We used texture-mapped images on a wire-frame anatomical head model combined with synthetic speech to enhance realism.

4.1. RESULTS & DISCUSSION

Using target image training, a parameter database that maps phonemes to facial expressions was created. Files containing speech broken into phonemes can be fed to the system which will fabricate the corresponding animation. Figures 1- 3 shows several frames from an animation sequence of the word "teeth" [3].

Figure 1: First 2 key frames from the word "teeth".

Figure 2: Ending 2 key frames from the word "teeth".

Unlike other optimization problems where the best solution is required, sub-optimal solutions often provide very decent solutions. Although distortion decreases with addition iterations, this error may be less noticeable to human eyes, particularly if implemented in real-time. From our preliminary experiments, 20 - 60 iterations generally produce synthesized facial images that looked strikingly similar to the target image.

We also considered adding more control points to improve lip shape resolution. While adding more control points permit more accurate distinction between images, the complexity of the system increases. There was a consensus to keep five points so real-time performance would not be handicapped. Other variations of population selection, such as steady-state reproduction, were also evaluated but did not yield any significant improvements.

5. REFERENCES

[1] D. H. Klatt, "Review of Text-To-Speech Conversion for English," *J. Acoustic Soc. of Am.* 82(3), pp. 737-793, Sept. 1987.

Figure 3: Middle 2 key frames from the word "teeth".

[2] A. Peng and M. Hayes, "Iterative Human Facial Expression Modeling". *ICASSP '95 Proceedings*– Detroit,vol. 4, pp. 2627-2631.

[3] A. Peng, "Speech Expression Modeling and Synthesis". *PhD Dissertation.* School of Electrical and Computer Engineering, Georgia Institute of Technology. Atlanta, Georgia USA. May 1996.

[4] M. Mitchell and S. Forrest, "Genetic Algorithms and Artificial Life". *Artificial Life.* Santa Fe Institute. Santa Fe, New Mexico.

[5] S. Forrest and M. Mitchell, "What Makes a Problem Hard for a Genetic Algorithm? Some Anomalous Results and Their Explanation". Department of Computer Science, University of New Mexico. Albuquerque, New Mexico .

[6] P. Ekman and W.V. Friesen, "Facial Action Coding System," *Consulting Psychologist.* 1977.

EVOLUTIVE STRATEGIES FOR ADAPTIVE COLOR QUANTIZATION

A. I. Gonzalez, M. Graña, A. D´Anjou, P.Larrañaga, J.A. Lozano, F.X. Albizuri

Dept. CCIA Univ. Pais Vasco/EHU[1],Aptdo 649, 20080 San Sebastián, España
e-mail: ccpgrrom@si.ehu.es

ABSTRACT

Color quantization of still images can be easily stated as a clustering problem. Color quantization of sequences of images becomes a non-stationary clustering problem. In this paper we propose a very simple and effective evolutive strategy to perform adaptively the computation of the color representatives for each image in the sequence. Salient features of the evolutive strategy proposed here are: individuals correspond to individual cluster centres, to approach real-time response we impose one-generation adaptation for each image, only mutation operators are applied and these mutation operators are guided by the actual covariance matrices of the clusters. Experimental results on a sequence of indoor images are presented.

1 INTRODUCTION

Color Quantization [1, 2, 3, 4] is an instance of the more general technique of Vector Quantization [5] in the space of colors. Although it is well known [6] that the euclidean distance in the RGB space does not preserve the perceptive distance between colors, most of the approaches applied in practice and reported in the literature work in the RGB space using the euclidean distance as the clustering similarity metric. Some works [4] show that the tradeoff between computational efficiency and visual quality justify this decision. In the works reported in this paper we will stick to this common practice. Color Quantization has application in visualization, color image segmentation, data compression and image retrieval [7]. In visualization and compression applications the typical size of the color palette (codebook, color representatives) is 256, whereas for segmentation and retrieval tasks the size of the color palette is smaller. If the number of color representatives is not set beforehand, the problem of discovering the "natural" number of representatives becomes much more involved, and it is a line of research of its own. In [8] we reported the application of steady state genetic algorithms to this kind of problems.

In this paper, the emphasis is in performing near real-time adaptive clustering for a non-stationary population. Therefore, we will consider the palette size fixed for each experiment. Although sequences of images (video) lead naturally to the consideration of time varying clustering problems, the usual approaches consider time invariant distributions of either colors [9] or image blocks [10], and apply conventional clustering methods. This may be due to the nature of the video sequences considered. Most of the works are applied to video recording of talking heads, which show little variation of color distribution. Our experimental work involves the color quantization of a sequence of images that show a smooth but clear variation over time of the distribution of colors. Our experimental image sequence is a subsample of a panning sequence of an indoor scene (looking around the laboratory). Some heuristical efforts [11,12] have been reported that try to cope with the time varying characteristics inherent to image sequences. Our approach is to state the problem as a time-varying clustering problem, and to propose an evolutive strategy as the adaptation mechanism.

Evolutive strategies [13, 14, 15] have been developed mainly by Schwefel since the sixties. They belong to the broad class of algorithms inspired by natural selection: genetic algorithms, genetic programming, etc. Under this design philosophy (population based, fitness guided, genetic-like operators) a vast host of algorithms have been proposed and applied. The features most widely accepted as characteristic of evolutionary strategies are: (1) vector real valued individuals, (2) the main genetic operator is mutation, (3) individuals contain local information for mutation so that adaptive strategies can be formulated to self-regulate the mutation operator. However, it is widely recognized [15] that a lot of hybrid algorithms can be defined, so that it is generally difficult to assign a definitive "label" for a particular algorithm. Nevertheless, we classify the algorithm employed here as an evolutive statrategy because it fits in the above characterization. Our work differs from previous attempts [16] to apply evolutive strategies to clustering problems in several technical points (definition of individuals, fitness, selection) and in a mayor philosophical point: we are looking for an adaptive strategy that can be applied in near real time, whereas, for most of the evolutive/genetic

[1]This work is being supported by a research grant from the Dpto de Economía of the Excma. Diputación de Guipuzcoa, and a predoctoral grant and project PI94-78 of the Dept. Educación, Univ. e Inv. of the Gobierno Vasco

literature, the aim is to outperform other clustering algorithms without any regard for time constraints.

The paper is organized as follows. Section 2 introduces time varying clustering notation. Section 3 discusses the evolutive strategy used in the experimental work. Section 4 presents the experimental results, and section 5 gives our conclussions and lines for further work.

2 TIME VARYING CLUSTERING

Cluster analysis and the related Vector Quantization design problem are important techniques in many engineering and scientific disciplines [5,17,18,19,20,21]. Color Quantization is Vector Quantization in the RGB unit cube. In their most usual formulation it is assumed that the underlying stochastic process is stationary and that a given set of sample vectors properly characterizes this process. This paper tries to address the case when the underlying stochastic process is assumed to be time dependent (such as it happens in general image sequences), and propose an evolutive strategy that can be of use. Another important assumption is that no knowledge of a model for the time variation of the population is known. If a model is known, a predictive approach [5] would reduce the problem to a stationary one.

A time variant formulation of the clustering problem must start with the explicit assumption of a time varying population described by an stochastic process $\{\mathbf{X}_t,\ t=0,1,..\}$ (Note that we have jumped into the discrete time case). In this framework, a working definition of the time varying Clustering problem could read as follows: Given a sequence of sets of vectors $\aleph(t)=\{\mathbf{x}_1(t),..,\mathbf{x}_n(t)\}$ obtain a corresponding sequence of partitions of each of them into a sequence of sets of disjoint clusters $\{\aleph_1(t),..,\aleph_c(t)\}$ that minimizes a criterium function $C=\sum_{t\geq 0}C(t)$ The related time varying Vector Quantization can be stated as the search for a sequence of representatives $\mathbf{Y}(t)=\{\mathbf{y}_1(t),..,\mathbf{y}_c(t)\}$ that minimizes the error function (distortion) $E=\sum_{t\geq 0}E(t)$. Functions C and E coincide when the criterium function is the within cluster variance and the error function is based in the euclidean distance. Color Quantization of image sequences obviously is a time varying Vector Quantization problem. The stochastic minimization problem that must be considered in order to derive adaptive algorithms can be stated as follows:

$$\min_{\{\mathbf{Y}(t)\}}\sum_{t\geq 0}\sum_{j=1}^{n}\sum_{i=1}^{c}\left\|\mathbf{x}_j(t)-\mathbf{y}_i(t)\right\|^2\delta_{ij}(t)$$

$$\delta_{ij}(t)=\begin{cases}1 & \mathrm{i}=\mathrm{argmin}\left\{\left\|\mathbf{x}_j(t)-\mathbf{y}_k(t)\right\|^2 k=1,..,c\right\}\\ 0 & \text{otherwise}\end{cases}$$

A reasonable simplifying assumption is that the minimization of the sequence of time dependent error function can be done independently at each time step:

$$\min_{\{\mathbf{Y}(t)\}}\{E(t)\ t=0,1,..\}=\left\{\min_{\{\mathbf{Y}(0)\}}E(0),\min_{\{\mathbf{Y}(1)\}}E(1),..\right\}$$

Under another reasonable assumption, that of smooth (bounded) variation of optimal set of representatives at succesive time steps (i.e. $\left\|\mathbf{y}_i(t)-\mathbf{y}_i(t-1)\right\|^2<\varepsilon$) the set of representatives obtained after adaptation in a time step could be used as the initial conditions for the next time step. Smooth variation of color representatives can be assumed for image sequences if the frame rate is enough and color persistence is relatively high. In the experiments we will work with a time subsample, so the reader must keep in mind that smooth variation is an assumption that may not be satisfied.

3 THE EVOLUTIVE STRATEGY

The idea behind evolutive strategies is to mimic natural selection to solve difficult optimization problems. A population of feasible solutions is proposed. The best ones are selected (sometimes randomly) to built up a new proposition. New solutions (children) are built up from previous ones (parents) by the application of genetically inspired operators: recombination and mutation. A widely accepted pseudocode representation of the algorithm is as follows [14]:

```
t:= 0
initialize P(t)
evaluate P(t)
while not terminate do
        P'(t):= recombine P(t)
        P"(t):= mutate P'(t)
        evaluate P"(t)
        P(t+1):= select (P"(t) U Q)
        t:= t+1
end while
```

The population P(t) is a set of solutions proposed at generation t. The algorithm iterates until some time or optimality condition is met. The **evaluate** operator computes the objetive function for each individual. The **recombine** operator finds mates and recombines them producing a set of offsprings P'(t). This operator is seldom defined in evolutive strategies, we have not defined it in our strategy. The **mutate** operator produces

new individuals by the application of random perturbations. Usually evolutive strategies define these random perturbations as samples of normally distributed random variables. The set Q can be either the set of parents or the empty, depending of the strategy.

In order to define an evolutive strategy, the first decision is the appropriate definition of the population. The common approach is to define each individual as a whole solution. In the case of Clustering, each individual would correspond to a partition of the sample represented by the cluster centres. The fitness of the individual could be straightforwardly defined as the objective function. Individuals would compete to survive as the best solution. We have taken a radically different approach. We have defined each individual as a single cluster center, so that $P(t)=\{\mathbf{y}_i(t); i=1..c\}$. The local fitness of the individual is, then, its local distortion $F_i(t)=\sum_{j=1}^{n}\|\mathbf{x}_j(t)-\mathbf{y}_i(t)\|^2\delta_{ij}(t)$. The solution proposed at time t is given by the entire population. The population as a whole can be evaluated to measure its fitness $F(t)=\sum_{i=1}^{c}\sum_{j=1}^{n}\|\mathbf{x}_j(t)-\mathbf{y}_i(t)\|^2\delta_{ij}(t)$ which corresponds to the objective function to be minimized. The risky hypothesis is that the local optimization of individual cluster distortions will lead to the global optimization of the entire set of cluster centres.

As said before, we have not defined any kind of recombination operator. The mutation operator follows the basic philosophy of evolutionary mutation operators, it is a random perturbation that follows a normal distribution. There are two design questions to answer at this point: (1) Which individuals will be mutated? and (2) How many mutations will be allowed?. The uniform random selection of individuals to be mutated does not seem to be reasonable. Therefore, we have decided to perform a guided selection of the individuals to be subjected to mutation. The guide is to obtain mutations from the individuals with the highest distortion. We have considered two possibilities: the selection of the individuals whose local distortion is greater than the mean of the local distortions in its generation S^1, and the selection of the individual with the highest local distortion S^2. More formally:

$$S^1(t)=\{i|F_i(t)\geq\overline{F}(t)\};S^2(t)=\{k=arg\,max\{F_i(t)\}\}$$

As to the number of mutations we have decided to perform a fixed number of mutations in any case, so that the number of mutations per individual will depend on the selection strategy chosen. The mutation itself is performed adding to the selected individuals pseudorandom samples of a normal random variable: $P''(t)=\{\lambda=\mathbf{y}_i+\mathbf{u};\mathbf{u}\approx N(0,\hat{\Sigma}_i(t))|\mathbf{y}_i\in S^o(t)\}$.

The estimation of the covariance matrix is based on the actual cluster elements assigned to the mutated individual.

$$\hat{\Sigma}_i(t)=(n-1)^{-1}\sum_{j=1}^{n}(\mathbf{x}_j(t)-\mathbf{y}_i(t))(\mathbf{x}_j(t)-\mathbf{y}_i(t))^t\delta_{ij}(t)$$

Finally, to define the selection of the next generation individuals we have followed the so called ($\lambda+\mu$)-strategy. We pool together parents and children:

$$P''(t)\cup Q=\{\mathbf{y}_1,..,\mathbf{y}_c,\mathbf{y}_{c+1},..,\mathbf{y}_{c+m}\}$$

Where m is the number of individual generated by mutation. The fitness function used for selection of an indivual is the distortion when the sample is codified with the codebook given by $P''(t)\cup Q-\{\mathbf{y}_k\}$, more formally:

$$F_k^s(t)=\sum_{i=1;i\neq k}^{c+m}\sum_{j=1}^{n}\|\mathbf{x}_j(t)-\mathbf{y}_i(t)\|^2\delta_{ij}(t)$$

The selection operator selects the c best individuals according to the above fitness. To define formally the selection operator, first consider the set:

$$P'''(t)=\{\mathbf{y}_{i_1},..,\mathbf{y}_{i_{c+m}}|i_j<i_k\Rightarrow F_{i_j}^s(t)>F_{i_k}^s(t)\}$$

Then the specification of the selection operator is: $P(t+1)=\{\mathbf{y}_i\in P'''(t);i=1..c\}$ This selection involves the fitness of the whole population with the addition of the mutatations generated. This makes the algorithm sensitive to the number of mutations generated, forcing the above mentioned restriction to a fixed number of them.

The last critical decision in the design of the evolutive strategy is the mapping of the generation number into the frame number of the image sequence. The more "conventional" approach would consist of computing several generations for each frame. Given our desire to approach almost real time performance, we have also made a critical decision at this point. We have defined a one to one mapping. That is, for each image only one generation of the evolutive strategy is performed. In other words, t is the image number in the sequence.

4 THE EXPERIMENT

The sequence of images used for the experiment is a panning of the laboratory taken with an electronic Apple Quicktake camara. Original images have an spatial resolution of 480x640 pixels. Each two consecutive images overlap 50% of the scene. Figure 1 shows the distribution of the pixels in the RGB unit cube for some of the images in the sequence. This representation clearly demonstrates the time varying nature of the data.

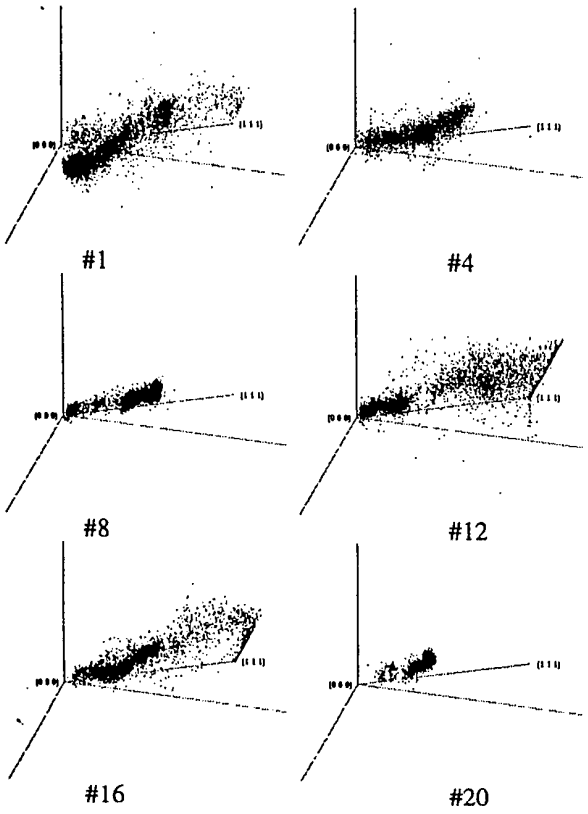


Figure 1. Distribution of pixel colors of some of the images in the experimental sequence

Figures 2 to 5 show the results of the application of the evolutive strategy to random samples of 1600 pixels of each image. As a reference algorithm we have used a variation of the algorithm proposed by Heckbert [1] as implemented in MATLAB. This algorithm recursively partitions the RGB unit cube along the axis of maximum variance. The partition is performed by an orthogonal plane chosen so as to minimize the sum of the residual variances [22]. Color representatives are computed as the center of mass of the resulting partition cubes. This algorithm has been applied in two ways. First it has been applied independently to each image (point line). Second, the set of color representatives obtained with the Heckbert algorithm for the first image has been used to color quantize the entire sequence (dashed line). The difference between both lines gives another view of the time varying nature of the data. As the goal is to show the adaptive properties of the proposed evolutive strategy, we have used as initial cluster centres the first Heckbert palette. The solid line in the figures shows the mean result of 30 replications of the evolutive strategy along with the 95% confidence intervals.

5 CONCLUSIONS AND FURTHER WORK

We have proposed an evolutive strategy for the adaptive computation of color representatives for Color Quantization that can be very efficiently implemented and reach almost real time performance for highly variable color populations. Experimental works show a good response for a realistic sequence of images. Further work will be address to define more robust mutation operators that could reduce the variance of the resuts. Particularly, deterministic mutation operators seem to be desirable.

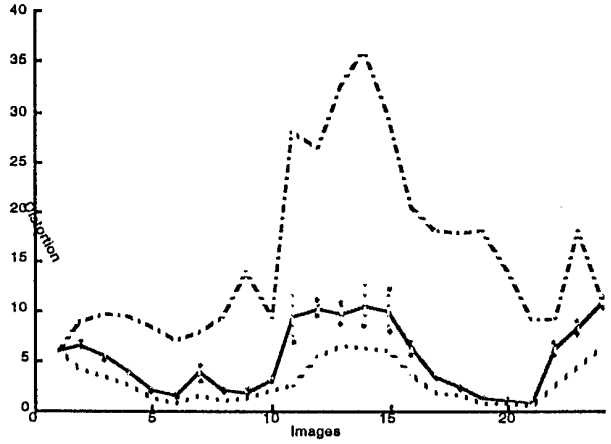


Figure 2: Results of the evolutive strategy with c=16, m=16, and selection S^1 of the mutated individuals.

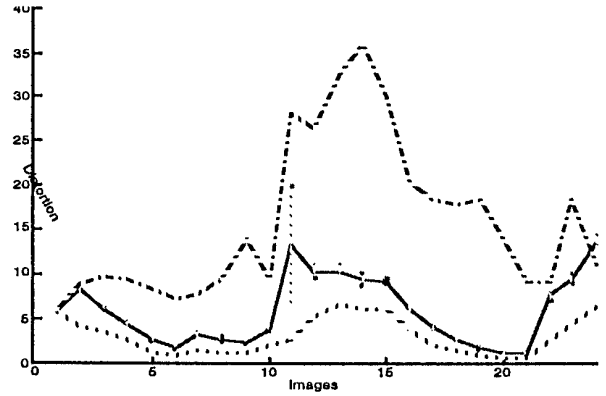


Figure 3: Results of the evolutive strategy with c=16, m=16, and selection S^2 of the mutated individuals.

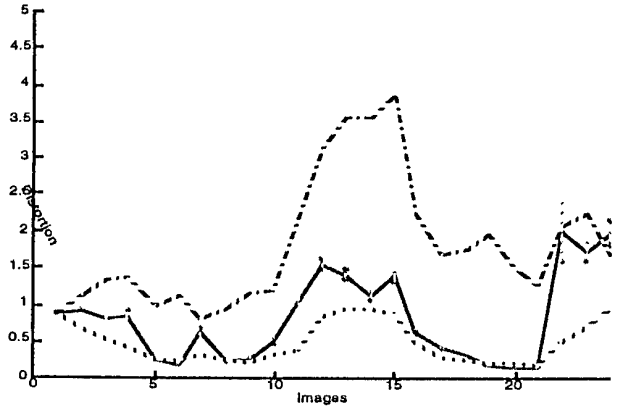


Figure 4: Results of the evolutive strategy with c=256, m=128, and selection S^1 of the mutated individuals.

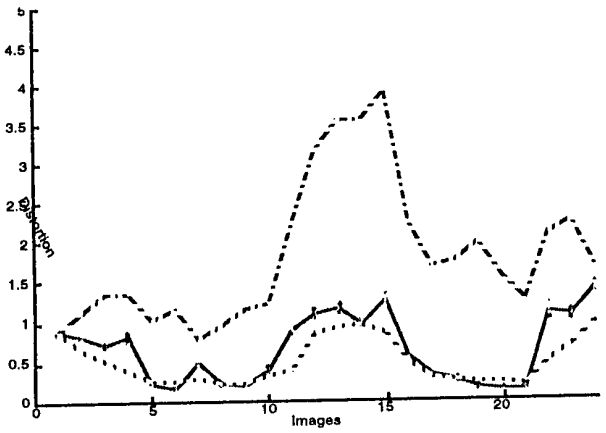


Figure 5: Results of the evolutive strategy with c=256, m=256, and selection S^1 of the mutated individuals.

REFERENCES

[1] P. Heckbert "Color image quantization for frame-buffer display" *Computer Graphics* 16(3) pp.297-307 (1980)

[2] M.T. Orchard, C.A. Bouman Color quantization of images *IEEE trans. Signal Processing* 39(12) pp.2677-2690 (1991)

[3] T.S. Lin, L.W. Chang "Fast color image quantization with error diffusion and morphological operations" *Signal Processing* 43 pp.293-303 (1995)

[4] T. Uchiyama, M. A. Arbib "Color image segmentation using competitive learning" *IEEE trans. Patt. Anal. and Machine Intelligence* 16(12) pp.1197-1206 (1994)

[5] A. Gersho, R.M. Gray "Vector Quantization and signal compression" *Kluwer Acad. Pub.* (1992)

[6] W.K. Pratt "Digital Image Processing" *Wiley* (1991)

[7] M.S. Kankanhalli, B.M. Mehtre, J.K. Wu "Cluster based color matching for image retrieval" *Pattern Recognition* 29(4) pp.701-708 1996

[8] J.A. Lozano, P. Larrañaga, M. Graña "Partitional cluster analysis with Genetic Algorithms: searching for the number of clusters" IFCS-96 Kobe (Japan) March 96

[9] R.J. Chen, B.C. Chien Three dimensional morphological pyramid and its application to color image sequence coding. *Signal Processing* 44 pp.163-180 1995

[10] H.H. Chen, Y.S. Chen, W.H. Hsu "Low rate sequence image coding via vector quantization" *Signal Processing* 26 pp.265-283 1995

[11] Y. Gong, H. Zen, Y. Ohsawa, M. Sakauchi "A color video image quantization method with stable and efficient color selection capability" *Int. Conf. Pattern Recognition* 1992 vol 3 pp.33-36

[12] O.T. Chen, B.J. Chen, Z. Zhang "An adaptive vector quantization based on the gold-washing method for image compression" *IEEE trans cirsuits & systems for video techn.* 4(2) pp.143-156 1994

[13] T. Back, H.P. Schwefel "An overview of Evolutionary Algortihms for parameter optimization" *Evolutionary Computation* 1 pp.1-24 (1993)

[14] T. Back, H.P. Schwefel "Evolutionary computation: an overview" *IEEE Int. Conf. Evolutive Computation* pp.20-29 (1996)

[15] Z. Michalewicz "Evolutionary computation: practical issues" *IEEE Int. Conf. Evolutive Computation* pp.30-39 (1996)

[16] G. P. Babu, N.M. Murty "Clustering eith evolution strategies" *Pattern Recognition* 27 pp.321-329 (1994)

[17] J. Hartigan Clustering Algorithms Wiley 1975

[18] E. Diday, J.C. Simon, Clustering Analysis In K.S. Fu (ed) Digital Pattern Recognition pp.47-94 Springer Verlag 1980

[19] R.D. Duda, P.E. Hart Pattern Classification and Scene Analysis, Wiley 1973

[20] A. K. Jain, R.C. Dubes Algorithms for clustering data Prentice Hall 1988

[21] K. Fukunaga Statistica Pattern Recognition Academic Press 1990

[22] X. Wu "Efficient Statistical Computations for Optimal Color Quantization" In J. Arvo (ed) Graphics Gems II *Academic Press Professional* (1991) pp.126-133

NEAR-PR DESIGN OF NON-UNIFORM FILTER BANKS

F. Argenti, B. Brogelli and E. Del Re

Dipartimento di Ingegneria Elettronica, Universitá di Firenze
Via di Santa Marta, 3 - 50139 Firenze - Italy
Tel.: +39 55 4796 424 - Fax: +39 55 4796 485
e-mail: argenti@cosimo.ing.unifi.it

ABSTRACT

In this work a new method to design filter banks with rational decimation factors is proposed. It aims at the cancellation of the main component of aliasing in the output signal; this imposes a set of conditions on the filters of the analysis/synthesis banks. If cosine-modulation of different linear phase prototypes is used, the aliasing cancellation condition constrains the prototypes relative to adjacent branches to become dependent on each other. A procedure to design the prototypes based on these constraints is proposed and examples of cosine-modulated non-uniform filter banks are presented.

1. INTRODUCTION

Splitting the spectrum of a digital signal can be useful in several applications, for example data compression. Most of the literature in the field of subband coding filter banks design is concerned with uniform width subbands. However, in some cases a non uniform splitting is more suitable, for example in audio coding [1], where non-uniform width subbands could match better the *critical bands* of the human auditory system.

The problem of designing non-uniform filter banks has been addressed, for example, in [2]-[5]. In this work filter banks with rational decimation factors are considered, so extending the work done in [6] related only to integer decimation factors.

The method can be considered a Near Perfect Reconstruction (Near-PR) one since it is based on the cancellation of the main component of aliasing, like in Pseudo-QMF banks [7]. In the case of rational decimation factors banks, however, more than one coupling of the aliasing components of adjacent branches that lead to their cancellation is possible. If cosine modulation is used, the aliasing cancellation constraints involve the prototype filters of each branch. A design procedure is proposed and numerical examples are presented to show the effectiveness of the method.

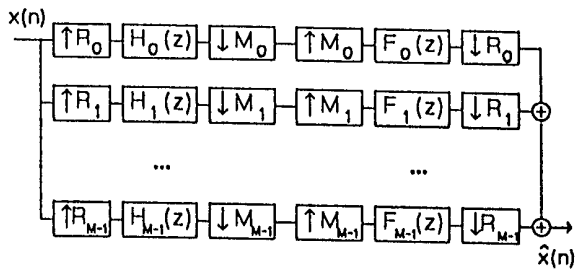


Figure 1: Non-uniform bank with rational decimation factors

2. ALIASING CANCELLATION IN FILTER BANKS WITH RATIONAL SAMPLING FACTORS

Consider the system in Fig. 1, where a non-uniform bank having rational sampling factors R_m/M_m, $m = 0, \ldots,$ M-1, is depicted. The input-output relationship in the z-domain is given by:

$$\hat{X}(z) = \sum_{m=0}^{M-1} \frac{1}{R_m}\frac{1}{M_m} \sum_{p=0}^{R_m-1} F_m(z^{\frac{1}{R_m}} W_{R_m}^{p}) \cdot H_m(z^{\frac{1}{R_m}} W_{R_m}^{p}) X(z) + \sum_{m=0}^{M-1} \frac{1}{R_m}\frac{1}{M_m} \sum_{l=1}^{M_m-1} \sum_{p=0}^{R_m-1} F_m(z^{\frac{1}{R_m}} W_{R_m}^{p}) \cdot H_m(z^{\frac{1}{R_m}} W_{R_m}^{p} W_{M_m}^{l}) X(z W_{M_m}^{l R_m}) \quad (1)$$

where $W_M = e^{-j2\pi/M}$. Eq. (1) highlights the reconstruction transfer function and the aliasing components.

Consider the analysis stage of each branch shown in

Fig. 1. Real coefficients filters are taken into account and, therefore, the frequency response of each filter has passbands located at positive and negative frequencies, symmetrically with respect to the origin. The passband at positive (negative) frequencies has width π/M_m and is centered in $(k_m + 0.5)\pi/M_m$ $(-(k_m + 0.5)\pi/M_m)$. The value k_m is an integer and selects which part of the spectrum of the R_m-fold upsampled input signal must be extracted. For example, to extract the spectrum in the frequency interval $[\pi/5, 3\pi/5]$ $([-3\pi/5, -\pi/5])$ we use R_m=3, M_m=5 and k_m=2. If we consider the frequency response of each filter of the analysis/synthesis banks approximately equal to zero in their stopbands, then the filters transfer functions can be expressed as:

$$H_m(z) = U_m(z) + V_m(z) \tag{2}$$

$$F_m(z) = \hat{U}_m(z) + \hat{V}_m(z) \tag{3}$$

where $U_m(\omega)$ and $\hat{U}_m(\omega)$ have a passband for $\omega > 0$, while $V_m(\omega)$ and $\hat{V}_m(\omega)$ have a passband for $\omega < 0$.

Due to the M_m-fold upsampler in the synthesis stage, *images* of the *m-th* subband spectrum are filtered by $F_m(z)$. The main aliasing terms are created at the high-frequency and at the low-frequency edges of the passband of $F_m(z)$. These components have been described for a cosine-modulated uniform bank in [7]. If we consider that in a rational decimation factors bank each branch operates on an R_m-fold upsampled version of $x(n)$ and we retain only the more relevant terms, then the aliasing terms can be written as:

$$A_m^{(low)}(z) = \frac{1}{M_m}\left[\hat{U}_m(z)V_m(zW_{M_m}^{k_m})X(z^{R_m}W_{M_m}^{k_m R_m}) + \right. \\ \left. + \hat{V}_m(z)U_m(zW_{M_m}^{-k_m})X(z^{R_m}W_{M_m}^{-k_m R_m})\right] \tag{4}$$

$$A_m^{(high)}(z) = \frac{1}{M_m}\left[\hat{U}_m(z)V_m(zW_{M_m}^{(k_m+1)})X(z^{R_m}W_{M_m}^{(k_m+1)R_m}) + \right. \\ \left. + \hat{V}_m(z)U_m(zW_{M_m}^{-(k_m+1)})X(z^{R_m}W_{M_m}^{-(k_m+1)R_m})\right] \tag{5}$$

In [7] it is shown that for uniform cosine-modulated filter banks the component $A_m^{(high)}(z)$ of the *m-th* branch is canceled by the component $A_{m+1}^{(low)}(z)$ of the *(m+1)-th* branch. In the non-uniform case, we have to consider that the cancellation may occur also by coupling the *(high)-(high)*, *(low)-(low)* or *(low)-(high)* aliasing terms coming from the *m-th* and the *(m+1)-th* branch, i.e., the following cases must be taken into account:

a) $A_m^{(high)}(z) \downarrow R_m + A_{m+1}^{(high)}(z) \downarrow R_{m+1} = 0$

b) $A_m^{(low)}(z) \downarrow R_m + A_{m+1}^{(low)}(z) \downarrow R_{m+1} = 0$

c) $A_m^{(low)}(z) \downarrow R_m + A_{m+1}^{(high)}(z) \downarrow R_{m+1} = 0$

d) $A_m^{(high)}(z) \downarrow R_m + A_{m+1}^{(low)}(z) \downarrow R_{m+1} = 0$

where $Q(z) \downarrow M$ stands for the z-transform of the M-fold subsampled version of $q(n)$. For example, consider the bank $\{1/5, 3/5, 1/5\}$ that can be implemented using filters having a passband equal to $\pi/5$ and centered, on the positive frequency axis, in $\pi/10$, $\pi/2$ and $9\pi/10$. The aliasing term $A_0^{(high)}(z)$ produced in the *m=0* branch at the synthesis stage must be canceled by $A_1^{(high)}(z) \downarrow 3$.

Consider the *(high)-(high)* case: substituting (5) into case *a)* equation yields an expression that can be split into two systems: if $W_{M_m}^{(k_m+1)R_m} = W_{M_{m+1}}^{(k_{m+1}+1)R_{m+1}}$ then the following must be verified

$$\begin{cases} \frac{1}{M_m}[\hat{U}_m(z)V_m(zW_{M_m}^{(k_m+1)})] \downarrow R_m + \\ + \frac{1}{M_{m+1}}[\hat{U}_{m+1}(z)V_{m+1}(zW_{M_{m+1}}^{(k_{m+1}+1)})] \downarrow R_{m+1} = 0 \\ \\ \frac{1}{M_m}[\hat{V}_m(z)U_m(zW_{M_m}^{-(k_m+1)})] \downarrow R_m + \\ + \frac{1}{M_{m+1}}[\hat{V}_{m+1}(z)U_{m+1}(zW_{M_{m+1}}^{-(k_{m+1}+1)})] \downarrow R_{m+1} = 0 \end{cases} \tag{6}$$

otherwise, if $W_{M_m}^{(k_m+1)R_m} = W_{M_{m+1}}^{-(k_{m+1}+1)R_{m+1}}$ then the following must be verified

$$\begin{cases} \frac{1}{M_m}[\hat{U}_m(z)V_m(zW_{M_m}^{(k_m+1)})] \downarrow R_m + \\ + \frac{1}{M_{m+1}}[\hat{V}_{m+1}(z)U_{m+1}(zW_{M_{m+1}}^{-(k_{m+1}+1)})] \downarrow R_{m+1} = 0 \\ \\ \frac{1}{M_m}[\hat{V}_m(z)U_m(zW_{M_m}^{-(k_m+1)})] \downarrow R_m + \\ + \frac{1}{M_{m+1}}[\hat{U}_{m+1}(z)V_{m+1}(zW_{M_{m+1}}^{(k_{m+1}+1)})] \downarrow R_{m+1} = 0 \end{cases} \tag{7}$$

Similar equations can be written also for the other possible couplings of aliasing components.

3. COSINE-MODULATED NON-UNIFORM BANKS

The use of cosine modulation simplifies the fulfillment of the aliasing cancellation condition. Suppose that each filter of the analysis/synthesis banks is obtained as follows:

$$\begin{aligned} h_m(n) &= 2g_m(n)\cos((2k_m+1)\tfrac{\pi}{2M_m}(n - \tfrac{N_m-1}{2}) + \theta_m) \\ f_m(n) &= 2g_m(n)\cos((2k_m+1)\tfrac{\pi}{2M_m}(n - \tfrac{N_m-1}{2}) - \theta_m) \\ &= h_m(N_m - 1 - n) \end{aligned} \tag{8}$$

for $m = 0, 1, \ldots, M-1$. N_m is the length of $g_m(n)$. The prototypes $g_m(n)$ have a linear phase and satisfy $g_m(n) = g_m(N_m - 1 - n)$. The phase terms θ_m are chosen to satisfy the aliasing cancellation constraints.

In the case of cosine-modulated banks, the following relationships hold:

$$\begin{aligned}
U_m(z) = \; & G_m(zW_{2M_m}^{(k_m+(1/2))})W_{2M_m}^{(k_m+(1/2))(N_m-1)/2}\cdot \\
& \cdot e^{j\theta_m} \\
V_m(z) = \; & G_m(zW_{2M_m}^{-(k_m+(1/2))})W_{2M_m}^{-(k_m+(1/2))(N_m-1)/2}\cdot \\
& \cdot e^{-j\theta_m} \\
\hat{U}_m(z) = \; & G_m(zW_{2M_m}^{(k_m+(1/2))})W_{2M_m}^{(k_m+(1/2))(N_m-1)/2}\cdot \\
& \cdot e^{-j\theta_m} \\
\hat{V}_m(z) = \; & G_m(zW_{2M_m}^{-(k_m+(1/2))})W_{2M_m}^{-(k_m+(1/2))(N_m-1)/2}\cdot \\
& \cdot e^{j\theta_m}
\end{aligned} \tag{9}$$

Consider, for example, the *(high)-(high)* case. Substituting the above expressions into the aliasing cancellation constraints (6) and (7) yields a relationship between the prototypes of adjacent branches. In [8] it is shown that for $W_{M_m}^{(k_m+1)R_m} = W_{M_{m+1}}^{\pm(k_{m+1}+1)R_{m+1}}$ the choice $e^{-j2\theta_m} + e^{\mp j2\theta_{m+1}} = 0$ allows aliasing cancellation if the following relationship holds (the same result is obtained considering the cases *b)-d)*, but with a different relationship between the phase terms θ_m):

$$\begin{aligned}
& \frac{1}{M_m}\frac{1}{R_m}\sum_{p=0}^{R_m-1} G_m(z^{1/R_m}W_{R_m}^p)\cdot \\
& \cdot G_m(z^{1/R_m}W_{R_m}^p W_{2M_m}) = \frac{1}{M_{m+1}}\frac{1}{R_{m+1}}\sum_{p=0}^{R_{m+1}-1} \\
& G_{m+1}(z^{1/R_{m+1}}W_{R_{m+1}}^p W_{4M_m}^{R_m/R_{m+1}}W_{4M_{m+1}}^{-1})\cdot \\
& \cdot G_{m+1}(z^{1/R_{m+1}}W_{R_{m+1}}^p W_{4M_m}^{R_m/R_{m+1}}W_{4M_{m+1}})
\end{aligned} \tag{10}$$

Moreover, it is possible to demonstrate the following facts [8], which outline also the steps of the procedure to design rational sampling factors filter banks.

By using the zero-phase representation of the prototypes, i.e.,

$$G_m^{(zp)}(\omega) = G_m(\omega)e^{j\frac{N_m-1}{2}\omega} \tag{11}$$

and by imposing the condition

$$\frac{N_m-1}{R_m} = \frac{N_{m+1}-1}{R_{m+1}} \tag{12}$$

on the lengths of the prototypes, the constraint of aliasing cancellation reduces to the following relationship between the zero-phase frequency responses of the prototype filters

$$\begin{aligned}
& \frac{1}{M_m}\frac{1}{R_m}G_m^{(zp)}(\frac{\omega}{R_m})G_m^{(zp)}(\frac{\omega}{R_m}-\frac{\pi}{M_m}) = \\
& \frac{1}{M_{m+1}}\frac{1}{R_{m+1}}G_{m+1}^{(zp)}(\frac{\omega}{R_{m+1}}-\frac{\pi R_m}{2M_mR_{m+1}}+\frac{\pi}{2M_{m+1}})\cdot \\
& \cdot G_{m+1}^{(zp)}(\frac{\omega}{R_{m+1}}-\frac{\pi R_m}{2M_mR_{m+1}}-\frac{\pi}{2M_{m+1}})
\end{aligned} \tag{13}$$

Let $\omega_{c,m} = \pi/(2M_m)$ be the cut-off frequency of $G_m(\omega)$. Suppose the transition band, having width $\Delta\omega_m$, is centered in $\omega_{c,m}$ and let $\omega_{p,m} = \omega_{c,m}-(\Delta\omega_m/2)$ and $\omega_{s,m} = \omega_{c,m} + (\Delta\omega_m/2)$ be the upper bound of the passband and the lower bound of the stopband, respectively, of $G_m(\omega)$. Therefore, the constraint (13) is satisfied if $R_m\Delta\omega_m = R_{m+1}\Delta\omega_{m+1}$ and if $G_{m+1}^{(zp)}(\omega)$ is chosen as follows

$$G_{m+1}^{(zp)}(\omega) =$$

$$= \begin{cases}
0 & -\pi < \omega \le -\omega_{s,m+1} \\
\sqrt{\frac{R_{m+1}M_{m+1}}{R_mM_m}}\, G_m^{(zp)}(-\omega_{p,m}+(\omega+\omega_{p,m+1})\cdot & \\
\cdot\frac{\omega_{s,m}-\omega_{p,m}}{\omega_{s,m+1}-\omega_{p,m+1}}) & -\omega_{s,m+1} < \omega \le -\omega_{p,m+1} \\
\sqrt{R_{m+1}M_{m+1}} & -\omega_{p,m+1} < \omega \le \omega_{p,m+1} \\
\sqrt{\frac{R_{m+1}M_{m+1}}{R_mM_m}}\, G_m^{(zp)}(\omega_{p,m}+(\omega-\omega_{p,m+1})\cdot & \\
\cdot\frac{\omega_{s,m}-\omega_{p,m}}{\omega_{s,m+1}-\omega_{p,m+1}}) & \omega_{p,m+1} < \omega \le \omega_{s,m+1} \\
0 & \omega_{s,m+1} < \omega \le \pi
\end{cases} \tag{14}$$

Assuming the aliasing components have been completely eliminated, the input-output relationship shown in (1) can be expressed by

$$\begin{aligned}
\hat{X}(z) = \; & \Big[\sum_{m=0}^{M-1}\frac{1}{R_m}\frac{1}{M_m}\sum_{p=0}^{R_m-1}F_m(z^{\frac{1}{R_m}}W_{R_m}^p)\cdot \\
& \cdot H_m(z^{\frac{1}{R_m}}W_{R_m}^p)\Big]X(z) = T(z)X(z)
\end{aligned} \tag{15}$$

Phase error is absent if the synthesis filters are a time reversed version of the analysis filters, while the magnitude error is maintained at low levels if $T(z)$ is approximately allpass. The reconstruction error is reduced choosing prototype filters with high stopband attenuation and also with a proper behavior in the transition band.

A first prototype is designed (by using known techniques, for example those shown in [9][10][5]). This prototype is relative to the *m-th* branch, where m must be chosen so that R_m/M_m=min$\{R_k/M_k,\ k$=0,...,M-1$\}$. Its cut-off frequency is $\omega_{c,m} = \frac{\pi}{2M_m}$ and $\sqrt{R_mM_m}$ is the gain in the passband. $G_m(\omega)$ must have a *power complementary* transition band, i.e., satisfies

$$\begin{aligned}
& |\, G_m(\omega)\,|^2 + |\, G_m(\tfrac{\pi}{M_m}-\omega)\,|^2 = R_mM_m \\
& for\ \omega_{p,m} < \omega < \omega_{s,m}
\end{aligned} \tag{16}$$

The prototypes in the other branches are obtained by using (14) and by adding the correct linear phase

term to determine $G_{m+1}(\omega)$; $g_{m+1}(n)$ is obtained by the inversion of $G_{m+1}(\omega)$.

It can be shown that prototypes designed by using this procedure make $T(z)$ approximately allpass.

4. EXPERIMENTAL RESULTS

To show the effectiveness of the design procedure described in the previous section we consider three examples of non-uniform banks. Example 1 and 2 are relative to banks with rational sampling factors, while Example 3 refers to an integer sampling factors bank, suitable for audio coding applications, that has been proposed in [11]. We indicate with K and Θ the sets { k_m, m=0,...,M-1 } and { θ_m, m=0,...,M-1 }, respectively.

Example 1: Bank { 1/5, 3/5, 1/5 }. Two prototypes need to be designed ($g_0(n) = g_2(n)$); the couplings of aliasing components that must be considered are *(high)-(high)* and *(low)-(low)* between the branches *0-1* and *1-2*, respectively; K={ 0, 2, 4}; Θ = { $\pi/4$, $\pi/4$, $\pi/4$ }.

Example 2: Bank { 2/7, 2/7, 2/7, 1/7}. Two prototypes have to be designed ($g_0(n) = g_1(n) = g_2(n)$). In this example more than one choice is possible for K. We will use K={ 0, 5, 4, 6 } to show the largest variety of couplings of aliasing components (*(high)-(high)*, *(low)-(high)*, *(low)-(low)*, in the order). In this case Θ = { $\pi/4$, $\pi/4$, $-\pi/4$, $-\pi/4$ }.

Example 3: Non-uniform bank having 16, 32 and 64 as possible decimation factors and allowing the splitting of an audio signal sampled at 48 kHz as shown in Fig. 2.

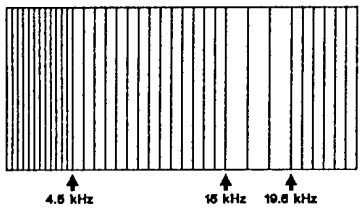


Figure 2: Subband splitting relative to Example 3

The performance of the presented design method is evaluated in terms of both the overall distortion function $T(\omega)$ and the residual aliasing error. As to the latter error, a global measure relative to the whole structure is used in this work. According to the input-output relationship in (1), the aliasing contribution relative to $X(zW_{M_m}^{lR_m})$ can be written as

$$A_{l,m}(z) = \frac{1}{R_m}\frac{1}{M_m}\sum_{p=0}^{R_m-1} H_m(z^{\frac{1}{R_m}} W_{R_m}^{p} W_{M_m}^{l}) \cdot F_m(z^{\frac{1}{R_m}} W_{R_m}^{p}) \tag{17}$$

with m=0,...,M-1, l=0,...,$M_m - 1$. The functions $A_{l,m}(\omega)$ are 2π-periodic functions. All the aliasing terms $A_{l,m}(z)$ that refer to the same shifted version of $X(z)$, i.e., having the same value of $W_{M_m}^{lR_m}$, must be summed up, so that the following aliasing error can be defined:

$$E_a(\omega) = \sqrt{\sum_{r=1}^{M_{max}-1} \left| \sum_{m=0}^{M-1} \sum_{l=1,(lR_m) \bmod M_m = r}^{M_m-1} A_{l,m}(\omega) \right|^2} \tag{18}$$

where $M_{max} = \max\{M_m, m=0,...,M-1\}$ and where the inner summation in (18) is evaluated only for the values of l and m satisfying the condition $(lR_m) \bmod M_m = r$.

Therefore

$$E_{p-p} = \max_{0\le\omega\le\pi} |T(\omega)| - \min_{0\le\omega\le\pi} |T(\omega)| \tag{19}$$

$$E_{a,max} = \max_{0\le\omega\le\pi} E_a(\omega) \tag{20}$$

can be used as measures of the quality of the designed banks.

Tables 1 and 2 report the results obtained for Example 1 and 2, respectively, for different lengths of the prototypes. As can be seen, both the magnitude distortion and the aliasing error are kept small

In Fig. 3 the frequency responses of the final cosine modulated analysis filters relative to Example 2 and obtained with prototypes having 82 and 163 coefficients are shown: from the inspection of this figure it can be seen that the design based on (14) does not degrade the passband and the stopband characteristics of the new prototypes.

In Fig. 4 the final bank relative to Example 3 and obtained with filter lengths equal to 512 is shown: the reconstruction and the aliasing error are $E_{p-p} = 3.88E-03$ and $E_{a,max} = 8.99E-03$, respectively.

5. CONCLUSIONS

In this work a method to design non-uniform filter banks with rational sampling factors has been presented. Aliasing cancellation constraints have been applied to cosine-modulated banks. A simple procedure, that requires numerical optimization of only one prototype, being the others derived in a straightforward way from this one, has been proposed.

6. REFERENCES

[1] N. Jayant, J. Johnston and R.Safranek, "Signal Compression Based on Models of Human Perception", *Proceedings of the IEEE*, Vol. 81, no. 10, pp. 1385-1422, Oct. 1993.

[2] P.Q. Hoang and P.P. Vaidyanathan, "Non-uniform Multirate Filter Banks: Theory and Design", in *Proc. Int. Symp. Circuits Syst.*, pp. 371-374, May 1986.

[3] J. Kovacevic and M. Vetterli, "Perfect Reconstruction Filter Banks with Rational Sampling Factors", *IEEE Trans. Signal Processing*, Vol. 41, no. 6, pp. 2047-2066, Jun. 1993.

[4] S. Wada, "Design of Nonuniform Division Multirate FIR Filter Banks", *IEEE Trans. Circuits Syst. II*, Vol. 42, no. 2, pp. 115-121, Feb. 1995.

[5] J. Princen, "The Design of Nonuniform Modulated Filterbanks", *IEEE Trans. Signal Processing*, Vol. 43, no. 11, pp. 2550-2560, Nov. 1995.

[6] F. Argenti and E. Del Re, "Non-uniform filter banks based on a multi-prototype cosine modulation", *IEEE ICASSP'96*, Atlanta, May 1996, pp. 1511-1514.

[7] R.D Koilpillai and P.P. Vaidyanathan, "A Spectral Factorization Approach to Pseudo-QMF Design", *IEEE Trans. Signal Processing*, Vol. 41, no. 1, pp. 82-92, Jan. 1993.

[8] F. Argenti B. Brogelli and E. Del Re, "Design of filter banks with rational sampling factors based on a multi-prototype cosine modulation", submitted to *IEEE Trans. Signal Processing.*

[9] R.D Koilpillai and P.P. Vaidyanathan, "Cosine-Modulated FIR Filter Banks Satisfying Perfect Reconstruction", *IEEE Trans. Signal Processing*, Vol. 40, no. 4, pp. 770-783, Apr. 1992.

[10] T.Q. Nguyen, "Near-Perfect-Reconstruction Pseudo-QMF Banks", *IEEE Trans. Signal Processing*, Vol. 42, no. 1, pp. 65-76, Jan. 1994.

[11] F.Argenti, V.Cappellini, E.Del Re, A.Fiorilli, "Non-uniform subband analysis banks for the compression of audio signals", Proc. 1st Workshop on Sampling Theory and Applications, Jurmala, Latvia, 20-22 Sept. 1995, pp. 285-289.

Table 1: Results relative to Example 1

N_0, N_2	N_1	E_{p-p}	$E_{a,max}$
36	106	3.42 E−03	1.82 E−02
46	136	4.33 E−03	3.79 E−03
56	166	1.52 E−03	2.93 E−04

Table 2: Results relative to Example 2

N_0, N_1, N_2	N_3	E_{p-p}	$E_{a,max}$
67	34	6.67 E−02	3.00 E−02
83	42	3.06 E−02	1.20 E−02
123	62	6.45 E−03	4.36 E−03
163	82	4.04 E−03	9.15 E−04

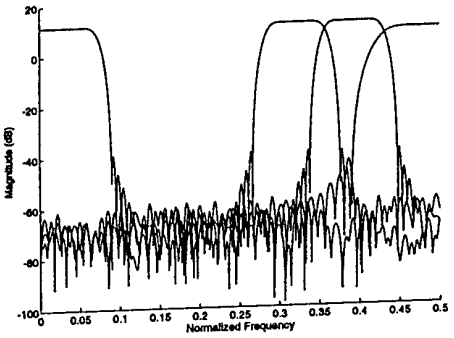


Figure 3: Cosine-modulated bank relative to Example 2 ($N_0 = N_1 = N_2 = 163$, $N_3 = 82$)

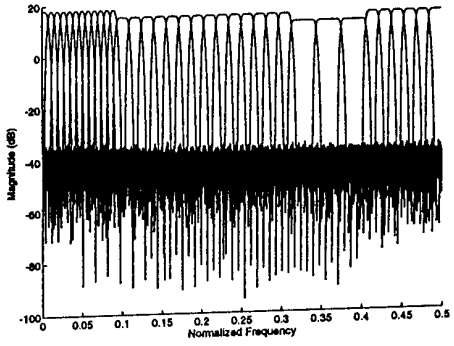


Figure 4: Filter bank relative to Example 3 (N=512)

ORTHOGONALIZATION OF A FRAME-BASED WAVELET SUBSPACE FOR SIGNAL COMPRESSION AND NOISE REDUCTION

L. Rebollo Neira, J. Fernandez Rubio and A. Constantinides†

Departament de Teoria del Senyal i Comunicacions, Escola Tecnica Superior
d'Enginyers de Telecomunicació, Campus Nord, UPC,
Edifici D-4, Apdo 30002, 08080 Barcelona, Spain.

† Signal Processing Section, Department of Electrical and Electronic Engineering,
Imperial College of Science, Technology and Medicine, London SW7 2BT, UK.

ABSTRACT

A mathematical scheme for simultaneous signal compression and noise reduction is presented in this contribution. Initially the use of well-localized wavelet is proposed as derived from the general theory of frames [1, 2, 3, 4], in order to generate a representation subspace capable of reproducing the original signal while excluding the additive noise.

The representation subspace however, is shown to be efficient for noise reduction but in its initial form creates an ill-conditioned inverse problem. This is related to the norm of the wavelet expansion coefficients which may be very large in magnitude. In our treatment we show clearly that the ill-conditioned problem can be avoided simply by adopting an orthonormal representation for the wavelet-generated subspace. The mathematical framework of our approach allows us to develop a method to construct explicitly the orthonormal representation in a natural way. The new representation preserves the signal norm and improves the compactness of the subspace with respect to its compression properties.

1. FRAMEWORK

To represent a signal, $f(t)$, we fix a subspace, S, by choosing a finite set of wavelet functions as those propose in [4], pp 79, (see Fig 1), and expand the signal as:

$$f(t) = \sum_{m\in Z_M}\sum_{n\in Z_N} c_{m,n}\phi_{m,n}(t) \quad ; \quad t\in[0,T] \qquad (1)$$

where

$$Z_M = \{m\in Z \ ; \ m_1 \le m \le m_2 \ ; \ M = |m_2+1| - m_1\} \qquad (2)$$

$$Z_N = \{n\in Z \ ; \ n_1 \le n \le n_2 \ ; \ N = |n_2+1| - n_1\}. \qquad (3)$$

The transformation (1) has no inverse in a formal sense. However, a solution of minimum norm may be found as:

$$c_{m,n} = \sum_{k\in Z_M}\sum_{j\in Z_N}\sum_{l=1,\lambda_l\neq 0}^{MN} \langle m,n|\psi_l\rangle \frac{1}{\lambda_l}\langle\psi_l|k,j\rangle\langle\phi_{k,l}|f\rangle \qquad (4)$$

where the vectors $|\psi_l\rangle$ satisfy the eigenvalue equations [5]:

$$\sum_{k\in Z_M}\sum_{j\in Z_N}\langle\phi_{m,n}|\phi_{k,j}\rangle\langle k,j|\psi_l\rangle = \lambda_l\langle m,n|\psi_l\rangle \quad ; \quad (5)$$
$$l = 1,\ldots,MN \ ; \ m\in Z_M \ ; \ n\in Z_N$$

The inner products $\langle\phi_{m,n}|\phi_{k,j}\rangle$ and $\langle m,n|\psi_l\rangle$ are performed in $L^2([0,T])$ and in the space of square summable sequencies respectively, with $\langle m,n|k,j\rangle = \delta_{m,k}\delta_{n,j}$. In order to gain accuracy in the determination of the eigenvalues to be considered as different from zero, instead of solving (6) we calculate directly the singular values $\beta_l = \lambda_l^{1/2}$ and singular vectors $|\psi_l\rangle$ of the matrix $\phi_{m,n}(t_i)$; $m\in Z_M$; $n\in Z_N$; $i=,\ldots,N_T$, where N_T is the number of samples that are taken in discretizing the $[0,T]$ interval. Unfortunately, as it can be seen from (4), when the spectrum λ_l ; $l=1,\ldots,MN$ has fast decay rate, the coefficients $c_{m,n}$ are of large magnitude and hence the representation (1) becomes "non-economical". This problem can be easily overcome by noticing that the eigenvectors $|\psi_l\rangle$, which belong to the square summable sequences, can be transformed to provide an orthogonal set of functions $\bar{\varphi}_l(t)$, that span the wavelet-generated subspace. These are

calculated in the form [6]:

$$\bar{\varphi}_l(t) = \frac{1}{\lambda_l} \sum_{k \in Z_M} \sum_{j \in Z_N} \langle \psi_l | k, j \rangle \phi_{k,j}(t) \ ; \ \lambda_l \neq 0. \quad (6)$$

Through this new set of functions, we have an alternative representation for the signal as:

$$f(t) = \sum_{l=1, \lambda_l \neq 0}^{MN} c_j \bar{\varphi}_l(t) \quad (7)$$

where

$$c_l = \frac{1}{\lambda_l} \sum_{k \in Z_M} \sum_{j \in Z_N} \langle \psi_l | k, j \rangle \langle \phi_{k,j} | f \rangle \ ; \ \lambda_l \neq 0. \quad (8)$$

Since the transformation (7) is unitary, the norm of the coefficients c_j is equal to the signal norm and this is a more economical representation.

It can be shown that, for a signal outside the subspace S, both, representation (1) in terms of wavelets and the orthogonal representation (7), are identical approximations for such a signal [6].

Let us denote by $f^*(t)$ the signal when it is corrupted by zero mean random noise of variance σ^2. In order to reduce the noise effect, we seek an approximation of $f^*(t)$ in S within a degree of precision that takes into account the assumed known variance of the noise. This precision has to be used to fix the dimension of the representation subspace S. Notice that, according to definitions (2) and (3), the subspace is fixed by given the numbers m_1, m_2, n_1, n_2. From these four numbers, only m_1 has to be precise, since the functions $\phi_{m,n}(t)$ become sharper as m decreases and hence large negative value of m render the functions susceptible to reproducing random noise. The upper bound m_2 may be overestimated without causing undue effects. The bounds n_1 and n_2 for Z_N, are merely estimated so that the considered time interval is covered by the support of the functions involved. The crucial value of m_1, is proposed to be fixed as the maximum value for which the following is satisfied:

$$E\{||f^*(t) - f(t)||^2\} \leq \sigma^2. \quad (9)$$

$E\{.\}$ denotes the mean value operation, $f^*(t)$ is the noisy signal and $f(t)$ corresponds to the equivalent approximations (1) or (7).

2. NUMERICAL TEST

Fig 2 shows 500 samples of noisy data which are simulated by adding Gaussian noise of variance $\sigma^2 = 0.2$ to the original signal represented as a continuous curve in the same figure. In Fig 3, the continuous line represents the original clean signal for comparison and the dotted line represents the reconstruction obtained through both, the wavelet (1) and orthonormal representation (7).

Fig.4 shows the coefficients of the wavelet representation, notice that these are very large. On the other hand, the triangles in Fig.5 correspond to the values of the coefficients of the orthonormal expansion, and this is clearly a more economical representation.

Fig 6 shows 500 samples of noisy data which are simulated by adding Gaussian noise of variance $\sigma^2 = 0.4$ to a sinusoid whose phase changes randomly at $t = 0.25, 0.5, 0.75$. The continuous curve of Fig 7 represents the original clean signal for comparison and the dotted line represents the reconstruction obtained through both, the wavelet (1) and orthonormal representation (7).

Fig 8 shows the coefficients of the wavelet representation for this case, and they clearly have the same feature as above, namely, that they are very large numbers. A more reasonable set of coefficients is obtained through the orthonormal representation as shown in Fig 9.

3. CONCLUSIONS

In the scheme for data compression and noise reduction we have presented, the used of frame-based wavelets is proposed as a starting point. However, the fact that we deal with a finite subset of frame elements, and that we build the dual vectors in this subspace, implies that the inversion becomes an ill-conditioned problem whereby the norm of the coefficients, used to represent the signal, is very large. In out treatment the problem is avoided simply by adopting an orthonormal representation for the wavelet-generated subspace. In addition to preserving the signal norm, the proposed orthogonal representation eliminates redundancy improving compression performance.

4. REFERENCES

[1] R. J. Duffin, A. C. Shaffer, "A Class of Nonharmonic Fourier Series", *Trans. Amer. Math. Soc.*, Vol 72, pp 341-366, (1952).

[2] R. M. Young, "An introduction to Nonharmonic Fourier Series", *Academic Press*, New York, (1980).

[3] I. Daubechies, "The Wavelets Transform, Time Frequency Localization and Signal Analysis",

IEEE Trans. Inform. Theory Vol 36, pp 961-1005, (1990).

[4] I. Daubechies, "Ten Lectures on Wavelets" *CBMAS-NSF*, SIAM, Philadelphia, (1992).

[5] M. Reed, B. Simon, "Functional Analysis", *Academic Press*, New York, (1980).

[6] L. Rebollo Neira, A. Constantinides, "Signal representation for compression and noise reduction through frame-based wavelets". *To be published*

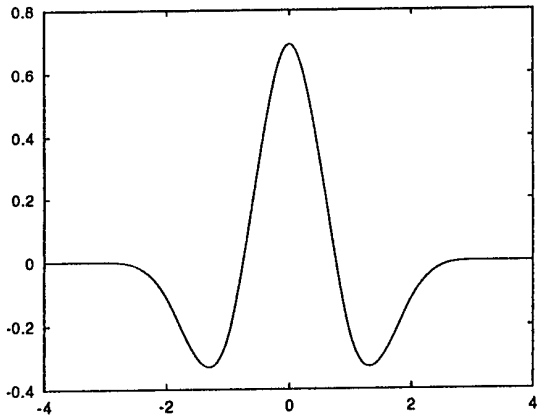


Figure 1: Mother Wavelet $\phi(t)$

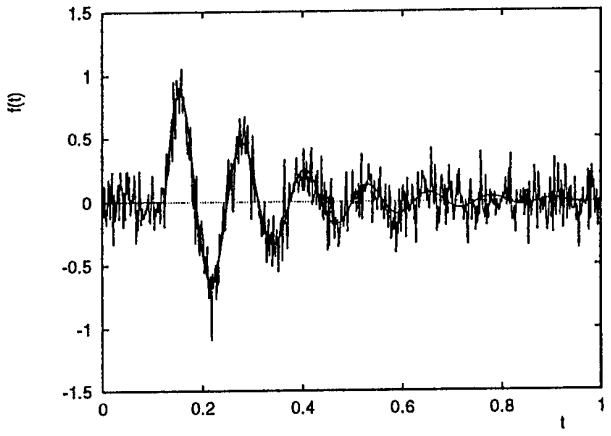


Figure 2: Input noisy data

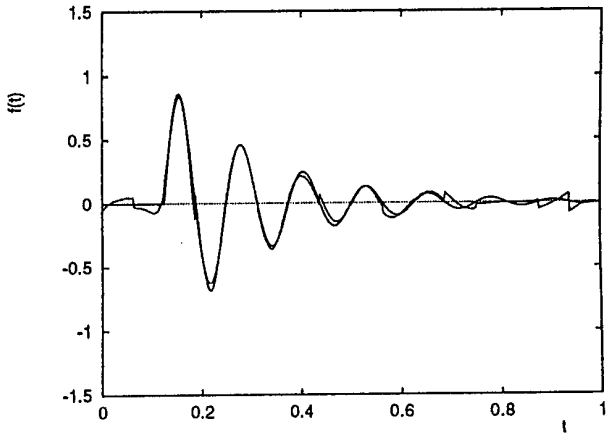


Figure 3: The continuous curve plots the original signal. The dotted line plots the reconstruction obtained through both, the wavelet and orthonormal representations

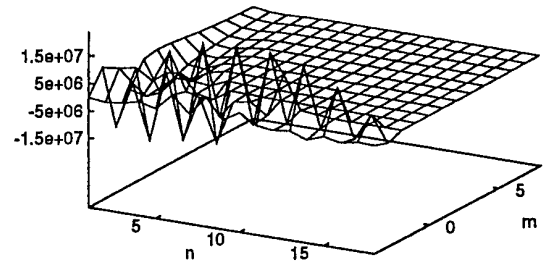


Figure 4: The vertical axis shows the wavelet representation coefficients $c_{m,n}$

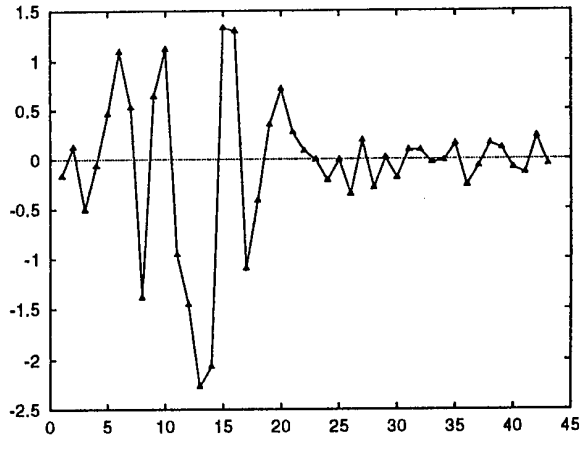


Figure 5: Orthogonal representation coefficients c_l

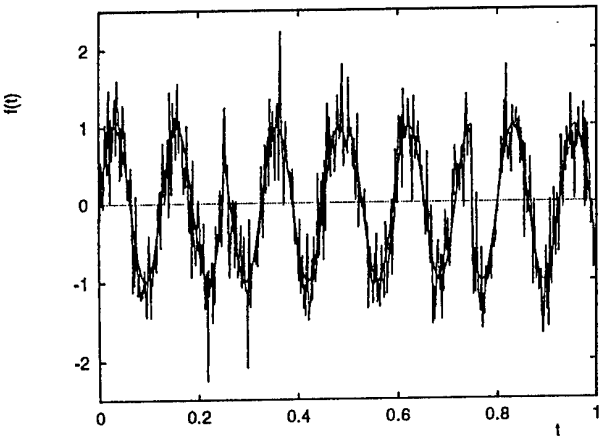


Figure 6: Input noisy data.

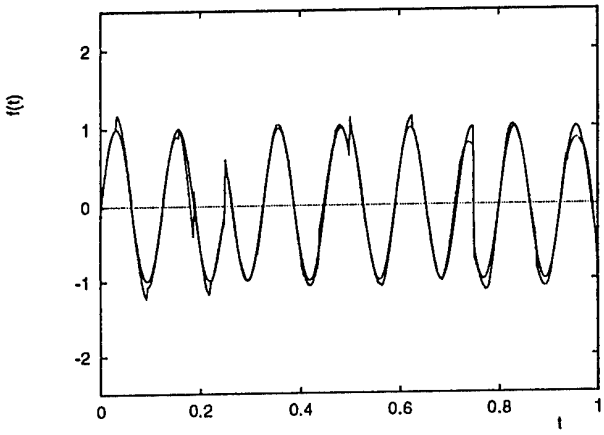


Figure 7: The continuous curve plots the original signal. The dotted line plots the reconstruction obtained through both, the wavelet and orthonormal representations.

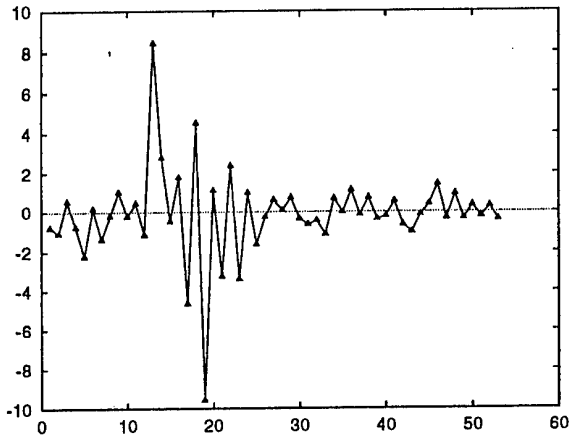


Figure 9: Orthogonal representation coefficients c_l

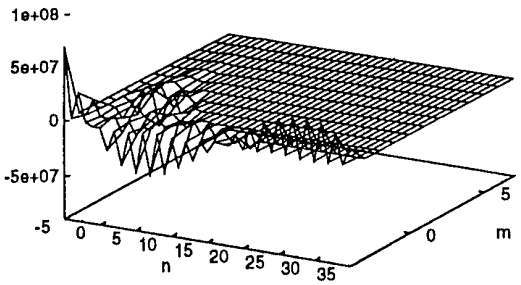


Figure 8: The vertical axis shows the wavelet representation coefficients $c_{m,n}$

PHASE DIVERSITY REGULARIZATION

Richard A. Carreras, Sergio R. Restaino, Gordon D. Love,
Gregory L. Tarr, Janet S. Fender.

PL/LIM Imaging Technologies Division
U.S. Air Force, Phillips Laboratory
3550 Aberdeen Avenue SE
Kirtland AFB, New Mexico, 87117-5776, USA

ABSTRACT

This article derives theoretical results of the addition of a regularization term to the phase diversity algorithm. Also derived is the image estimate using the regularization term which accompanies the phase diversity algorithm. Phase diversity and its advantages are described. The phase diversity method is outlined and theoretically developed to show the reader how it is implemented using an error metric and nonlinear optimization methods. As the theoretical development is performed the image reconstruction from the phase diversity image estimate is discussed. The phase diversity image estimate is shown to be mathematically ill-posed, thus, the idea of regularization is introduced and developed further.

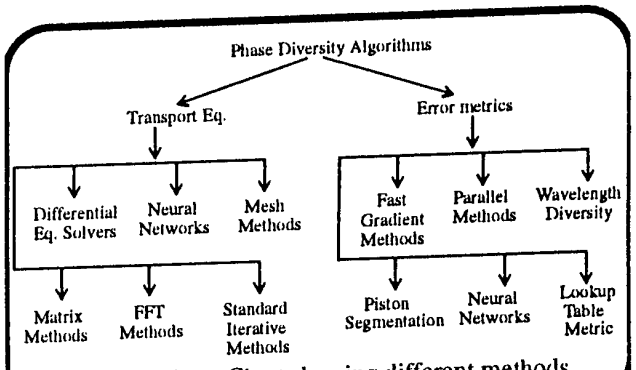
1.0 INTRODUCTION

The Air Force Phillips Laboratory has been interested in advanced concepts and devices for optical wavefront phase sensing and detection for many years. The phase diversity technique extracts the wavefront phase from two simultaneous images. Typically one image is the best focused image and the second image has a known, induced defocus aberration. From these two collected images, the optical system can be fully characterized.

Phase diversity techniques hold immense advantages over conventional optical wavefront phase detection. The first advantage is that this method is scene insensitive. That is, the images do not have to be that of a point source, but can be of any type of image. Another advantage is its simplicity in implementation. Both the in focus and the out of focus images can even be collected on the same camera if desired. Thus, this simple configuration requires less maintenance. An additional advantage is that phase diversity as a wavefront sensor can detect a piston error between two adjacent telescopes, or two adjacent segments of a single telescope. Thus, being useful in phased array telescopes, segmented telescopes and interferometer designs. Other wavefront sensors such as the Shack-Hartman or shearing interferometer wavefront sensors are not able to detect piston directly. Phase diversity also relies on an external, common reference, that of the image, which makes the techniques more robust, and less susceptible to systematic errors induced by optical hardware. Finally, this technique uses the same photons to form the image and to do aberration detection. This could be advantageous to splitting the valuable photons from the image to a separate wavefront sensor to perform the aberration detection as conventional wavefront sensors do.

Phase diversity can be divided into two distinct categories. This first category uses nonlinear optimization methods to minimize a selected error metric. This error metric utilizes the in focus image and the out of focus image in its search to find the optimal optical transfer function (OTF) which best minimizes the error metric. The second category utilizes the Transport Equation of Phase (TEP) to calculate the optical phase of the wavefront. The following chart summarizes the different methods within these two categories. The method which was chosen for study in this article was the error metric category in which the error metric is defined as the Gonsalves error metric.

Thus, phase diversity solves the problem of phase information retrieval from modulus data only, with the hypothesis that an additional modulus measurement with an induced diverse phase term suffices to solve the set of equations uniquely. In our case the induced diverse term is



Figure(1) Flow Chart showing different methods.

a defocus term, although the algorithms are not limited to focus as the diverse term. Therefore, the first image is the best focused image which can be collected on a detector plane. The second image is taken on the same or similar detector plane with a defocused image. Thus, an alternate viewpoint is that phase diversity can be seen as taking advantage of the 3-dimensional characteristics of the diffraction field within an optical system.

2.0 IMAGE RECONSTRUCTION USING GONSALVES ERROR METRIC

The framework for the development of the mathematical model requires the assumptions that the system be continuous, linear, and shift invariant. The imaging system used here is shown in Figure(2). All variables shown in Figure(2) are in the spatial domain, where $o(x,y)$ is the original object and can be seen to go in two directions. The first direction is to the in focus optical system and the second to the out of focus, diverse optical system. Now, $i(x,y)$ is the received image from the in focus optical system, similarly, $i_d(x,y)$ is the diverse image from the out of focus optical system. Next, $h(x,y)$ is the point spread function (PSF) for the in focus optical system and $h_d(x,y)$ is the point spread function of the out of focus optical system. For this particular model, the box in Figure(2) labeled $h(x,y)$ is a model that represents both the optical imaging system, and the propagation medium, under the aforementioned assumptions.

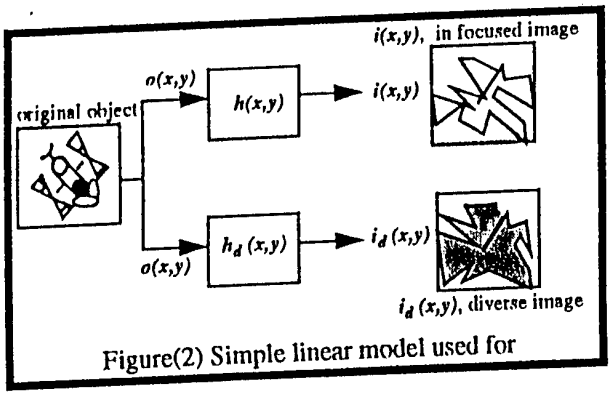
The input output relationship for the in focus optical system is the following equation[1,5].

$$i(x, y) = o(x, y) * h(x, y) \tag{1}$$

The equation in the Fourier domain is the following.

$$I(u, v) = O(u, v)H(u, v) \tag{2}$$

Where $I(u,v)$ is the image spectra; $O(u,v)$ is the object spectra; and $H(u,v)$ is the Optical Transfer Function (OTF). The OTF, is defined as the following autocorrelation.



Figure(2) Simple linear model used for

$$H(u, v) = C(u, v) \otimes C(u, v) \tag{3}$$

Where $C(u,v)$ is referred to as the complex pupil function and is defined as the following equation.

$$C(u, v) = A(u, v)e^{i\phi(u, v)} \tag{4}$$

The $A(u,v)$ is defined as the pupil function of the optical system and $\phi(u,v)$ can be expanded in a series as follows.

$$\phi(u, v) = \sum_{n=1}^{N} \alpha_n \phi_n$$

$$\phi(u, v) = \sum_{n=1}^{N} \alpha_1 + \alpha_2 \cos\theta + \alpha_3 \sin\theta + \alpha_4 R^2 + \ldots + func(R, \theta) \tag{5}$$

The ϕ_n is the basis function used in optics to describe the various aberrations and it is composed of the discretized Zernike polynomials. The Zernike coefficients are the α_n in front of each polynomial function, where α_1 is piston coefficient, α_2 and α_3 are x-tilt and y-tilt coefficients, α_4 is focus coefficients, etc. The following definitions are used for R and θ.

$$R = \sqrt{u^2 + v^2} \quad \text{and,} \quad \theta = \operatorname{atan}\left(\frac{v}{u}\right) \tag{6}$$

The following are the corresponding companion equations for the diverse system.

$$i_d(x, y) = o(x, y) * h_d(x, y) \tag{7}$$

$$I_d(u, v) = O(u, v)H_d(u, v) \tag{8}$$

$$H_d(u, v) = C_d(u, v) \otimes C_d(u, v) \tag{9}$$

Where the complex pupil function for the diverse system is the following equation.

$$C_d(u, v) = A(u, v)e^{i\{\phi(u, v) + \Delta\varepsilon(u, v)\}} \tag{10}$$

Notice that the complex pupil function for the diverse optical system has the known defocus term added to the exponential and is designated as $\Delta\varepsilon(u,v)$.

Starting with the two basic equations in the frequency domain for the in focus and diverse optical systems.

$$I(u, v) = O(u, v)H(u, v) \tag{11}$$

$$I_d(u, v) = O(u, v)H_d(u, v) \tag{12}$$

Both $I(u,v)$ and $I_d(u,v)$ are computed from the measured image data. $H(u,v)$, $H_d(u,v)$ and $O(u,v)$ are unknowns. Setting up an error metric between the measured and diverse images to get the following equation.

$$E = \sum_u \sum_v |I(u, v) - O(u, v)H(u, v)|^2 + \ldots + \sum_u \sum_v |I_d(u, v) - O(u, v)H_d(u, v)|^2 \tag{13}$$

Taking the derivative of the above error metric E with respect to the object, $O(u,v)$ and set equal to zero, then solve for $O(u,v)$. After some algebraic manipulations $O(u,v)$ is found as a function of $H(u,v)$, $H_d(u,v)$, $I(u,v)$ and $I_d(u,v)$.

$$O(u,v) = \frac{H^*(u,v)I(u,v) + H_d^*(u,v)I_d(u,v)}{|H(u,v)|^2 + |H_d(u,v)|^2} \quad (14)$$

Substituting this equation into the above error metric with appropriate algebraic manipulations yields the following error metric equation.

$$E = \sum_u \sum_v \frac{|I(u,v)\hat{H}_d(u,v) - I_d(u,v)\hat{H}(u,v)|^2}{|\hat{H}(u,v)|^2 + |\hat{H}_d(u,v)|^2} \quad (15)$$

We have replaced the OTF's, H and H_d by $\hat{H}$ and $\hat{H}_d$ to indicate that these are to be estimated values. Remarkably, this expression is independent of the original object $O(u,v)$ and was first derived by Gonsalves[2,3]. Throughout the rest of this report this equation is the only error metric used and is referred to as the Gonsalves error metric.

Given the above error metric, a possible strategy is to minimize the Gonsalves error metric, by the estimation of the coefficients of the OTF. The minimization can be accomplished by using nonlinear optimization techniques, such as conjugate gradient or simplex algorithms. The conjugate gradient is suggested as the nonlinear optimization algorithm since it is well understood, is fast, and its memory requirements are workable. In addition there are many conjugate gradient routines which are available in various scientific software packages. The conjugate gradient suggested for this research is from the IMSL libraries and uses finite difference methods to calculate the gradient.

From this estimated OTF (i.e. $\hat{H}(u,v)$) the in focus image was used in an inverse filter calculation derived from Equation(2). The calculated Fourier transform of the object is given by,

$$\hat{O}_{inv}(u,v) = \frac{I(u,v)}{\hat{H}(u,v)} . \quad (16)$$

The inverse filter suffers from being mathematically ill-posed and numerically instable. The problem arises when there are zeros in the OTF, which will cause singularities in the calculation of the object estimate, $\hat{O}_{inv}$. The mathematically ill-posed condition can be solved by the use of regularization[8]. The parametric Wiener filter is an optimized least-square filter which essentially has a regularization term in the denominator to perform the image recovery[9]. The calculated Fourier transform of the object is given using the parametric Wiener filter below.

$$\hat{O}_{wien}(u,v) = \frac{\hat{H}(u,v)^* I(u,v)}{|\hat{H}(u,v)|^2 + \beta\left[\frac{P_N(u,v)}{P_O(u,v)}\right]} \quad (17)$$

where $P_O(u,v)$ is the power spectrum of the object, $P_N(u,v)$ is the power spectrum of the noise and the β, is a user selectable parameter. In actual reconstruction of data, the researcher does not have the object from which to get the power spectrum, (the object is what we are seeking). Therefore, for many applications the Signal to Noise (SNR) of the image is used in replacement of the power spectra.

$$\hat{O}_{wien}(u,v) = \frac{\hat{H}(u,v)^* I(u,v)}{|\hat{H}(u,v)|^2 + \beta\left[\frac{1}{SNR_I(u,v)}\right]} \quad (18)$$

These two estimates of the object can be calculated on every iteration of the phase diversity calculation or can be calculated at the end of the iteration. For most applications the above estimates of the OTF are just calculated at the end when the results of the phase diversity have resulted in the most optimum estimate of the OTF. However, these calculations can be made at each iteration if there is a reason to believe that the calculation of the estimated OTF can be incorporated into the iteration of the phase diversity optimization and will result in a more favorable results. More favorable results can be either a better estimate of the OTF or faster convergence or similar criterion.

The above equations (Equation(16), (17) and (18)) are well understood and well documented. The authors choose to concentrate on Equation(14), rewritten below and referred to as the Gonsalves object estimate.

$$\hat{O}_{gons}(u,v) = \frac{H^*(u,v)I(u,v) + H_d^*(u,v)I_d(u,v)}{|H(u,v)|^2 + |H_d(u,v)|^2} \quad (19)$$

One of the first characteristics which is observed in Equation(19) is the similarity to the Wiener filter, Equation(17) and Equation(18). In fact, Equation(19) also suffers from being numerically ill-posed, the same drawback which is found in the inverse filter, Equation(16). The Wiener filter solved the ill-posed nature of the inverse problem by its addition of the extra term in the denominator. Thus, a possible solution which at times has been proposed was the addition of a similar Wiener type term in the denominator[6,7]. Thus, rewriting Equation(19).

$$\hat{O}_{gons}(u,v) = \frac{H^*(u,v)I(u,v) + H_d^*(u,v)I_d(u,v)}{|H(u,v)|^2 + |H_d(u,v)|^2 + \beta\left[\frac{P_N(u,v)}{P_O(u,v)}\right]} \quad (20)$$

However, this equation was ad-hoc and there was not any analytical derivation or mathematical justification for the addition of the extra term in the denominator. Also, there is no indication of the effect of this addition term would have on the actual error metric which is being minimized. The next section will address these shortcomings.

3.0 REGULARIZATION FOR THE GONSALVES ERROR METRIC

This section shows a mathematically tractable derivation using the method of regularization to solve for the Gonsalves error metric and the Gonsalves object estimate. The method used here is similar the Wiener filter development in the text by Andrews and Hunt[11]. Using the methods of Lagrange multipliers we can add the regularization term to the original Gonsalves error metric in Equation(13) to get the following new metric.

$$E = \sum_u \sum_v |I(u,v) - O(u,v)H(u,v)|^2 + \ldots + |I_d(u,v) - O(u,v)H_d(u,v)|^2 + \beta |QO(u,v)|^2 \quad (21)$$

Where β is the Lagrange multiplier value and Q is a weighting matrix the same dimension as the spectrum of the object. The matrix Q could be a smoothing matrix or a matrix associated with SNR. Rewriting Equation(21) and dropping the spatial frequency variables *(u,v)*.

$$E = \sum_u \sum_v (I - OH)(I - OH)^* + (I_d - OH_d)(I_d - OH_d)^* + \ldots + \beta (QO)(QO)^* \quad (22)$$

Now taking the derivative with respect to the object O.

$$\frac{dE}{dO} = \sum_u \sum_v \left[\frac{d}{dO}(I - OH)\right](I^* - [OH]^*) + \ldots + (I - OH)\left[\frac{d}{dO}(I^* - [OH]^*)\right] + \left[\frac{d}{dO}(I_d - OH_d)\right](I_d^* - [OH_d]^*) + \ldots + (I_d - OH_d)\left[\frac{d}{dO}(I_d^* - [OH_d]^*)\right] + \ldots + \beta\left[\frac{d}{dO}(QO)\right](QO^*) + \beta(QO)\left[\frac{d}{dO}(QO^*)\right] \quad (23)$$

After some algebraic manipulations the following equations result.

$$\frac{dE}{dO} = \sum_u \sum_v 2RE[H(HO - I)^* + H_d(H_dO - I_d)^* + \beta Q(QO)^*] \quad (24)$$

Taking the conjugate of the above equation.

$$\frac{dE}{dO} = \sum_u \sum_v 2RE[H^*(HO - I) + H_d^*(H_dO - I_d) + \beta Q^* QO] \quad (25)$$

Expanding the above equation.

$$\frac{dE}{dO} = \sum_u \sum_v 2RE[H^*H + H_d^*H_d + \beta Q^* Q]O - H^*I - H_d^*I_d \quad (26)$$

Now as before, the above equation is set to zero and solved for the object O.

$$O = \frac{H^*I + H_d^*I_d}{|H|^2 + |H_d|^2 + \beta Q^* Q} \quad (27)$$

As shown in Andrews and Hunt[11], the Q matrix may be defined as the power spectra shown below, thus making Equation(27) very Wiener filter like. Thus, if $Q = [P_O]^{-1/2}[P_N]^{1/2}$, implies; $Q^*Q = P_N/P_O$.

An alternate method for finding appropriate values for the Q matrix are shown in Andrews and Hunt[11]. The Q matrix may be chosen to minimize the second (or higher) difference energy of the estimated object. In this case, for the second difference, $Q = [Q_1] \otimes [Q_1]$, where $\oplus$, is a matrix multiply, thus, Q_1 is defined as the following tridiagonal matrix.

$$[Q_1] = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 & \ldots \\ 1 & -2 & 1 & 0 & 0 & \ldots \\ 0 & 1 & -2 & 1 & 0 & \ldots \\ 0 & 0 & 1 & -2 & 1 & \ldots \\ 0 & 0 & 0 & 1 & -2 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & 1 & -2 \end{bmatrix} \quad (28)$$

Such operator constraints guarantee that the object estimate does not oscillate wildly in the constraint solution by minimizing higher order differences.

Continuing with the analysis, from Equation(27) get its complex conjugate,

$$O^* = \frac{HI^* + H_dI_d^*}{|H|^2 + |H_d|^2 + \beta |Q|^2},$$

and define the denominator as D and the two numerators as the following.

$$D = |H|^2 + |H_d|^2 + \beta |Q|^2 \quad (29)$$

$$N = H^*I + H_d^*I_d \quad (30)$$

$$N^* = HI^* + H_dI_d^* \quad (31)$$

Then substitute into the regularized Gonsalves error metric, using the slightly rearranged Equation(22) shown below.

$$E = \sum_u \sum_v |I|^2 + |I_d|^2 - O^*(IH^* + I_dH_d^*) - O(I^*H - I_d^*H_d) + \ldots + |HO|^2 + |HO_d|^2 + \beta |QO|^2 \quad (32)$$

to get the following equation.

$$E = \sum_u \sum_v |I|^2 + |I_d|^2 - N^*\frac{(IH^* + I_dH_d^*)}{D} - N\frac{(I^*H - I_d^*H_d)}{D} + \ldots + |O|^2(|H|^2 + |H_d|^2 + \beta |Q|^2) \quad (33)$$

Notice that the last quantity in parenthesis is the definition of the defined denominator, D. Thus, after some algebraic manipulations we arrive at the following equation.

$$E = \sum_u \sum_v \frac{D(|I|^2 + |I_d|^2) - |I^*H - I_d^*H_d|^2}{D} \quad (34)$$

Substituting in for the D and expanding with more

algebraic manipulations to finally arrive at the regularized error metric.

$$E = \sum_u \sum_v \frac{|IH_d - I_d H|^2 + \beta |Q|^2 [|I|^2 + |I_d|^2]}{|H|^2 + |H_d|^2 + \beta |Q|^2} \qquad (35)$$

The same values which are used for Q in this error metric should also be used for the reconstruction using the phase diversity object estimate with the regularization term, Equation(27), rewritten here.

$$\hat{O}_{PD} = \frac{H^* I + H_d^* I_d}{|H|^2 + |H_d|^2 + \beta |Q|^2} \qquad (36)$$

4.0 CONCLUSIONS USING THE PHASE DIVERSITY REGULARIZATION ERROR METRIC

It can now be hypothesized that the regularized Gonsalves object estimate should be used with its accompanying regularized Gonsalves error metric to produce good results. Computer simulations should follow to verify or disprove this hypothesis. The authors challenge new users to incorporate any *a-priori* knowledge of the specific problem being addressed via the β parameter and the Q matrix to produce improved results. The regularized Gonsalves object estimate should also be compared to Wiener filter reconstructed results. We believe there should also be more concise signal to noise analysis on the phase diversity regularized filter to see if results can be improved.

5.0 REFERENCES

[1] Gaskill, J.D., Linear Systems, Fourier Transforms and Optics, John Wiley and Sons. New York

[2] Gonsalves, R. A., Phase Retrieval From Modulus Data, J. Opt. Soc. Am. Vol. 66, No. 9, September 1976

[3] Gonsalves, R. A., Chidlaw, R., Wavefront Sensing by Phase Retrieval, SPIE. Vol. 207, Applications of Digital Image Processing III,1979

[4] Gonsalves, R. A., Phase Retrieval and Diversity in Adaptive Optics, Optical Engineering, Vol. 21, No. 5, September/October, 1982

[5] Goodman, J.R., Introduction to Fourier Optics, McGraw-Hill, N.Y

[6] Paxman, R. G., Fienup, J. R., Optical Misalignment Sensing and Image Reconstruction Using Phase Diversity, J. Opt. Soc. Am. A/Vol. 5, No. 6, June 1988

[7] Paxman, R. G., Schulz, T. J.,Fienup, J. R., Joint estimation of Object and Aberrations by Using Phase Diversity, J. Opt. Soc. Am. A/Vol. 9, No. 7, July 1992

[8] Craig, I. J. D., Brown, J. C., Inverse Problems in Astronomy, Adam Hilger Ltd, 1986, Bristol, England

[9] Gonzalez, R.C., Wintz, P., Digital Image Processing, Addison-Wesley, 1987,

[10] Andrews, H.C., Hunt, B.,R. Digital Image Restoration, Prentice-Hall, N.Y, 1977

AN APPROACH TO A SPEECH DETECTION SYSTEM BY MEANS OF HIGHER-ORDER SPECTRA

[1]Juan L. Navarro-Mesa, [2]Asunción Moreno-Bilbao & [2]Antonio Bonilla-Aguilera

[1]Escola Universitària Politècnica de Mataró. Universitat Politècnica de Catalunya.
Avda. Puig i Cadafalch, 101-111. 08303 Mataró. Barcelona
[2]Dept. of Signal Theory and Communications. Universitat Politècnica de Catalunya.
C./ Gran Capitán, s/n. Campus Nord UPC. Módulo D5. 08034 Barcelona. Spain
Tel.: 343 4016454 Fax.: 343 4016447 e-mail: navarro@gps.tsc.upc.es

ABSTRACT

The properties of Higher-Order Statistics (HOS) open a new sight in the problem of signal detection and speech presence detection specifically. In essence, both the frequency wealth of speech and the ability of HOS to suppress additive symmetrically distributed noises as well as discern and extract information about deviations of Gaussianity and non-stationarities is exploited. Theoretical and experimental results have led us to two functions. One is obtained from the principal domain of the bispectrum, and the other one is the integrated polyspectrum. In this paper we propose how to do a proper use of these functions and apply to a simple speech detection system.

1. PROBLEM STATEMENT

Methods for the automatic detection of the beginning and ending points of an utterance are required in many speech processing applications, either for isolated or continuos speech. The most classic methods for speech presence detection use data frames from which estimate and extract some kind of feature. The most used features are short time energy and its variations, zero crossings, combinations of autocorrelation lags, etc. [7]. Once the information is extracted it can be applied to a detection system. When there is not present any disturbance at the signal, the detection scores are quite good. However, when the signal is embedded in additive noise and the SNR is low, poor results arise due to the impossibility to discern speech from noise. This is specially the case with sounds like plosives and fricatives since they are very often low energy and noise affects them strongly. Therefore, in a frame-based method, it is difficult to discriminate them from noise.

2. SPEECH DETECTION BY MEANS OF HIGHER-ORDER SPECTRA

On a particular level, an important and attractive property of the HOS is that the HOS of two independent random processes equals the sum of the individual ones. A key characteristic of the HOS, from the detection point of view, is their ability to suppress any kind of Gaussian process. Practically speaking, this means that when HOS-based methods are applied to detect non-Gaussian signals corrupted by additive Gaussian noise they automatically improve the results at a given signal-to-noise ratio compared against classical autocorrelation based methods. In particular , odd-order HOS (e.g., third-order spectra or bispectra) suppress any kind of symmetrically distributed process. Moreover, HOS have the ability of detecting non-stationarities.

Many real world processes are non-symmetrically distributed, and measurement noise can often be realistically described as a stationary symmetrically distributed (e.g., Gaussian) process. Furthermore, a process of interest in a stationary noise background becomes a non-stationarity which can be detected by means of HOS-based methods. Due to the wealth of speech in the frequency dimension, a method that explores the signal in the frequency dimension should lead to improved detection results over classic ones. Moreover, since speech-silence transitions carry non-stationarities, a representation specifically designed to detect non-stationarities should improve results even more. Therefore, it is possible to go a step further in the way we extract information from speech by explicitly exploiting its polyfrequency content.

In this paper we do this by means of three bispectral-based functions. Firstly, the Integrated Polyspectrum (IP) [8] has been proposed for detecting an unknown, random, stationary, non-Gaussian signal in Gaussian noise. The IP can be seen as the integration in one frequency dimension of the polyfrequencies. It shares with HOS all their general properties. Also, the IP estimators are robust and consistent, and computationally efficient. We use the integrated bispectrum for speech detection. And secondly, we can also use the bispectrum by exploiting its ability of discerning and extracting information about deviations from Gaussianity and non-stationarities. This is done using two regions in the principal domain, the Inner Triangle (IT) and the Outer Triangle (OT), respectively. The basic concepts related with the use of HOS for speech detection are as follows.

2.1. GAUSSIANITY AND NON-STATIONARITY TESTS

Be s(n) a discrete-time real, zero mean, stationary and non-Gaussian random process. Its third-order cumulant and Fourier transform, the bispectrum, are [4],

$$C_{3s}(j,k) = E\{s(n)s(n+j)s(n+k)\} \quad (1)$$

$$B(f,g) = \sum_j \sum_k C_{3s} e^{-i2\pi(fj+gk)} \quad (2)$$

It is necessary a consistent estimator of B(f,g). There are two methods, indirect and direct. We use the direct one for

which a signal frame is divided in K records of M points, computed the individual bispectrum and averaged for the K estimates.
Paying attention in the principal domain of B(f,g) we can differentiate between two regions [1,2,3]. One is the Inner Triangle where for continuos-time, stationary, non-Gaussian and unaliased processes the bispectrum is non-vanishing. The other is the Outer Triangle where the bispectrum will usually be nonzero when the process is either non-stationary or aliased.

In [3] the authors study the ability of bispectrum for detecting non-Gaussian signals masked by either Gaussian or non-Gaussian stationary noise. They propose a detection test using the bicoherence function evaluated in the IT region. In this function the noise effect is mitigated by extracting its bispectrum from the signal bispectrum. Signal presence and transitions can also be detected by testing changes of stationarity [1]. When there is silence alone or stationary noise is present the expected value of B(f,g) in OT is zero even for non-Gaussian noise. The importance of restricting the attention to the OT triangle is that we can detect the presence of non-stationarities in a stationary noise. This is just the situation when there is a (noisy) silence/signal transition.

2.2. SIGNAL DETECTION USING THE IP

In parallel with the development of the work presented in [1,2,3] there has appeared another bispectral (polyspectral in general) function of potential application to signal [8] and speech detection [5]. This is the integrated bispectrum which is defined in the following way. Be $T_b(w_m)$ a consistent estimator of the integrated bispectrum obtained by averaging over K individuals estimations from non-overlapping records. Then the detection function T_b is defined as;

$$T_b = \sum_{\forall w_m} |T_b(w_m)|^2 \qquad (3)$$

where μ is the noise variance. T_b is related with the detection functions from IT and OT in the sense that it integrates both information at once.

Our previous work over real speech embedded in synthetic noise [5,6] demonstrate that the functions mentioned in this section are suitable for speech detection. However, it is necessary to study the best way to use them with real noises and validate the initial results. In the next sections we present our progresses in this direction by applying the detection functions to a threshold-based detection system.

3. SPEECH DETECTION SYSTEM

Since the information about non-stationarities and deviations from Gaussianity can be used separately we obtain two detection functions [1,3] as a summation of the squared module of the bispectrum at each bifrequency in the OT and IT, let's call F_IT (associated to speech presence) and F_OT (associated to speech presence and non-stationarities), respectively.

$$F\_IT = \sum_{f,g \in IT} \sum \left| \hat{B}(f,g) \right|^2 \qquad (4)$$

$$F\_OT = \sum_{f,g \in OT} \sum \left| \hat{B}(f,g) \right|^2 \qquad (5)$$

We have experimentally found that both functions should be jointly used if we want to obtain the best detection scores. In [5,6] we proposed a quotient F = F_OT / F_IT since it seemed the best choice in noisy environments. However, as SNR increases it is very difficult to apply a threshold because the quotient becomes greater during silences than at the speech parts and there appear too many false detection. The best way to use F_OT and F_IT is by means of the difference F_IOT = F_IT - F_OT. Ideally, this detection function is zero during noisy silences if the noise is symmetrically distributed and stationary thanks to the statistical properties of the bispectrum. In practice, it has very low and very smooth levels because F_OT eliminates the effect of very short local non-stationarities avoiding false detection. At the instants of speech/silence transitions there is a mismatch in the time-frequency content of speech associated to both a non-stationarity and speech presence. This is shown as an abrupt change in the function which is easily detectable with a threshold. In consequence, after applying a proper threshold to this function, speech presence and transitions detection becomes reliable and precise.

Of course, one can wonder what happen for more realistic noises. That is, if the noise is not symmetrically distributed and/or non-stationary, or simply unknown, is F_IOT suitable too ?. The answer is not obvious because non-stationary noises produce the worst cases we can meet. At a first glance, from the properties of HOS, it is expected to be affirmative since F_OT is sensitive to the non-stationarities and smoothes F_OIT. In practice, this expectation is nearly correct due to inability of F_OIT. For instance, short-time strong noise peaks may cause false alarms. Experiments in the next section show that even in this case F_OIT is more reliable and accurate than the energy.

The (poly)frequency dimension suppose an additional degree of freedom from which we can take benefits. For instance, if we assume stationary noise for long intervals (which is the case in many realistic situations) it is possible to perform a polyspectral subtraction with the noise polyspectrum obtained during silences, e.g., before the beginning of an utterance. Thus, noise effect could be reduced even if it is not a symmetrically distributed processes. Our experiments show that bispectral subtraction for F_OIT is computationally expensive and degrades detection when the noise is not stationary. However, the behavior of T_b is reinforced by noise subtraction because false alarms are strongly reduced.

$$T_b = \sum_{\forall w_m} |T_b(w_m) - T_n(w_m)|^2 \qquad (6)$$

We use a simple speech detection scheme where detection is on/off when the detection function is over/under a given threshold. Information is extracted from overlapped frames of signal. For comparison purposes, the reference detection function is the one based on energy and we study the possibility of replacing the energy-based (En) function by the HOS-based one. Other than energy functions could be used, e.g., zero-crossing rate, autocorrelation lags, etc. However, in noisy environments they impair detection. Therefore, energy seems to be good for comparisons.

4. EXPERIMENTS AND RESULTS

The experiments have been made with a data base of 1083 English isolated digits sampled at 8 KHz. In this database there are several examples of fricative and occlusive sounds such as, /s/, /f/, /t/, /th/, etc. The noises cover a wide range of cases. These are, synthetic stationary Gaussian and exponential, internal and external telephone line, air conditioning, car engine at 2000 and 3000 rpm, keyboard and fan. All noise realizations are independent each other.

Before showing and comment results , some aspects must be pointed out. Firstly, thresholds are adaptive computed and depend on the mean and variance of the detection function in some preceding estimates (about 10 is enough). Adaptation stops when the detection is on and starts again when a new silence is detected. Secondly, the analysis frames are 37´5 ms. (300 samples) long and the time shift between frames is 6´25 ms (50 samples) each time. These values are a compromise between acceptable time precision, good use of the functions properties and computation time. And thirdly, there is an indetermination of the detection time inside the frame. The analytic solution to this problem appears quite complex. We have chosen to obtain an heuristic solution which states that detection is acceptable when the signal inside the frame (sliding from left to right) fills up at least a third of its length for F_OIT and T_b and a half for the energy function.

Comparisons between energy-based and HOS-based functions have been made taking into account precision and reliability. For this purpose we have distinguished between beginning and ending points. The graphics show, for different SNR (from to 0 dB), the average of results over all noises. In them, the energy, integrated bispectrum and F_OIT based detection functions are represented by '--', '..', and '.-', respectively.

Reliability is measured by means of the number of lost beginning (graphic 3) and ending (graphic 4) points. We consider that a beginning or an ending point is lost if the error is greater than three frames (800 samples). As we can see, for all SNR, the HOS-based methods perform better than the energy-based. The tendency of the losses is to increase when the SNR decreases. However, for T_b and F_OIT when the SNR is between 20 and 10 dB the losses tend to decrease. This is because for clean speech some human noises (e.g., breaths, clicks) in the neighborhood of the endpoints cause false detection. When the SNR is lower the noise obscure these human noises and the detection functions suppress their effect. For SNR lower than 10 dB noise is the main cause of losses.

From the accuracy point of view we have distinguished between detection of silence as speech (+) and detection inside speech (-) or speech segment loss. In graphic 1 we can see the scores for the beginning points. The tendency of the HOS-based functions, specially F_OIT, is more conservative. We mean that the HOS-based functions tend to detect beginning points during the silence before the begging. In general, all detection functions have a similar accuracy (about one frame). The energy-based function is a more accurate than the others. However, since the mean error is computed for the detected beginnings, if we take into account this fact, the HOS-based functions appear to be a better compromise between reliability and accuracy. Specially if our strategy were conservative. In graphic 2 we show the scores for the ending points. The comments about the ending point detection are similar to the ones for beginning points.

5. CONCLUSIONS

The main objective of the present work is to study the possibility of substituting the energy-based detection function by one based on bispectral measures. The results obtained from the experiments show that this possibility is very acceptable. Some aspects are still to be studied. For instance, the performance for non-symmetrically distributed and, specially, for non-stationary noises. These cases appear to be the more complicated for any kind of detection function. However, the HOS-based seem to perform quite well.

6. ACKNOWLEDGEMENT

This work has been granted by the Spanish Ministry of Education and Science (MEC) TIC 95-1022-C05-03.

7. REFERENCES

[1] M. J. Hinich . " Detecting a Transient Signal by Bispectral Analysis ". IEEE Trans. on ASSP and Signal Processing, vol. 38,No. 7,July 1990.

[2] M. J. Hinich, Hagit Messer. "On the Principal Domain of the Discrete Bispectrum of a Stationary Signal". IEEE Trans. on SP, vol. 43, n° 9, pp 2130-34, September 1995.

[3] M. J. Hinich and G. R. Wilson. " Detection of Non-Gaussian Signals in Non-Gaussian Noise Using the Bispectrum ". IEEE Trans. on ASSP and Signal Processing, vol. 38, No. 7, July 1990.

[4] J. M. Mendel. "Tutorial on Higher-Order Statistics (Spectra) in Signal Processing and System Theory:

Theoretical Results and Some Applications". Proc. of the IEEE, vol. 79, n° 3, March 1991.
[5] J. L. Navarro-Mesa & A. Moreno. "Skewness and Nonstationarity Measures Applied to Reliable Speech Endpoint Detection". Proc. Eurospeech, vol. 1, pp 1423-1426, Madrid, Spain, 1995.
[6] J. L. Navarro-Mesa & A. Moreno, E. Lleida. "Bispectral-based Statistics Applied to Speech Endpoint Detection". Proc. of the IEEE Signal Processing ATHOS Workshop on Higher-Order Statistics, pp 280-283, Spain, 1995.
[7] L.R.Rabiner & R.W.Schafer. "Digital Processing of Speech Signals".Prentice-Hall,1978.
[8] J. K. Tugnait. " Detection of Non-Gaussian Signals Using Integrated Polyspectrum ". IEEE Trans. on Signal Processing. vol. 42, No. 11. November 1994.

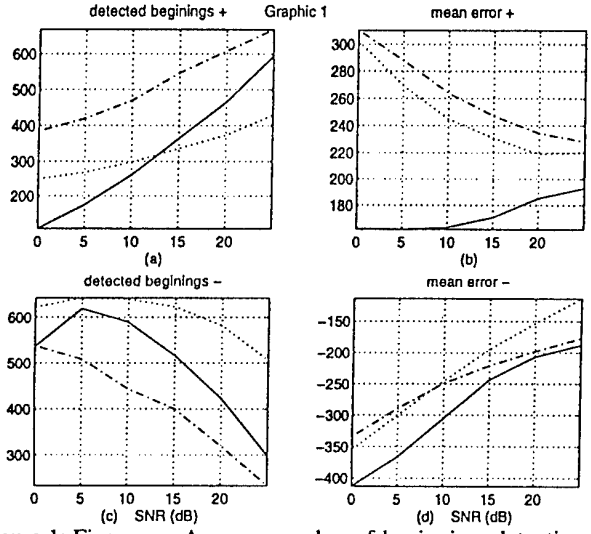


Figure 1: Fine error. Average number of beginning detection (error < 100 ms) in terms of the SNR. Energy (--), Integrated Bispectrum (..) and F_OIT (.-).

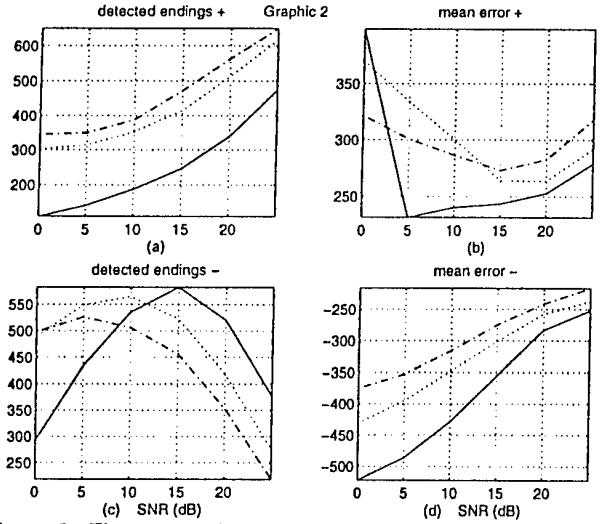


Figure 2: Fine error. Average number of ending detection (error < 100 ms) in terms of the SNR. Energy (--), Integrated Bispectrum (..) and F_OIT (.-).

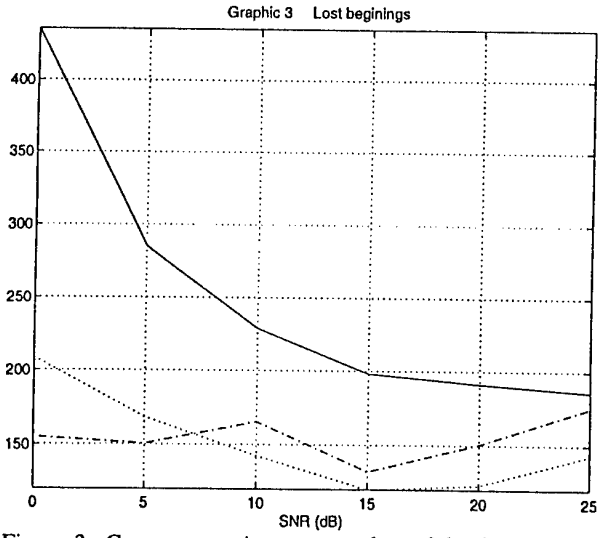


Figure 3: Gross error. Average number of beginning loss (error > 100 ms) in terms of the SNR. Energy (--), Integrated Bispectrum (..) and F_OIT (.-).

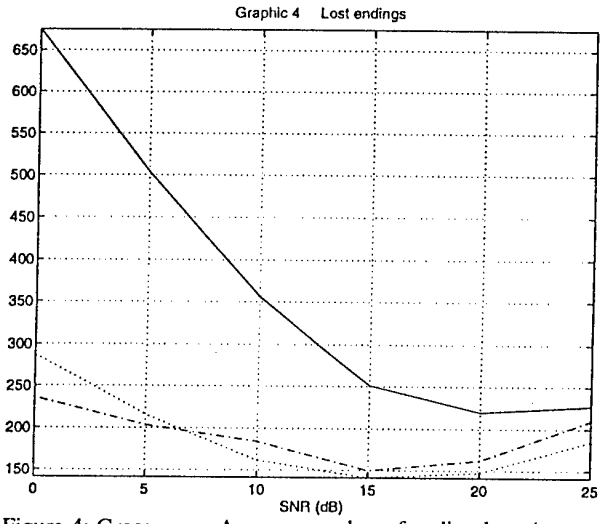


Figure 4: Gross error. Average number of ending loss (error > 100 ms) in terms of the SNR. Energy (--), Integrated Bispectrum (..) and F_OIT (.-).

Recognition of Information Symbols Using Modified Rapid Transform

*J. Turán - *K. Fazekas - L. Kövesi - M. Kövesi*

Department of Radioelectronics
Technical University of Košice
Park Komenského 13
04021 Košice
Slovakia
Tel./Fax: +42 95 6335692
E-mail: J.TURAN@CCSUN.TUKE.SK

**Department of Microwave Telecommunications*
Technical University of Budapest
Goldmann Tér 3
1111 Budapest
Hungary
Tel./Fax: +36 12 043289
E-mail: T-FAZEKAS@NOV.MHT.BME.HU

Abstract

Application of fast translation invariant modified rapid transform (MRT) in feature extraction stage of Information Symbols recognition system are described. Experimental results are given of applying the proposed recognition system to recognition Airport Passenger Orientation Symbols and Meteorological Symbols, including the dependence of recognition efficiency on the number of selected features and noise.

Keywords

Rapid Transform, Modified Rapid Transform, Pattern Recognition, Feature Selection, Invariant Features, Information Symbol Classification

1. Introduction

Transformation methods can be used to obtain alternative description of signals. These alternative descriptions have many uses such as classification redundancy reduction, coding, etc., because some of these tasks can be better performed in the transform domain [1].

Various transformations have been suggested as a solution of the problem of high dimensionality of the feature vector and long computation time. More recently the modified rapid transform (MRT) [2] was presented to break undesired invariances of the rapid transform (RT)[3].

In the paper, a new method of recognition Information Symbols using MRT will be presented. We apply the MRT in feature extraction stage of Information Symbols recognition process. Some properties of the RT and MRT will be first reviewed, then the new method of recognition of Information Symbols will be presented. Finally, the experimental results will be given in applying of the proposed pattern recognition method to recognition of Airport Passenger Orientation Symbols and Meteorological Symbols, including dependence of recognition efficiency on number of selected features and noise.

2. Modified rapid transform

Transforms which do not change with cyclic shifts in the sequence are called translation invariant. Fast translation invariant transforms are valuable tool for pure shape-specific feature extraction in pattern recognition problems. The transforms may be used to extract features of one- or two-dimensional patterns, which are invariant under cyclic permutations to characterize objects independent of their position. In the field of pattern recognition and also scene analysis is well known the class of fast translation invariant transforms - certain transforms (CT) [4] based on the original rapid transform (RT) [3] but with choosing of other pairs of simple commutative operators. The RT results from a minor modification of the Walsh-Hadamard transform (WHT). The signal flow graph for the RT is identical to that of the WHT, except that the absolute value of the output of each stage of the iteration is taken before feeding it to the next stage. This is not an orthogonal transform, as no

inverse exists. With the help of additional data, however, the signal can be recovered from the transform sequence, i.e. inverse rapid transform can be defined [5]. RT has some interesting properties such as invariance to cyclic shift, reflection of the data sequence, and the slight rotation of a two-dimensional pattern. It is applicable to both binary and analogue inputs and it can be extended to multiple dimensions. More recently was introduced the modified rapid transform (MRT) [2] which can distinguish many more patterns from one another that the original RT can. The MRT was presented to break undesired invariances of the RT which leads to a loss of information about the original pattern. This is achieved, by combining the RT with preprocessing steps using a asymmetric neighbor operator α. This operator is used to break undesirable invariances but keep the shift invariance of the MRT. Using the symbolic notation we can introduce MRT as follows:

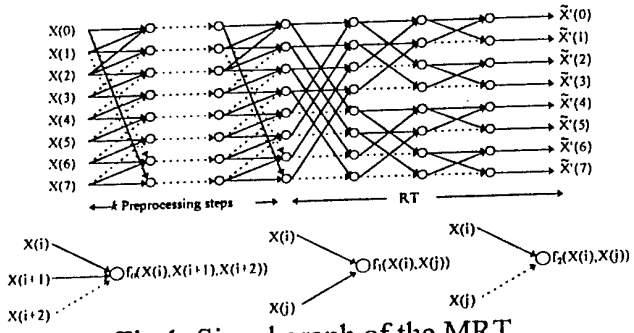


Fig.1. Signal graph of the MRT

Signal graph of MRT (Fig.1) results from signal graph of RT with adding in general k preprocessing steps $\mathbf{x}'=\alpha\mathbf{x}$. This maps the element $x(i)$ of input vector $\mathbf{x}$ to element $x'(i)$ of vector $\mathbf{x}'$ by working on the elements $x(i)$, $x(i+1)$ and $x(i+2)$

$$x'(i)=f_0(x(i),x(i+1),x(i+2)) \qquad (1)$$

It is important that the operator f_0 be asymmetric because we want to destroy the invariance of RT under reflection. Operator f_0 may be realized in the following simple manner

$$x'(i)=f_0(x(i),x(i+1),x(i+2))=x(i)+|x(i+1)-x(i+2)| \qquad (2)$$

The transform process of MRT (Fig.1) - identical to the transform process of RT requires $N=2^n$ input pixels, where n is a positive integer. Each column of the transform process in Fig.1 corresponds to a particular computational step; n steps are required. In general the variables $x^{(r)}$ in any column (r) are calculated from variables $x^{(r-1)}$ in the preceding column $(r-1)$ by

$$x^{(r)}(i+2js)=f_1(x^{(r-1)}(i+2js),x^{(r-1)}(i+(2j+1)s))$$
$$x^{(r)}(i+(2j+1)s)=f_2(x^{(r-1)}(i+2js),x^{(r-1)}(i+(2j+1)s)) \qquad (3)$$

where operators f_1, f_2 for MRT (or RT) are

$$f_1(a,b)=a+b;$$
$$f_2(a,b)=|a-b| \qquad (4)$$

and $s=2^{n-r}$; $t=2^{r-1}$; $i=0,...,s-1$; $j=1,..,t-1$ and $x \equiv x^{(0)}$ are input data (pixels) and $\mathbf{x}^{(n)} \equiv \tilde{\mathbf{x}} = MRT\{\mathbf{x}\}$ are spectral coefficients of MRT.

MRT can be applied in all areas where the RT (or any transform from class CT) can be used. Some undesired invariances of RT can be destroyed applying only one preprocessing step. Experiments from use of MRT [2,6] in character recognition showed, that MRT can distinguish many more patterns from one another than the RT or the Fourier power spectrum.

3. The Information Symbols Recognition System Model

The recognition system is simulated in digital computer using program package CT-CAD [7]. It contains the following sub-systems(Fig.2):

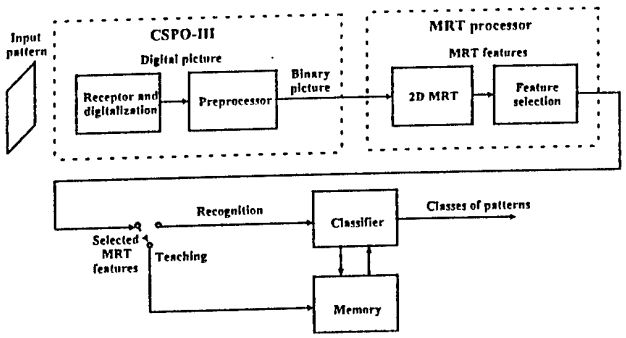


Fig.2. The MRT recognition system

1. Original digital picture preprocessing system CSPO-III was used to accepts the physical input picture and then transduce it into a measurable matrix. CSPO-III divides a visual pattern into small elements and after suitable preprocessing produces an NxN matrix over the binary field; the

element becomes 1 or 0 depending upon whether it is black or white.

2. The MRT processor according to its function may be also called a feature extractor. A 2D MRT of all binary prototypes is taken in this stage. Than feature selection is carried out in the MRT "spectral" domain on various basis (maximum value of spectral coefficients, variance zonal sampling and interclass standard deviation).

3. The selected MRT features of binary pictures (symbols) are in the teaching process feeded into the memory. Thus the memory unit learn the a priori knowledge of each class before the system can be used to make any decision. In the recognition process the selected MRT features are feeded into the classifier, which discriminates each pattern (symbol) and assigns a category (a class) to it by some decision rule. We use a simple classifier based on cross responses d_{kl} between two different patterns from class k and l, defined in the next section.

4. *Recognition of Information Symbols*

The proposed Information Symbols recognition system was tested on the two classes of selected symbols:

1. Airport Passenger Orientation Symbols (class consist of M=11 independent symbols) (Fig. 3).

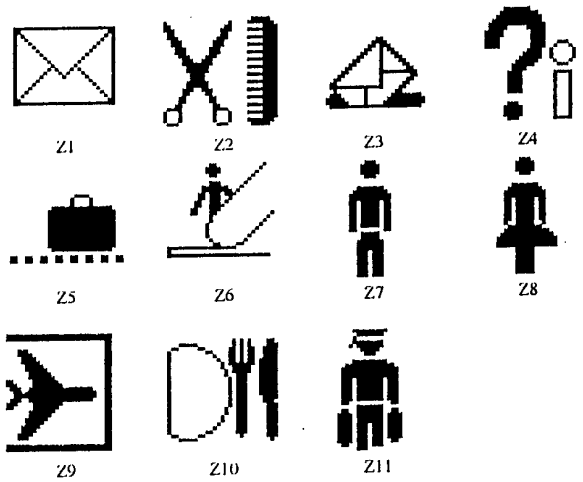


Fig.3. The Airport Passenger Symbols

2. Meteorological Symbols (class consist of M=16 independent symbols) (Fig. 4).

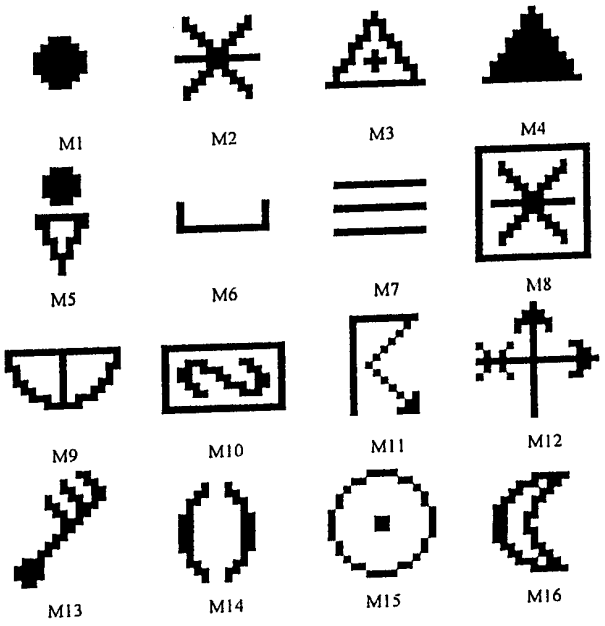


Fig.3. The Meteorological Symbols

We implemented feature extraction with MRT at the both sets of Information Symbols. In general, the efficiency of feature extraction can be assessed by the system confusion matrix $D=\{d_{kl};\ k, l=1,...,M\}$ where d_{kl} are cross responses (or the distances between any two different symbols k, l in the feature space) and M is the number of classes or number of different symbols. The confusion matrix can be calculated in two steps shown as follows:

A. All M prototypes of Information Symbols, each represented by a binary NxN matrix $(x_k(i,j)$, with $i, j=1,...N;\ k=1,...,M$ and M=11 or M=16) are transformed to the MRT transform domain

$$\tilde{x}_k(i,j) = \tau\{x(i,j)\} \qquad (5)$$

where $\tau \equiv MRT$.

B. The cross response $d^{(1)}_{kl}$ between two different symbols from class k and l is defined as follows:

$$d^{(1)}_{kl} = \sum_{i,j=1}^{N} |\tilde{x}_k(i,j) - \tilde{x}_l(i,j)| \qquad (6)$$

The results of experiments of dependence of recognition efficiency on number of selected features and influence of noise are shown on Tab.1 and Tab.2. A set of 165 symbols were used for testing and teaching purposes, testing set used on Tab.1 contains 5 noised symbols for each Airport Passenger Orientation Symbol. A set of

Tab.1 Recognition of Airport Passenger Orientation symbols (class consist of M=11 independent symbols)

Noise	Noised teach.set	Number of coefficients	Recogn. Effic.	Comment
0%	No	1 (0.098%)	100%	
1%	No	10 (0.98%)	100%	
1%	Yes	8 (0.78%)	100%	
2%	No	≥40 (≥18.75%)	95%	Z6->Z4...3x, Z7->Z1...1x
2%	Yes	8 (0.78%)	98.18%	Z8->Z7...1x
	Yes	16 (1.56%)	100%	
3%	Yes	8 (0.78%)	100%	
4%	Yes	18 (1.76%)	100%	

240 symbols were used for testing and teaching purposes, testing set used on Tab.2 contains 5 noised symbols for each Meteorological Symbol.

Tab.2 Recognition of Meteorological symbols (class consist of M=16 independent symbols)

Noise	Noised teach. Set	Number of coefficients	Recogn. Effic.	Comment
0%	No	1 (0.4%)	100%	
1%	No	10 (4.0%)	96.25%	M13 ->M3...1x
1%	No	13 (5.0%)	100%	
1%	Yes	8 (3.125%)	100%	
2%	No	≥48 (≥18.75%)	95%	M6->M4...3x, M7->M1...1x
2%	Yes	8 (3.125%)	100%	
3%	Yes	≥40(≥15.625%)	97.5%	M3->M10...1x, M14->M2...1x

The results of both experiments may be summarized as follows:

A. Only one preprocessing step in MRT signal graph is sufficient to destroy the undesired invariances and improve significantly capability of MRT distinguish many more patterns from one another than the original RT can.

B. Even if a very simple classifier was used, the recognition efficiency 97%-100% can be obtained with selecting only a couple of features (0.4%-5% of the number of MRT coefficients) in the MRT spectral domain, even if the symbols are corrupted by (1%-3%) noise.

5. Conclusion

We apply the MRT in feature extraction stage of Information Symbols recognition system. Experiments with recognition of two classes of symbols (Airport Passenger Orientation Symbols and Meteorological Symbols) demonstrate that even if very simple classifier was used, the very high recognition efficiency can be obtained with selecting only a couple of features in the MRT spectral domain, even if the symbols are corrupted by noise.

References

[1] Chmúrny, J. - Turán, J.: Two-dimensional Fast Translation Invariant Transforms and Their Use in Robotics. Electronic Horizon, Vol.15, No.5, 1984, 211-220.

[2] Fang, M. - Häusler, G.: Modified Rapid Transform. Applied Optics, Vol.28, No.6, 1989, 1257-1262.

[3] Reitboeck, H. - Brody, T.P.: A Transformation with Invariance Under Cyclic Permutation for Application in Pattern Recognition. Inf. and Control, Vol.15, 1969, 130-154.

[4] Wagh, M. D. - Kanetkar, S.V.: A Class of Translation Invariant Transforms. IEEE Trans. on Acoustic, Speech and Signal Proc., Vol. ASSP-25, No.3, 1977, 203-205.

[5] Turán, J. - Chmúrny, J.: Two-dimensional Inverse Rapid Transform. Computers and Art. Intelligence, Vol.2, No.5, 1983, 473-477.

[6] Turán, J.: Recognition of Printed Berber Characters Using Modified Rapid Transform. Journal on Communications, Vol.XLV, 1994, 24-27.

[7] Turán, J. - Kövesi, L. - Kövesi, M.: CAD System for Pattern Recognition and DSP with Use of Fast Translation Invariant Transform. Journal on Communications, Vol.XLV, 1994, 85-89.

A EXPONENTIAL OPEN HASHING FUNCTION BASED ON DYNAMICAL SYSTEMS THEORY

Bradley J Smith Gregory Heileman Chaouki Abdallah

Department of Electrical & Computer Engineering
University of New Mexico
Albuquerque, NM 87131

ABSTRACT

In this paper an efficient open hashing function is developed using a combination of dynamic systems analysis and number theory. The new hash function appears to nearly match the optimal double divide hash function for uniform data distributions, and performs significantly better for clustered data distributions. A higher integer Lyapunov exponent for initial data probes is indicative of this improved cluster hashing behavior. The number of mathematical operations per probe in the new hash function matches that of double division hashing.

1. INTRODUCTION

Hash functions are ubiquitous in the field of Computer Science. They are widely used to gain rapid access to databases, operating systems, compilers, and a ranges of business and scientific applications. Despite the importance of these functions, only a primitive theoretical understanding of what makes a good hash function exists. This research began with the premise that hash functions might better be analyzed using measures from the study of non-linear, chaotic systems. The results presented indicate that it is possible to build an efficient open hash function that performs better on clustered data than the commonly used double divide hash function.

A hash table is a well-known data structure used to maintain *dynamic dictionaries*. A dynamic dictionary is used to manage a collection of data items (each of which has a unique key value) that can be accessed according to the following operations:

1. *Search*(k, S). Returns the data item with key k in dynamic dictionary S.
2. *Insert*(x, S). Adds data item x to dynamic dictionary S.
3. *Delete*(k, S). Removes the data item with key k from dynamic dictionary S.

The hash table data structure consists of an array T whose N slots are used to store the collection of data items. When implementing the above operations, an index is computed from the key value using an *ordinary hash function* h, which performs the mapping

$$h : U \rightarrow \{0, 1, \ldots, N-1\}$$

where U denotes the set of all possible key values (i.e., the universe of keys). Thus, $h(k_i)$ denotes the index, or *hash value*, computed by h when it is supplied with key $k_i \in U$. Furthermore, one says that k_i *hashes* to slot $T[h(k_i)]$ in hash table T.

Since $|U|$ is generally much larger than N, h is unlikely to be a one-to-one mapping. In other words, it is very probable that for two keys k_i and k_j, where $i \neq j$, $h(k_i) = h(k_j)$. This situation, where two different keys hash to the same slot, is referred to as a *collision*. Since two items cannot be stored at the same slot in a hash table, the *Insert* operation must resolve collisions by relocating an item in such a way that it can be found by subsequent *Search* and *Delete* operations.

One method of resolving collisions, termed *open addressing* by Peterson [5], involves computing a sequence of hash slots rather than a single hash value. This sequence is successively examined, or *probed*, until an empty hash table slot is found in the case of an *Insert* operation, or the desired item is found in the case of *Search* or *Delete* operations.

Typically, in open addressing, the ordinary hash function discussed above is modified so that it uses both a key, as well as a probe number when computing a hash value. This additional information is used to construct the probe sequence. That is, in open addressing, hash functions perform the mapping

$$h : U \times \{0, 1, \ldots, N-1\} \rightarrow \{0, 1, \ldots, N-1\}$$

and produce the *probe sequence* $< h_0(k), h_1(k), h_2(k), \ldots >$. Because the hash table contains N slots, there can be

at most N unique elements in a probe sequence. A *full length probe sequence* is defined to be a probe sequence $< H(k,1), H(k,2) \ldots H(K,N) >$ which visits all N table entries after only N probes.

A general form for dynamical systems is given by the first order recurrence relation

$$x_{n+1} = f(x_n) \quad x_0 = c \tag{1}$$

where the constant c is the *initial condition*, and $f : \Re \rightarrow \Re$. The function f generally must be non-linear to generate complex behavior. This simple system is called an *iterator*. It is well-known that for some choices of even simple f in equation (1), a system that exhibits extremely complex behavior can be obtained. One such form of behavior is referred to as *chaos*. While a universally accepted definition of chaos does not exists, it is generally agreed that one characteristic is sensitive dependence on initial conditions, coupled with bounded behavior [4]. Qualitatively, an iterator is said to be sensitive to initial conditions if the orbits that result from two initial conditions, which are arbitrarily close, are distinctly different. The technique most often used to detect this type of behavior involves computing the *Lyapunov exponent* of system (1), which will be defined in section 3.

2. OPEN HASHING

Open hashing is an insertion strategy for resolution of data collisions based on probing of hash table entries until an empty table slot is found. The hash function $H(k,i)$ is used to denote a probing hash function where k is the key associated with the data being inserted and i is the probe index. Knuth [3] notes that the desirable properties of an open hash function include:

- Efficient hash function evaluation time.
- A long probe sequence to accommodate tables near capacity.
- Different probe sequences for each data item to avoid primary and secondary clustering.
- Even data distribution over the entire table size for both initial and subsequent probes. This property is widely known as the *uniform hashing* property [1].

3. CHAOTIC MEASURES AND DYNAMICAL SYSTEMS

The assertion that hash functions and chaotic iterators share some of the same desired properties was first put forth in [2]. The authors suggest that a chaotic iterator which exhibits sensitive dependence on initial conditions might also perform well as a hash function. The authors introduce the notion that hash functions can be transformed into chaotic iterators in the real domain, allowing some measures from the field of non-linear dynamics to be applied. This was done by converting the hash functions to iterators in the continuous domain, and then applying the continuous Lyapunov exponent to the resulting iterator [4]. The results showed that the corresponding double hashing iterator had a positive Lyapunov exponent in the real domain, indicating that this iterator has sensitive dependence on initial conditions. Similar tests for linear hashing indicated that it had a zero Lyapunov exponent, or no sensitive dependence on initial conditions.

Additional work done in the integer domain by the authors indicates that measurement of the Lyapunov exponent, modified slightly for the integer domain, does provide an indicator of distance between iterations, and therefore provides a useful measure. The details of this measurement and analysis are too long to include in this abbreviated paper. The relevance of this Lyapunov measure to open hash functions, is that it provides a measure of the ability of the open hashing function to quickly distribute data during successive probes. Further, our analysis shows that the most commonly used open hash function, double divide hashing, actually has a low Lyapunov exponent as measured over the first few probe sequences. This indicates that the double divide hash function tends to place clustered data close together during successive probes, leading to poor performance for some clustered data configurations. This key result was then used to develop our exponential hash function.

4. AN EXPONENTIAL HASHING FUNCTION

The Lyapunov measurements above led to the development of a better hash function for use on clustered data. Two alternatives exist, either choose non-linear functions modulo N for $h_1(k)$ and $h_2(k)$ or create a non-linear modulo N probe function. The problem with the first approach is that the hash functions used in double hashing must be quickly evaluated, yet must also preserve uniform distribution of the hashed data in the table space, both described in section 2. It is difficult to create a non-linear modulo N function that meets all three criterion. A good choice would appear to be $h(k) = k^m \bmod N$ where the exponent m is chosen to be relatively prime to $N-1$, and N is prime. This function, similar to that used in public key encryption,

appears to be a good choice because it is a permutation of the values $[2 \ldots N-1]$, because it is non-linear, and because it uniformly distributes the data. However, the evaluation of the integer exponent k^m is much more expensive that the simple divide hash function $h(k) = k \bmod N$, requiring a multiplication and division for each bit of m versus a single division [7]. However, consider that two different hash functions must be evaluated for each key accessed. Clearly this would be a poor choice in terms of performance relative to commonly used methods.

What is needed is a hash probe function that has a large Lyapunov exponent as evaluated over the first few iterations, rather than the entire range of N. This is based on the fact that while double hashing has a nearly ideal Lyapunov exponent when evaluated over the whole table, its worst Lyapunov measure is in the first few probe iterations. In addition, the hash function should preserve all of the desirable characteristics of double hashing, including fast run time, long probe sequences, and no primary or secondary clustering for similar keys. The function could not be a linear function of i, or it would suffer from the same limitation as double hashing. Some introductory number theory is necessary to develop a new hash probing function based on the calculation of an exponent modulo N.

Definition 1 *(cyclic group, generator) If a group G contains an element a such that every element of G is of the form a^k for some integer k, then G is a* cyclic group, *and a is called a* generator *of G [6].*

The group Z_p^* consisting of the elements { [1 ... p] } and operator *, which is normal multiplication modulo p, p prime, forms a cyclic group. This is derived directly from the definition of a cyclic group. The hash function:

$$H(k,i) = h(k)^i \bmod N \tag{2}$$

where $h(k)$ is a hash function returning values in the range $[2, \ldots, N]$. Equivalently, expressed as an iterator using the x_i notation previously defined:

$$x_i = x_0^i \bmod p \tag{3}$$

where $p = N$ must be a prime hash table size. This function is similar to the RSA and ElGamal cryptosystems [7], in that a finite field exponent is used to create a non-linear permutation of values. This probe sequence has the following characteristics:

- It can be computed efficiently. The value x_i at the i'th step is simply the previous value x_{i-1} times x_0 modulo N. This requires the same number of mathematical operations as linear or double hashing.
- The probe sequence is non-linear. Small perturbations in the initial value x_0 become large differences after only two iterations.
- The probe sequence depends entirely on the initial hash value x_0, which may lead to primary and secondary clustering. Fortunately this can easily be remedied by adding a second hash function value $h_2(k)$ as will be demonstrated shortly.
- The probe sequence is not of length N for all values of x_0, since only cases where x_0 is a generator for Z_N^* will generate the full domain.

4.1. THEORETICAL PERFORMANCE

First, consider the probe length of the groups and subgroups that equation (3) generates:

Definition 2 *(order) The number of unique elements in a group is called the* order *of the group. The group Z_p^* has order $p-1$.*

Definition 3 *(subgroup) A subset H of a group G is a* subgroup *of G, if H is itself a group relative to the binary operation defined in G.*

Theorem 1 *(Lagrange's Theorem) If G is a group of order N, then the order of every subgroup H of G is a divisor of N.*

Lemma 1 *The number of generators for a cyclic group of order N is $\phi(N)$ where $\phi(N)$ denotes the Euler function — the number of integers less than N which are relatively prime to N.*

Ideally, all of the elements of Z_p^* should be generators of the entire group, for this would imply that every element could be generated starting with any element. This would mean that every element leads to a probe sequence of length $p-1$. Applying the above to the new hash function, it is readily apparent that Z_p^* is a group of order $p-1$, and that the number of generators for the group is $\phi(p-1)$. Unfortunately p must be prime for Z_p^* to be a cyclic group. This means that $p-1$ must be an even number, since p is odd. Therefore $p-1$ must have as one of its factors the number 2, which means at most only $p/2$ elements of Z_p^* as generators of the group. This can be easily verified empirically by trying any small group of prime size.

Next, apply Lagrange's theorem to partially correct this deficiency. Since Z_p^* is a group of order $p-1$ all subgroups H of Z_p^* must have orders that are divisors of $p-1$. However, carefully selecting $p-1$ such that it is the product of 2 and another prime number t will assure that all subgroups of Z_p^* have order of either 2

or t. In fact if the prime t is chosen carefully so that $p = 2t + 1$ is also prime, all elements in the group Z_p^* will either be generators of the entire group, generators of subgroups of order t, or generators of subgroups of order 2.

Since the subgroup of order 2 is also cyclic, it has only one generator in Z_p^*. It is easy to see that the value $x = (p-1)$ is in fact the only element in Z_p^* which can generate a subgroup of order 2. These results are summarized in theorem 2.

Theorem 2 *Given p and t primes, with $p - 1 = 2t$, the group $G = Z_p^*$ contains exactly $t - 1$ generators for the entire group G, t elements which are generators for subgroups of order t, and one element which generates a subgroup of order 2.*

Therefore, the following conclusions can be derived for the exponential probe function in equation (3) by applying theorem 2.

- Half $(t-1)$ of the choices for x_0 will be generators for Z_p^*, and will create probe sequences of length $p - 1$.
- Exactly t of the values for x_0 will generate probe sequences of length t.
- Only one value, $x_0 = p - 1$ will generate a poor probe length of 2. This value can be avoided by choosing the initial value $x_0 = (k \bmod (p-2))+1$.
- Different initial values x_0 will generate unique probe sequences in Z_p^*.
- The primes t and p can be efficiently generated with $p-1 = 2t$ using probabilistic primality testing. The Prime Number Theorem says that there are approximately $\log(N)$ prime numbers less than N. Therefore the expectation is that one would have to explore at most $\log^2(N)$ such pairs to find a suitable table size probabilistically. This only needs to be done during initialization of the table.

This new exponential probe function has many desirable characteristics. Except for the less than optimal probe length on 1/2 of the table elements, it has many of the characteristics of double hashing.

4.2. HASH TABLE EXPERIMENTS

To test the above hypothesis, we implemented both double hashing and the new exponential hash function. Table sizes were determined using the double prime criterion where $N = p = 2t + 1$ required for the exponential hash function where p and t are both prime. The Miller-Rabin probabilistic primality test was used to determine the next largest prime table size meeting this criterion given a target table size [7]. Based on the earlier analyses for both functions this should produce optimal probe lengths. All trial runs were done by creating two identical empty hash tables of the same size. Elements chosen at random from the data distributions described below were successively added to the table to achieve a load factor of 95% ($\alpha = 0.95$) of the table size. The measure of merit was the average number of probes required per element added. For example, if k elements take a total of m probes, the average probes per element is simply m/k. Samples of this metric were taken every 5% of the table load factor, from 5%...95% to determine the behavior as load factor increased.

Summaries of four experiments are presented here:

- Uniform data distribution over the entire table size — To show that the exponential hash function and double hash function have statistically equivalent performance for uniformly distributed random data.
- Clustered data distribution — To demonstrate improved performance of exponential hash function over double hash function for tightly clustered data.
- Variation of cluster size — To demonstrate sensitivity of improved exponential hash function performance to size of the data cluster.
- Variation of table size with fixed percentage cluster — To demonstrate sensitivity of exponential hash function to table size using a fixed data cluster size.

4.3. UNIFORM DATA DISTRIBUTION

The control case for this analysis was a series of runs done on a uniform data distribution with a fixed table size. Two identical tables of equal size were created and filled to 95% capacity, one using the double divide hash function and the other using the exponential hash function. The test was repeated 100 times using a different random number seed for each run to determine if any statistical difference in total number of probes to fill the table could be detected. A table size of 3023 was used for these runs. The summary results are presented in table 1.

With 100 runs, no statistically significant difference in performance could be detected. The difference in total number of probes between the two functions is only 1.76%, which is significantly less than the 15.51% standard deviation as measured between runs of the

Measure	*Double*	*Exponential*
Total Probes	1028281	1010148
Avg Probes Per Run	10282	10101
Std Deviation	305.0	257.7

Table 1: Uniform data comparison

same hash function. This means that the exponential hash function and double divide hash function are statistically equivalent for randomly chosen uniformly distributed data.

4.4. CLUSTERED DATA DISTRIBUTION

The second experiment involved clustering the data over a sub-interval of the total table size, in an attempt to simulate a dense data cluster. The dense data cluster represents many real world data sets where data is far from evenly distributed. A table size of $p = 3023$ was chosen, and all of the data was chosen at random from a single data cluster of approximately 300 elements from the beginning of the data space. Samples of the average number of probes per data element were taken for every 5% of table size, up to a total table load of 95%. Results of this test indicates that the exponential hash function out-performs the double divide hash function. For high table load the double divide hash function stores data in as little as half the number of probes.

4.5. VARIATION OF CLUSTER SIZE

This experiment was similar to the previous one, but in this case the cluster size was varied from 2% to 20% of the overall table size. Again, identical tables were created and populated using both the double divide hash function and the exponential hash function. Data was taken at random from a cluster of size varying from 2% to 20% of the table size, and the average number of probes per element inserted was sampled for each table to reach 95% capacity. A table size of 2027 elements was used for all experiments. The results, show that the exponential hash function uses far fewer probes than the double divide hash. Further the relative advantage seems to be larger for more tightly clustered data.

4.6. VARIATION OF TABLE SIZE

The next experiment varied the table size to see if it had any effect on relative performance of the two functions. Tables were created with from approximately 1000 entries to 15000 entries, and filled to 95% capacity using the double hash and exponential hash function. The exponential hash function performed better for all table sizes, and table size appeared to have little effect on the relative outcome.

5. CONCLUSIONS

The results presented here indicate a new relationship between chaotic iterator theory and non-uniform open hash function performance. Results indicate that the proposed exponential hash function does outperform double divide open hashing for some clustered data distributions, and performs as well for uniform data distributions. A number of avenues for future research are now open. It is likely that other measures from nonlinear systems theory can be applied to hash functions, and may also provide additional indicators of hash table performance, possibly leading to further improvements in the exponential open hash function presented here. Further, it may eventually be possible to apply chaotic measures to other hash function applications such as cryptographic signature verification to detect undesirable hash function characteristics. A full length version of this paper, with complete details of the Lyapunov exponent analysis and complete test results are available from the authors.

6. REFERENCES

[1] Gregory L. Heileman. *Data Structures, Algorithms and Object-oriented Programming.* McGraw-Hill, New York, NY, 1996.

[2] Gregory L. Heileman, Chaouki Abdallah, Donald R. Hush, and S. Baglio. Chaotic probe strategies in open address hashing. In *Proceedings of International Symposium on Nonlinear Theory and its Applications*, 1993.

[3] Donald E. Knuth. *The Art of Computer Programming*, volume 3. Addison-Wesley Publishing Co., 1973.

[4] Heinz-Otto Peitgen, Hartmut Jürgens, and Dietmar Saupe. *Chaos and Fractals: New Frontiers of Science.* Springer-Verlag, New York, 1992.

[5] W. W. Peterson. Addressing for random access storage. *IBM Journal of Research and Development*, 1(2):130–146, 1957.

[6] George W. Polites. *An Introduction to the Theory of Groups.* International Textbook Company, Scranton, Pa, 1968.

[7] Douglas Stinson. *Cryptography Theory and Practice.* CRC Press, Boca Raton, FL, 1995.

Hardware and Software Implementation

FACT™: A C++ Environment for Accurately Modeling Fixed-Point Digital Signal Processors

By

Anastasios S. Maurudis
Algorithm Research and Development Group
DSP Software Engineering, Inc.
175 Middlesex Turnpike
Bedford, MA 01730 USA
anastasios@dspse.com

ABSTRACT

In this paper, we present the DSP Software Engineering (DSPSE) Fixed Arithmetic C++ Tool (FACT™) method for modeling DSP fixed-point processors in a C++ environment and converting a floating point model to a given DSP fixed-point processor model. All DSP fixed-point processors have associated with them a set of fixed bit length data representations for the storage and manipulation of binary information. We define a C++ class for each distinct fixed bit length data representation of a given DSP fixed-point processor, such that, the behavior of the given DSP fixed-point processor can be achieved in a C++ environment using the library of classes. For our own development, we have created, most recently, a FACT™ library for the Texas Instrument TMS320C54x DSP fixed-point processor. The TMS320C54x library has been used in the development of the Japanese Vector Sum Excited Linear Prediction (JVSELP) algorithm and the International Telecommunications Union G.728 standard algorithm.

I. INTRODUCTION

With the explosive growth of the DSP market, we have seen a direct increase in the use of fixed-point digital signal processors in a variety of industries, such as telecommunications, speech/audio processing, instrumentation, military, graphics, image processing, control, automotive, robotics, consumer electronics and medical technology. In general, fixed-point DSPs compared to floating-point DSPs are less expensive, use less power, and less space. One advantage of a floating point DSP is a smaller development cost (i.e. man hours), however, you compromise a greater production cost. Thus, if possible, companies are using and will use fixed-point DSPs for their products. In the near-future, we will be faced more and more with the challenge of real-time implementations of complex DSP algorithms on fixed-point DSPs. The FACT procedure is the outcome of our desire to decrease the development time of fixed-point implementations.

We will assume the following software development cycle model for the real-time implementation of a given algorithm on a fixed-point DSP:

1) floating point model
2) fixed point model
3) real-time implementation.

At DSPSE, our development time is drastically reduced using a FACT™ procedure with the above development model. By decreasing development time, we have narrowed the advantage gap between floating point DSPs and fixed-point DSPs.

Besides being able to model a fixed-point DSP in a C++ environment, a FACT™ library expedites the conversion of an algorithm from a floating point model to a given fixed-point processor model; from step 1 to step 2 in our development model. For the FACT™ floating-point model we define a C++ class, say FLOAT. We attach various data members to our class FLOAT to keep track of pertinent information for transforming a floating point model to a fixed-point model. Moreover, suppose the floating point model of an algorithm calls N modules then we need a fixed-point model for each of the N modules under each fixed-point processor we wish to model.

Situations will also arise when we will want to convert only certain modules to a fixed-point processor model while leaving other modules as a floating point model, such as a fixed-point encoder and a floating-point decoder. In order to accomplish the dual existence of a fixed and floating point model, we create an C++ interface class, to do exactly that, interface a fixed-point module with a floating point module. In terms of linear algebra, our interface class acts as a transformation operator, transforming from a FACT™ fixed-point model space to a FACT™ floating-point model space.

The paper is organized as follows. In section II, we discuss the creation of a FACT™ library. Within section III, we further explain the FACT procedure by showing examples in modeling the TMS320C54x. We introduce a FACT Floating-point model in section IV and discuss the transformation from a floating point model to a FACT™ fixed-point model. While our conclusion is in section V.

II. Creating a FACT™ Library

A. Distinct Fixed Length Data Representation

All fixed-point DSPs have associated with them a set of fixed bit length data representations for the storage and manipulation of binary information. A fixed bit length data representation is considered distinct if any of the following three conditions are met: 1) the length is different; 2) if the length is the same then an operation exists which will produce a different result given the same input value(s) of identical length and under the same control conditions; 3) if the length is the same then an operation exists which can not be performed on a data representation of the same length. By control conditions, we mean all status fields, control fields, mode of operation and the like.

The reasoning for condition 1 is obvious, an L1 bit length can not exactly represent an L2 bit length, unless L1 = L2. Suppose L1 = 16 and L2 = 32, we can not use 16 bits to represent 32 bits. One might say, you can use 32 bits to represent 16 bits. For example, let us assume we are using the lower 16 bits of a 32-bit representation to simulate a 16-bit representation. From our point of view, the 16-bit simulation is not the same as the actual 16-bit representation since we are concerned with bit exact similarities. That is, the 16-bit simulation really is 16 zeros followed by 16 binary digits, as compared to just 16 binary digits. Condition 2 exists when L1=L2 and at least one operation will produce different results with the same identical inputs. For example, a fixed-point DSP could have more than one accumulator and depending which accumulator is an input and/or an output, an operation produces different results. Condition 3 exists when L1=L2 and an operation can not be performed on all representations of the same length, just some. Again, using the multi-accumulator example, at least one operation exists that will not accept all accumulators as an input. For example, on the TMS320C54x there are instructions which will produce different results depending on whether the source (input) or destination (output) accumulator is A or B, even if the input value and the control conditions are the same. And, as is for most processors, certain registers which are 16-bits in length can not be operated on as a 16-bit short data memory operand can.

The different fixed bit length data representations are grouped into a set. We will refer to the set of fixed bit length data representations, for a given fixed-point DSP, as the length set vector, $\Lambda = \{\lambda_1,\lambda_2,...,\lambda_M\}$, where each λ_i, for i = {1,2,...,M}, is a non-zero positive integer equal to the length of the distinct representation in bits. Thus, M is the total number of distinct representations of information possible on a given fixed-point processor. For example, on the TMS320C54x the length set vector, Λ_{C54x} = {40,40,32,16,16}. Hence, we can conclude that the TMS320C54x has five distinct data representations; two being 40 bits in length, one 32 bits in length, and the other two, 16 bits in length. The two 40-bit lengths, 32-bit length and two 16-bit length are due to the existence of 40-bit accumulator A, 40-bit accumulator B, the ability to address 32-bit operands, 16-bit registers and the ability to address 16-bit operands, respectively.

B. C++ Class Hierarchy

As stated earlier, each distinct fixed bit length data representation has an associated C++ class. Thus, each λ_i has an associated C++ class which, if possible, is derived from another class for the same bit-length. The actual procedure for deciding which, if any, class a given distinct fixed bit length data representation is derived from is developed in [2]. The authors in [2] use the projection theorem by representing each distinct fixed bit length data representation as a vector space.

The base class may be an abstract class, which allows pure virtual function declarations or, the base class can define a virtual standard set of operation definitions to be performed on the use of a base class object. The former choice is good in applications where the end-user must choose which DSP fixed-point processor to model since objects of an abstract class can not be created, while the later is used in situations where no specific processor is modeled but the standard DSP processor as determined by the library creator. That is, objects of the standard class are allowed. The concept of the base class becomes more clear as we explain the power structure of class inheritance

Suppose we want to create a library to model DSP fixed-point processors A, B and C. Let us assume that the length set vectors for DSP processors A, B and C are

$$\Lambda_A = [40,40,32,16,16], \quad M=5, \quad (1)$$

$$\Lambda_B = [64,40,32,16,16], \quad M=5, \quad (2)$$

and $$\Lambda_C = [64,40,32,32,16,16], \quad M=6, \quad (3)$$

respectively.

For sake of brevity, we will go through the details of creating only the class for the 64-bit length data representation of the DSP fixed-point processor B needed to create the library. The same procedure is applied to the other bit-length data representations. Furthermore, we will assume that we have already created a base 64-bit base class, say I64, with virtual operator definitions. Thus, we need to create a class, say I64_B, for the 64-bit length data representation of fixed-point processor B.

The operators(i.e. instructions) to be defined in the I64_B class can be grouped into two categories, (a) operators already defined in the I64 class and (b) operators not defined in the I64 class. You can think of the category (a) operators as the projection of the I64_B operators onto the I64 operators. Of course, if the projection was the empty set then

I64_B will not be derived from I64. Furthermore, the base class should not have any operators for which the I64_B class should not implement. Analogous to linear algebra, the previous statement implies that category (a) accompanies all of class I64, the base class, such that I64_B is the direct sum of I64 plus category (b) operators. That is,

$$I64\_B = I64 \oplus \text{category (b)}. \quad (4)$$

For our case, let us assume that the projection was not the empty set and that all I64 operators are to exist in the I64_B class, such that I64_B is derived from I64.

We can divide I64_B operators into two orthogonal sets of operators. The first set is accomplished by taking the operator projection of I64 onto I64_B. We will refer to the operator projection of class α onto class β in our study, as $O(\alpha,\beta)$. The other set is the rest of the I64_B instructions which need to be initially defined for the implementation of an I64_B object. Therefore, we can decompose our I64_B class into the following:

$$I64\_B = O(I64,I64\_B) \oplus (I64 \perp I64\_B) \quad (5)$$

The last term, $(I64 \perp I64\_B)$, is the set of operators which need to be added to the I64_B class, what we refer to in equation (4) as category (b) operators.

The same methodology is applied to the creation of the rest of the classes until all 16 classes are created. Once we have these 16 classes we have a library for modeling DSP fixed-point processors A, B and C in a C++ environment. One possible power structure of the class hierarchy for a library to model fixed-point processor A, B and C is shown, in figure 1, below. We show the structure with two 16-bit length standard base classes; I16 to mimic 16-bit length data operands and R16 to mimic 16-bit length registers.

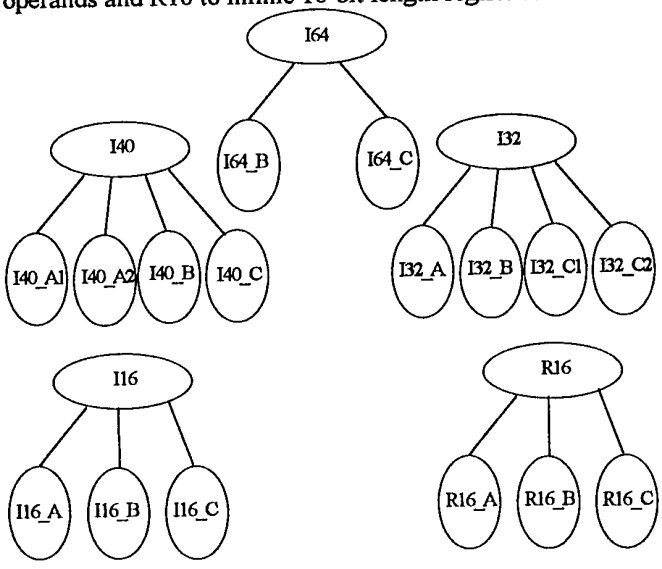


Figure 1: Power Structure of Class Hierarchy for Example Library

III. Examples of the TMS320C54x FACT™ Library

The length set vector for the TMS320C54x, Λ_{C54x} = {40,40,32,16,16}, contains five (5) elements. Let us focus on the accumulators. The TMS320C54x has two accumulators, referred to as accumulator A and accumulator B, with a 40-bit length. Each accumulator contains three memory-mapped registers: Guard bits (AG,BG), High-order bits, (AH,BH), and Low-order bits, (AL,BL). As shown in figure 6 and figure 7, the layouts for the accumulators are the guard bits which are 8-bits in length, while the high-order and low-order bits are 16-bits in length, bringing the total length to 40.

The TMS320C54x I40A/B class is used to declare and define operators and functions which utilize the TMS320C54x accumulator A or B. In other words, if you were to use an assembly instruction equivalent, equivalence with respect to an operator or function in the C++ model, the final result bit matches with the C++ model result. Moreover, in the C++ model, we are able to explicitly state whether a 40-bit variable resides in accumulator A or accumulator B, by creating two separate classes.

Our simulation is accomplished by using a 32-bit integer and an 8-bit unsigned character in tandem as the data members for our I40 structure, shown in figure 8. The 32-bit integer is called guardhi, while the 8-bit character is called low. As shown in the layout below, guardhi contains the 32 MSBs of the accumulator and low contains the remaining 8 LSBs. In other words, guardhi contains the guard bits, high-order bits, and 8 MSBs of the low-order bits and low incorporates just the 8 LSBs of the low-order bits.

As a reminder, the I40 layout, in figure 8, does not use accumulator specific notation (e.g. AH vs. H), since the I40 structure is accumulator independent. That is, the I40 class is a base class for the two accumulators. Simply stated, the ability to do 40-bit manipulation and operations is accomplished by telling guardhi and low what to do for each operator and function defined within this structure.

IV. FACT™ Floating-point Model

A DSPSE FACT™ Floating point model uses C++ classes for creating instances of variables. The FACT™ floating point data representation is implemented by a C++ class, say FLOAT. We attach various data members to our class FLOAT to keep track of pertinent information for transforming a floating point model to a fixed-point model. The preferred embodiment has the following data members:

Value = current value
Max_abs = running maximum of the absolute of Value
Min_abs = running minimum of the absolute of Value

Avg_abs = running average of the absolute of Value
Var_abs = running variance of the absolute of Value
Read_count = number of read accesses made of Value
Store_count = number of write accesses made of Value

We also declare global variables to keep track of the number of time a give function is called. In the preferred embodiment, we keep track of all mathematical operations (addition, multiplication, subtraction, division). Having the information provided by the preferred embodiment on any variable we declare as a FLOAT aids in determining the computational complexity, dynamic range, scaling effects, and Q storage format.

V. Converting a FACT™ Floating point Model to a FACT™ Fixed-point Model

Let us turn our attention to converting a floating point model of an algorithm to a given fixed-point processor model. Suppose the floating point model of the algorithm calls N modules then we would need a fixed-point model for each of the N modules under each processor we wish to model. Situations will also arise when we will want to convert certain modules to a fixed-point processor model while leaving other modules as a floating point model. One scenario could be a fixed-point encoder in tandem with a floating point decoder or you may want to convert only one module to a fixed-point model at a time and still be able to execute your algorithm with floating point modules.

In order to accomplish the dual existence of a fixed-point and floating point model, we create an interface class, to do exactly that, interface a fixed-point module with a floating point module. Let us call the interface class `ToInt` with a public data member, say `data`, of class type `FLOAT`. For sake of brevity, let N = 2 and say we want a fixed-point processor B model for the pure float model example algorithm shown in figure 2. In figure 3, we are testing a fixed-point model of `func1 ( )` with a floating point model of `func2 ( )`, while in figure 4 the roles of the modules are reversed. Then, in figure 5, we show both modules being fixed point models.

```
FLOAT func1(FLOAT);
FLOAT func2(FLOAT);
void main{
FLOAT a,b,c;
b=func1(a);
c=func2(b);
return 0;}
```

Figure 2: Pure FLOAT model.

```
ToInt func1(ToInt);
I64_B func1(I64_B);
FLOAT func2(FLOAT);
void main{
ToInt a,b;
FLOAT c;
b=func1(a);
c=func2(b.data);
return 0;}
ToInt func1(ToInt d)
{I64_B e(d);
 I64_B f;
 f=func1(e);
 return (ToInt)f;}
```

Figure 3: Mixed Model

```
FLOAT func1(FLOAT);
ToInt func2(ToInt);
I64_B func2(I64_B);
void main{
FLOAT a;
ToInt b,c;
b.data=func1(a);
c=func2(b);
return 0;}
ToInt func2(ToInt d)
{I64_B e(d);
 I64_B f;
 f=func2(e);
 return (ToInt)f;}
```

Figure 4: Mixed Model

```
I64_B func1(I64_B);
I64_B func2(I64_B);
void main{
I64_B a,b,c;
b=func1(a);
c=func2(b);
return 0;
```

Figure 5: Pure Fixed Model

By taking advantage of C++ function mangling, we create three definitions of a module(i.e. same function name): floating point definition, fixed-point definition, and interface definition. The interface definition accepts as arguments interface class objects with data members of class type FLOAT, then converts the objects to a fixed-point data representation class for the desired DSP, in our case a 64-bit length data representation for processor B, I64_B. Then, the interface definition calls the fixed-point definition, which returns a fixed-point class object to the interface definition. The returned fixed-point class object is converted to an interface class object upon return to the calling function of the interface definition. The main feature is that we can easily simulate the algorithm on another processor by replacing all instances of I40C objects with a data representation of the target processor. Furthermore, we can have assembly level characteristics in our C++ environment since we define the behavior of all operations under all control conditions. For example, our add operators can simulate sign extension mode, overflow mode, etc.

VI. Conclusion

Using the FACT™ approach, one can create a library, with an efficient class hierarchy, for accurately modeling various DSP fixed-point processors in a C++ environment. Furthermore, the FACT™ library is an adaptive library. Adaptive in the sense that other fixed-point processors may

be added in their entirety or for a current library fixed-point processor, its associated operators and their definitions may be added, removed or modified as needed. Once a FACT™ library is available for a given set of processors, any algorithm can be modeled under any fixed-point processor of the library. The multi-processor capability of a FACT™ library facilitates the comparison of an algorithm under different fixed-point processors without necessarily coding at assembly level. Moreover, by using a FACT™ library, the development time involved in going from a fixed-point model to an assembly level version for a given algorithm is dramatically reduced. The reduction is possible since a FACT™ fixed-point model has assembly level characteristics built into it.

VII. References

[1] A. Stevens, *Teach yourself ... C++*, MIS: Press, New York, 1995

[2] A. Maurudis, "FAT™: An Efficient Vector Space Method for Accurately Modeling Fixed-point Digital Signal Processors", to be published.

39	38	37	36	35	34	33	32
AG 7	AG 6	AG 5	AG 4	AG 3	AG 2	AG 1	AG 0

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16
AH 15	AH 14	AH 13	AH 12	AH 11	AH 10	AH 9	AH 8	AH 7	AH 6	AH 5	AH 4	AH 3	AH 2	AH 1	AH 0

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
AL 15	AL 14	AL 13	AL 12	AL 11	AL 10	AL 9	AL 8	AL 7	AL 6	AL 5	AL 4	AL 3	AL 2	AL 1	AL 0

Figure 6 Layout of TMS320C54x Accumulator A

39	38	37	36	35	34	33	32
BG 7	BG 6	BG 5	BG 4	BG 3	BG 2	BG 1	BG 0

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16
BH 15	BH 14	BH 13	BH 12	BH 11	BH 10	BH 9	BH 8	BH 7	BH 6	BH 5	BH 4	BH 3	BH 2	BH 1	BH 0

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
BL 15	BL 14	BL 13	BL 12	BL 11	BL 10	BL 9	BL 8	BL 7	BL 6	BL 5	BL 4	BL 3	BL 2	BL 1	BL 0

Figure 7 Layout of TMS320C54x Accumulator B

39	38	37	36	35	34	33	32
G 7	G 6	G 5	G 4	G 3	G 2	G 1	G 0
gh 31	gh 30	gh 29	gh 28	gh 27	gh 26	gh 25	gh 24

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16
H 15	H 14	H 13	H 12	H 11	H 10	H 9	H 8	H 7	H 6	H 5	H 4	H 3	H 2	H 1	H 0
gh 23	gh 22	gh 21	gh 20	gh 19	gh 18	gh 17	gh 16	gh 15	gh 14	gh 13	gh 12	gh 11	gh 10	gh 9	gh 8

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
L 15	L 14	L 13	L 12	L 11	L 10	L 9	L 8	L 7	L 6	L 5	L 4	L 3	L 2	L 1	L 0
gh 7	gh 6	gh 5	gh 4	gh 3	gh 2	gh 1	gh 0	low 7	low 6	low 5	low 4	low 3	low 2	low 1	low 0

Figure 8 Layout of I40 class data members

CONTROL OF A HANDS-FREE TELEPHONE SET

Christina Breining

Technische Hochschule Darmstadt
Fachgebiet Theorie der Signale
Merckstr. 25, 64283 Darmstadt, Germany
e-mail: breining@nesi.e-technik.th-darmstadt.de

ABSTRACT

In this paper, we discuss different methods of double talk detection used for echo cancellation in a hands-free telephone set. We deduce a general state-space representation of the set, which leads to state-dependent structures for stepsize control. The approach is further generalized by introducing fuzzy logic and fuzzy state memberships. Finally, first results are shown which are obtained with identical detection methods, but applying different control structures.

Introduction

A hands-free telephone set includes a loudspeaker and a microphone which are placed in the same room. One problem arising from this fact is that the far-end signal is retransmitted to the far-end speaker due to the acoustic coupling. This means that the far-end speaker hears his own voice after some delay, which can render the conversation impossible.

This effect can be avoided by echo cancellation: an adaptive filter imitates the acoustic transmission system of the room. An artificial "echo" can thus be produced and subtracted from the local signal. Echo cancellation can deal with double talk situations since the estimated echo is suppressed while the local speaker signal is transmitted.

Much work has been dedicated to the convergence speed of adaptation algorithms, often based on time-invariant room characteristics. In a real environment, however, the room impulse response is not constant over time, which is due to movement and temperature in the room, and the filter has to be adapted during the whole conversation, which includes noisy and double talk situations. Therefore, the convergence speed of real-time adaptation is strongly influenced by the ability of the system to adapt to the instantaneous situation. Without appropriate adaptation control, even the fastest algorithms cannot guarantee a sufficient adjustment of the filter. This will be explained at the example of the NLMS algorithm whose adaptation equation is

$$\underline{c}(k+1) = \underline{c}(k) + \alpha \frac{e(k)\underline{x}(k)}{||\underline{x}(k)||^2} \qquad (1)$$

where $\underline{c}(k)$ is the vector of N filter coefficients at time sample k, $\underline{x}(k)$ the vector that comprises the N latest far-end signal samples in a column, $e(k)$ the error that results from substracting $\underline{c}^T(k)\,\underline{x}(k)$ from the local signal at time k, and α the stepsize. The correction term $\frac{e(k)\underline{x}(k)}{||\underline{x}(k)||^2}$ will be large whenever the error is large, and small when the error is small. But $e(k)$ contains adaptation error, local noise and local speaker signal. When this term is large, the echo canceller might be badly adjusted or else the local speaker might be active. In the latter case correction should be small instead of large. The more disturbance is involved, the smaller the stepsize must be. Therefore, we need a control mechanism that can determine the situation and choose the stepsize accordingly. Since the detection algorithms are not fully reliable, combining them could be helpful.

This paper is organized as follows: In section 1 we present and discuss several methods for the determination of an appropriate stepsize. Section 2 introduces a state-space representation of the hands-free telephone set. General control structures including fuzzy concepts are presented in section 3 before we show first results in section 4.

1. METHODS FOR STEPSIZE CONTROL

Echo cancellation algorithms are usually deduced by modelling the far-end speaker signal and the local distortion as mutually uncorrelated white noise. The optimal stepsize can then be calculated as

The author is supported by the Graduiertenkolleg Intelligente Systeme in der Informations- und Automatisierungstechnik

$$\alpha_{opt} = \frac{E\{\epsilon^2(k)\}}{E\{e^2(k)\}} \tag{2}$$

where

$$\epsilon(k) = (\underline{g} - \underline{c}(k))^T \underline{x}(k) \tag{3}$$

is the part of the error signal $e(k)$ caused by misadjustment, $\underline{g}$ being the room impulse response (see [5]).

Hence, the optimal stepsize is one when there is no background noise or local speaker signal, and close to zero in the case of an active local speaker. Since we cannot measure $\epsilon(k)$ but only $e(k)$, the optimal stepsize has to be estimated by one of various methods.

One possibility is proposed in [5]: Two different FIR filters are used, an internal one for the determination of the coefficients and the stepsize, the other for the actual echo cancellation. The stepsize can then be estimated by adding artificial delay coefficients as first taps to the internal filter. These "noncausal" filter coefficients are to adapt to zero, so their adaptation quality can be measured and the optimal stepsize is calculated in the form:

$$\alpha_{opt}(k) = \frac{E\{((\underline{g} - \underline{c}(k))^T \underline{x}(k))^2\}}{E\{e^2(k)\}} \tag{4}$$

$$\approx \frac{N + N_T}{N_T} \frac{\sum_{i=0}^{N_T-1} c_i^2(k)\, \hat{\sigma}_x^2(k)}{\hat{\sigma}_e^2(k)} \tag{5}$$

with N_T the number of artificial delay coefficients introduced, and N the number of filter coefficients for imitation of the room impulse response.

This algorithm works quite well for time-invariant room impulse responses. It can deal with different levels of background noises and sets the stepsize to very small values during double talk. But the artificial delay coefficients will not detect changes in the room impulse response because their misadjustment to the zero vector remains the same whereas the misadjustment of the echo cancellation part of the filter increases.

Another control strategy is to set the stepsize to zero during double talk and to a constant value during single talk (considering the usual background noise, this value will be less than 1). The task is here to detect double talk or dominant background noise and switch the adaptation on or off. One method for double talk detection is proposed in [2] and uses the normalized correlation between the far-end signal and the local signal which consists of the echo and the local speaker signal. Stating that the room impulse response only slightly affects the correlation between the far-end speaker signal before and after passing the room, we assume a high correlation to be caused by low echo attenuation and therefore enable the adaptation, whereas with low correlation it is disabled. The correlation term must be estimated over a limited number of samples, so that there is a tradeoff between estimation quality and detection delay. During this delay double talk, although taking place, has not yet been recognized, which may lead to severe misadjustment. The importance of the damage depends on the convergence speed of the adaptation, so that with a small stepsize during adaptation, the stepsize control by correlation is satisfying, but not with fast algorithms or large stepsizes.

Other methods involve knowledge about room and speech characteristics, usually in the frequency domain. One detector recognizes variance of the room impulse response by analyzing the spectra of the error signal and the local signal. A large quantity of the speech signal power is situated with lower frequencies, whereas moving objects in a room cause high-frequency change of the room impulse response (see [4]). Therefore, an increasing error signal power, for which the power in the higher band is increasing in relation to that of the local signal, indicates that the room impulse has changed, else double talk is assumed. This criterion reacts rapidly because of the short filters used for the analysis, but it only works at the beginning of the variation of a room impulse and does not indicate when this situation ends. Additionally, this method requires a certain level of echo attenuation to function properly.

Even more speech characteristics are used for the so-called cepstral distance measure which is described in [1] and determines if the error signal comes from the far-end speaker or not by analyzing the cepstrum. Like most speech processing tools, it is quite complicated to calculate and implies a considerable delay. Many more methods can be adopted from speech processing, but they are usually not fast enough and will not be further discussed in this paper.

2. STATE-SPACE REPRESENTATION

These are only some of the known algorithms, but they show that the criteria usually do not distinguish between all the situations, or states, of the hands-free telephone set. The stepsize control should rely on different criteria for different states. Therefore, we interpret the hands-free telephone set as a finite-state machine whose relevant states are described by four parameters, i. e. sufficient/ insufficient amplitude of the far-end signal, sufficient/ insufficient adaptation, local noise/ negligible local noise, local speaker active/ inactive. These parameters lead to the distinction of 16 states which can be imagined as the corners of a four-dimensional cube. Since the basic condition for efficient adaptation is a sufficient excitation level, we can repre-

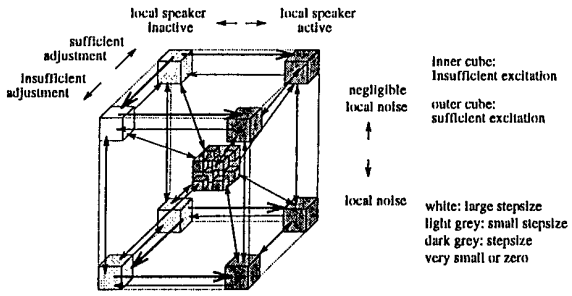


Figure 1: Representation of the hands-free telephone set in the state space. Bold arrows represent critical transitions

sent all states without sufficient excitation in one state. This state can easily be determined, as the excitation signal is known. The number of states is thus reduced from 16 to 9. In case of sufficient excitation, the state space can now be represented as a three-dimensional cube. In this representation, we assume that only one variable changes its value at a time, i. e. only the transitions situated at the edges of the cube are allowed. This is justified by the high sampling rate used for real-time application. If we decide to stop the adaptation entirely during double talk, we can define one state double talk comprising the four former states, so that, depending on the chosen control structure, the number of states can be further reduced.

The adaptation control should be fast and reliable so as to maximize convergence speed and optimize tracking behaviour. The critical transitions are therefore linked to the ability of the criteria to distinguish between the states. The critical transitions for several criteria are drawn as bold arrows in fig. 1.

The first control algorithm described here has only got to be restarted when the room characteristics change, i. e. there are two states to be distiguished, and the critical transitions lead from "sufficient adjustment" to "insufficient adjustment". But in general, as for the correlation coefficient method, the critical transitions are from single talk to double talk: If these transitions are not detected correctly, the adaptive filter can become completely misadjusted after only a few samples. On the other hand, if we try to detect every transition to double talk, we will also define many single talk situations as double talk and will reduce convergence speed considerably.

To compensate for the weakness of the algorithms in some situations, complete control algorithms combine the stepsize control by delay coefficients with the cepstral distance measure (see [1]), usually by logic AND, or one chooses a parameter set for the method so that the convergence of the echo cancellation is convienient for all states (see [2]).

To further improve the stepsize, we can extend both concepts: we can try to optimize either the parameter set for each state of the telephone set or the way of combining the methods, e. g. by nonlinear functions. We can generalize the concept in choosing an appropriate combination of reliable criteria for each state of the telephone set. In this case, we have to determine the instantaneous state, and choose the kind of combination to be applied accordingly. In order to extract more than a binary decision from the double talk detectors, we can utilize fuzzy logic so that there exists an individual fuzzy rule base for each state. The new stepsize is one of the clues for the determination of the new state.

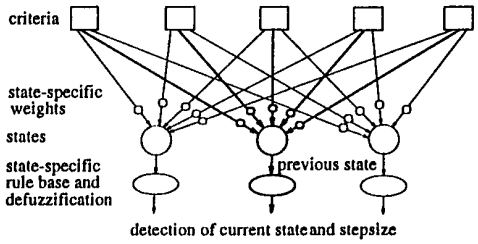


Figure 2: Concept for the logic of the control unit, discrete states. Bold lines mark the active path.

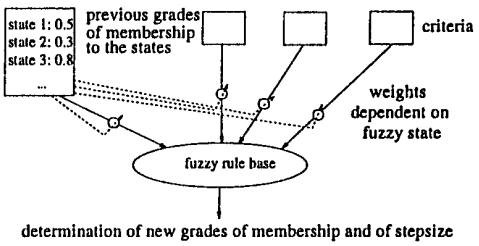


Figure 3: Concept for the logic of the control unit, fuzzy states. Dashed lines mark weight control.

We can also take into account the uncertainty about the states, i. e. how reliable the statement is that results from the combination of the possibly contradictory criteria. The uncertainty about the current state would then influence the stepsize, e. g. if we are not really sure whether double talk is taking place, the stepsize might be chosen smaller so as to reduce the damage in case the local speaker is active. The structure of this concept is shown in fig. 3. It reduces the propagation of errors, but consists of a more sensitive rule base which is also more difficult to verify. All the different structures can be shown in the space spanned

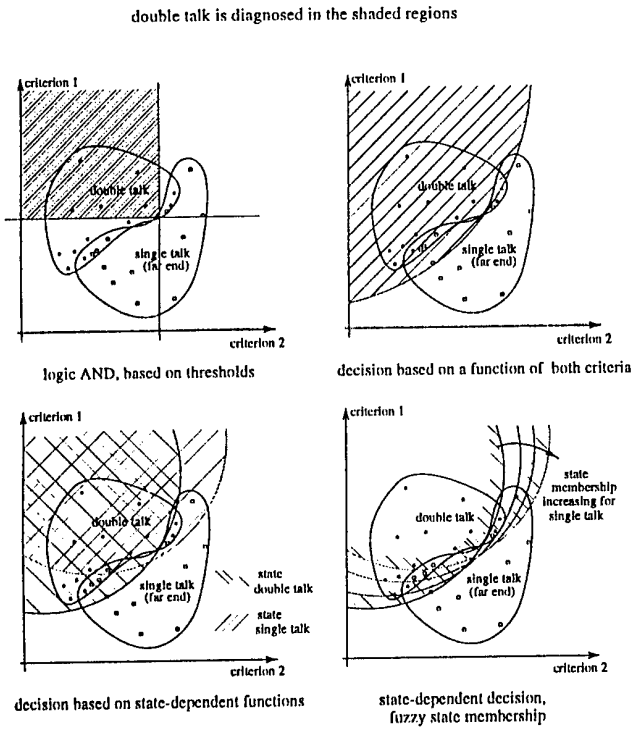


Figure 4: The different concepts in the space spanned by two criteria

by the criteria. This is done in fig. 4, with two criteria. Interpreting the results of the two methods as the coordinates of points, we will probably find that for a certain state points will mostly be found in a certain region, but we cannot be sure that those regions are well seperated from each other. The aim of the combination of criteria will be to construct a function that seperates the regions as well as possible, with regard to error probability and quality. Evidently, we will in general be more successful if more types of functions are allowed. The concept of state-dependent functions is then illustrated in the lower part of the figure.

3. RESULTS WHEN EMPLOYING THE CORRELATION METHOD

In this section we show some results that have been obtained by applying the different approaches to only one method, namely the correlation coefficient, with different parameter sets. The method was described in [2] as a double talk detector. To obtain a good estimation of the correlation coefficient between incoming and outgoing signal, one has to calculate it over as many samples as possible, but this will also lead to long delay for the double talk detection and is therefore unacceptable. The compromise used in the real-time implementation was then to set the number of samples to 72. When the correlation coefficient passes 0.9, single talk is assumed, and the stepsize is set to 1.0 in this paper. This concept is a special case of the state-space approach, wih one state, one criterium, and crisp logic, and will further be referred to as ρ_{72}. To extend this principle, we splitted the criterium: Since the correlation coefficient needs some time for the decision, due to the number of samples over which it is calculated, we reduced this number to 20, which makes it less reliable. We therefore used it in two forms: once in its original form so that it reacts fast (ρ_s), and once in a smoothed form obtained by an AR1 filter (ρ_l). Both criteria are worse than the one used before (see fig. 5 for the case of double talk) but contain different information which is combined by a fuzzy logic unit. The fuzzy logic block

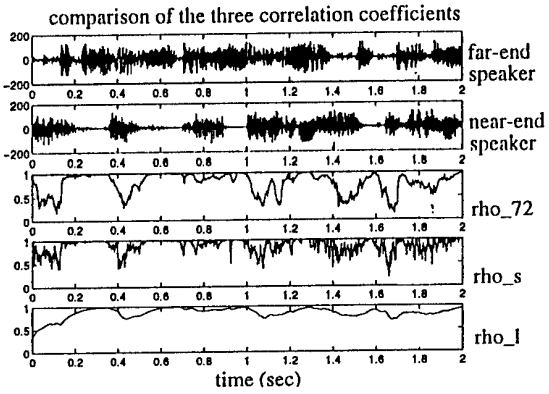


Figure 5: Comparison of correlation coefficients with three parameter sets

uses triangular membership functions, inference is realised by minimum and maximum (see [3]), and the defuzzification corresponds to a simplified "center of gravity", i .e a weighted average of the membership functions of the output:

$$\alpha = \alpha_{high}\mu_{\alpha_{high}} + \alpha_{med}\mu_{\alpha_{med}} + \alpha_{low}\mu_{\alpha_{low}} \qquad (6)$$

The rule base is shown below, and results are compared with those of the usual criterion in fig. 6.

IF ρ_s HIGH AND ρ_l HIGH, THEN α HIGH
IF ρ_s NOT HIGH AND ρ_l NOT HIGH OR IF ρ_s LOW, THEN α LOW
IF ρ_s NOT HIGH AND ρ_s HIGH OR IF ρ_l HIGH AND ρ_s MEDIUM, THEN α MEDIUM

Our second extension regards the number of states. In the simulations we observed that a relatively high threshold, in order to provide stability in double talk

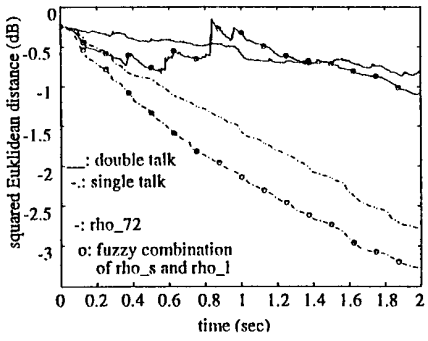


Figure 6: Comparison of the results with time-independent structures

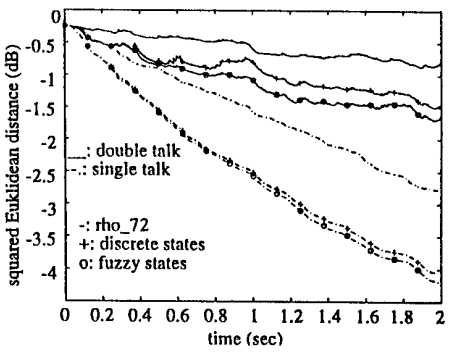


Figure 7: comparison of state-dependent approaches to the correlation method

situation, leads to slow convergence for single talk. So we try to infer the instantaneous state from the latest stepsize, i. e. if we calculate a large stepsize, we will assume single talk and thus facilitate large stepsizes for the next sample. It also means that in the case of a large stepsize for double talk, which is in itself an error, the damage will be amplified by encouraging more large stepsizes. Therefore we must be very certain that single talk is taking place when detecting it. Our new fuzzy logic unit now only contains different parameters for the membership functions and the defuzzification algorithm, but uses the same rule base as before. The states are determined by a state-dependent switch:

IF $\alpha < 0.5\alpha_{max}$ THEN SET DOUBLE TALK TRUE;
ELSE IF $\alpha > 0.8\alpha_{max}$ THEN SET DOUBLE TALK FALSE;

In the last step we transform the states into fuzzy numbers. The uncertainty about the state is thus integrated in the stepsize determination. The membership function of the state of double talk is calculated from:

$$\mu_{dt} = \mu_{dt} + \begin{cases} 0.6 - \frac{\alpha}{\alpha_{max}} & \alpha < 0.6\,\alpha_{max} \\ 0.8 - \frac{\alpha}{\alpha_{max}} & \alpha > 0.8\,\alpha_{max} \end{cases} \quad (7)$$

$$\text{with} \quad \mu_{dt} = \begin{cases} 1 & \mu_{dt} >= 1 \\ 0 & \mu_{dt} <= 0 \\ \mu_{dt} & 0 < \mu_{dt} < 1 \end{cases} \quad (8)$$

The state membership also influences the fuzzy numbers, as shown here for LOW:

$$LOW = \mu_{dt}\, LOW_{dt} + (1 - \mu_{dt})\, LOW_{st}$$

The results of the two state-dependent structures are shown in fig. 7. Both structures lead to better convergence than the state-independent approach, but the computational load and the optimization of the parameters become more important and more difficult to handle.

4. CONCLUSION

In this paper we presented several structures for stepsize control in adaptation algorithm. By using a generally optimized double talk detector and splitting it into two criteria with different parameter sets, we could extract more detailed information and combine it by state-dependent fuzzy logic. The results improved with every new extension of the structure. The number of parameters increases noticeably by including fuzzy logic, so that an optimization over a complete fuzzy state-dependent rule base might improve results even more, especially since better state determination will probably improve the results considerably for state-dependent control.

5. REFERENCES

[1] R. FRENZEL: *Freisprechen in gestörter Umgebung*, VDI-Fortschritt-Berichte, Reihe 10, Nr. 228, Düsseldorf, Germany, 1992

[2] P. HEITKÄMPER: *Optimization of an acoustic echo canceller combined with adaptive gain control*, Proc. ICASSP-95, Detroit, Michigan, 1995, pp. 3047 - 3050

[3] G. J. KLIR, B. YUAN: *Fuzzy Sets and Fuzzy Logic*, Prentice Hall PTR, New Jersey, 1995

[4] J. MARX: *Akustische Aspekte der Echokompensation in Freisprecheinrichtungen*, VDI-Fortschritt-Berichte, Reihe 10, Nr. 400, Düsseldorf, Germany, 1996

[5] U. SCHULTHEISS: *Über die Adaption eines Kompensators für akustische Echos*, VDI-Fortschritt-Berichte, Reihe 10, Nr. 90, Düsseldorf, Germany, 1988

DESIGN METHODOLOGY FOR VLSI IMPLEMENTATION OF IMAGE AND VIDEO CODING ALGORITHMS

Javier Bracamonte, Michael Ansorge and Fausto Pellandini*

Institute of Microtechnology, University of Neuchâtel
Rue A.-L. Breguet 2, 2000 Neuchâtel, Switzerland
Phone: +41 38 233210; Fax: +41 38 233201
Email: bracamonte@imt.unine.ch

ABSTRACT

The goal of this paper is to present a methodology for the design of VLSI circuits for image and video coding applications. The software environments for high/bit-true level simulations and hardware development are described. An example of an area efficient single-chip implementation of a JPEG coder is presented to illustrate the methodology.

1. INTRODUCTION

The basic steps of the methodology here reported are: a) high-level implementation of the algorithm b) VLSI architectures design c) bit-true level modeling d) data-path development and e) simulations. A brief description of the Joint Photographic Expert Group (JPEG) standard is given below to illustrate the procedure.

The JPEG standard describes an algorithm for the coding of continuous-tone still images [1]. It specifies four modes of operation: sequential, progressive, lossless and hierarchical. The sequential mode is by far the most used since it covers a wide range of applications and is hence the mode we address in this paper. The sequential (or baseline) mode is depicted in Figure 1. The algorithm is a DCT (Discrete Cosine Transform) -based process which transforms blocks of 8x8 pixels sequentially from the original image into 8x8 blocks of coefficients in the frequency domain [2]. The goal of this transformation is to decorrelate the original data and redistribute the signal energy among only a small set of transform coefficients in the low frequency zone. Based on psychovisual analysis, a normalization array can be defined. Its purpose is to quantize those DCT coefficients that are visually significant with relatively short quantization steps, while using large quantization steps for those coefficients which are less important. This large-step quantization associated with the energy packing effect of the transformation, results in general, in the zeroing out of many DCT coefficients. A long sequence of zero valued coefficients can then be efficiently abridged by runlength coding. Though the quantization is the main mechanism of data compression (and also of information loss), additional data compression can be obtained by entropy coding the output of the quantizer.

The high and bit-true level implementation of this algorithm are described in section 2 and 4 respectively, along with a description of the software environment. The VLSI architectures are discussed in section 3. Data-path development is discussed in section 5. Finally, the results and conclusions are given in section 6.

2. HIGH LEVEL IMPLEMENTATION

A modular high level implementation of the JPEG algorithm is shown in Figure 2. The top and bottom flow-graphs represent the encoder and decoder respectively. Each process of the algorithm was implemented with C code, using floating point precision for all the arithmetic operations. Each module was then converted into a Khoros routine of the Khoros software environment [3]. From a digital image processing point of view, this high-level implementation is an application program by itself [4]. Typical compression ratios are around 10, depending on the image activity and spatial resolution. A special toolbox (IMT TOOLBOX) was created and

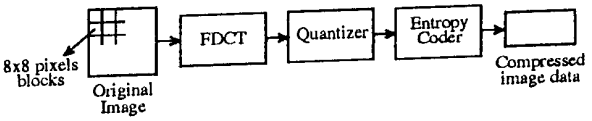


Figure 1: JPEG baseline algorithm

*This work was supported in part by the Laboratory of Microtechnology (LMT EPFL) common to the Swiss Federal Institute of Technology, Lausanne, and the Institute of Microtechnology, University of Neuchâtel.

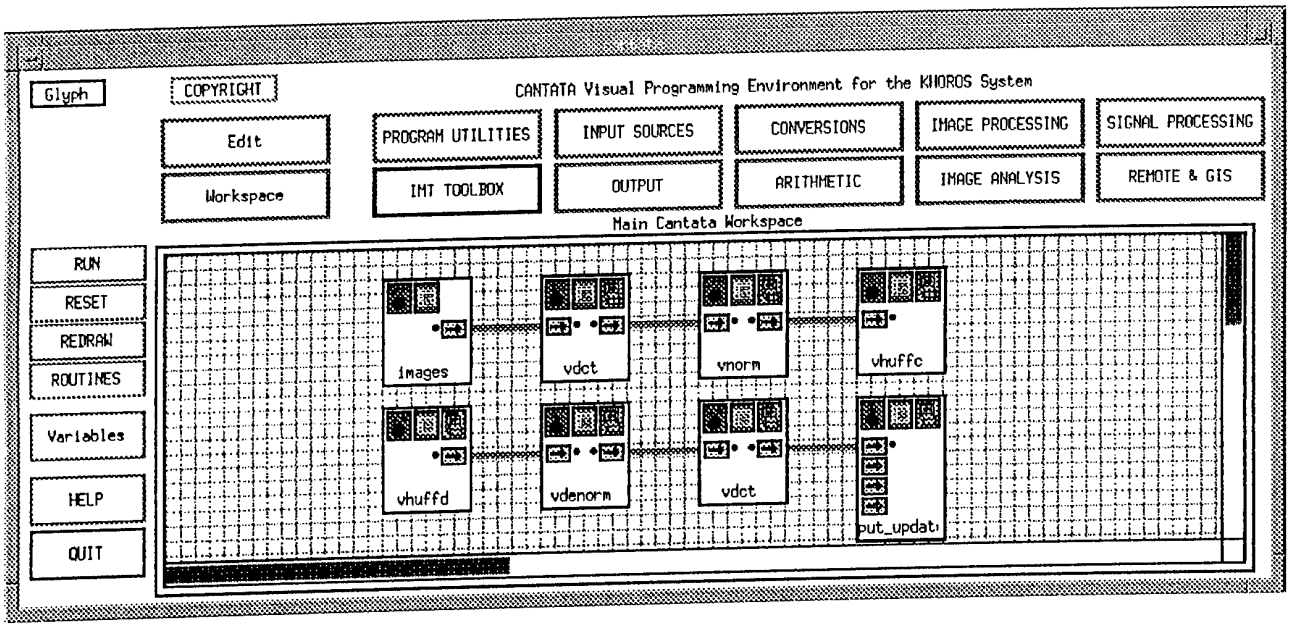


Figure 2: Implementation of JPEG in Khoros

integrated into the Khoros system. Besides the routines shown in Figure 2, this toolbox contains some mathematical functions and some programs for other methods of image compression (e.g., adaptive image compression).

3. ARCHITECTURES

Taking into consideration the constraints of the target application, a VLSI architecture for each block in Figure 2 was developed. Given their associated high data throughput, image and video coding applications require particularly high speed implementations. Using bit-serial architectures leads to tightly pipelined structures at the bit-level, which implies that a maximum clock-rate can be achieved [5]. Furthermore, bit-serial modules require less area than their parallel counterpart. Thus, whenever it was possible, we have chosen a full pipeline bit-serial approach. In the following paragraphs we describe the VLSI architectures for each of the main modules of the encoder in Figure 2.

a) FDCT: The forward 2-D DCT appropriate for JPEG is defined in [1] as:

$$S_{uv} = \frac{1}{4} C_u C_v \sum_{x=0}^{7} \sum_{y=0}^{7} s_{uv} \cos[(2x+1)u\pi/16] * \cos[(2x+1)v\pi/16];$$

for $u, v = 0, 1, \ldots, 7$, where $C_u, C_v = 1/2$ for $u, v = 0$; $C_u, C_v = 1$ otherwise. Given that the 2-D DCT is separable, it can be reformulated as two successive 1-D DCTs, which leads to a simpler hardware implementation. Each 1-D DCT is executed by a Distributed Arithmetic Processor (DAP) whose architecture is shown in Figure 3. Pre-addition/pre-subtraction operations can be used to exploit the symmetry of the 1-D DCT kernel. This results in halving the number of multiplications, or equivalently, in reducing the size of the DAP's ROM from 2^N to $2^{N/2}$ words [6]. The architecture also includes a transposition memory between the two 1-D transforms to store the results of the first 1-D DCT.

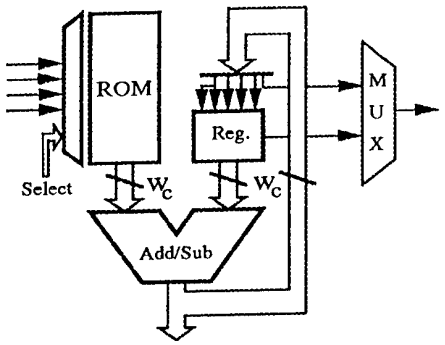


Figure 3: Distributed Arithmetic Processor

b) Quantizer: The second operation of the JPEG coder is the quantization of the 2-D DCT coefficients and is defined as: $Cq_{ij} = round(C_{ij}/N_{ij})$ for $i, j = 0, 1, ..., 7$, where C_{ij} denotes the ij-th element of an 8x8 DCT coefficients matrix, whereas N_{ij} denotes the

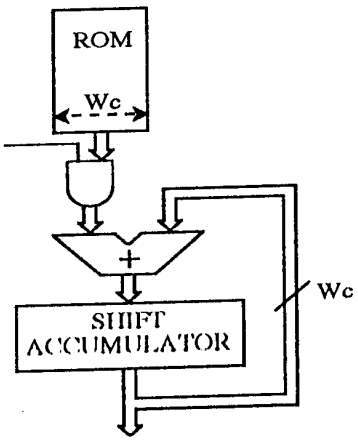


Figure 4: Architecture of the quantizer

ij-th element of an 8x8 normalization matrix. By reformulating the previous equation as $Cq_{ij} = round(C_{ij} * (1/N_{ij}))$ allows replacement of the rather complex circuitry of a divisor by that of a simple multiplier. The VLSI architecture of the quantizer is shown in Figure 4. Since the output of the 2-D DCT processor is bit serial, a serial-parallel multiplier was implemented. The parallel input to the multiplier being the output of a ROM containing the inverse of the normalization coefficients.

c) Entropy coder: The last operation of the JPEG algorithm is entropy coding. Its goal is to increase the compression performance of the encoder by taking advantage of the statistics of the symbols at the output of the quantizer. While both Huffman and arithmetic coders are supported by the JPEG standard, the baseline JPEG algorithm uses Huffman coding only. The circuit of the Huffman coder was based on [7]. Before assigning a variable length code to its input symbols, the Huffman coder must implement other operations, i.e., the runlength coding mentioned in the introduction, a category selection, etc. All these tasks were implemented with random logic and the Huffman table proposed in Annex K.3 in [1] was stored in a ROM of 2464 bits. A logic circuit packs the compressed bits into words of 32 bits at the output of the Huffman coder.

Between the quantizer and the entropy coder in Figure 1, some additional operations are defined by the JPEG standard: a) Raster to zigzag reordering: the reordering of the quantized 8x8 DCT coefficients 2-D array into a 1-D array, by order of increasing spatial frequency. This is implemented with a RAM with different read and write address sequences. b) DPCM: a DPCM coding of the DC DCT coefficients. This operation is implemented with a single subtractor and a register to store the prediction value. The latter operation is embedded at the input of the Huffman coder.

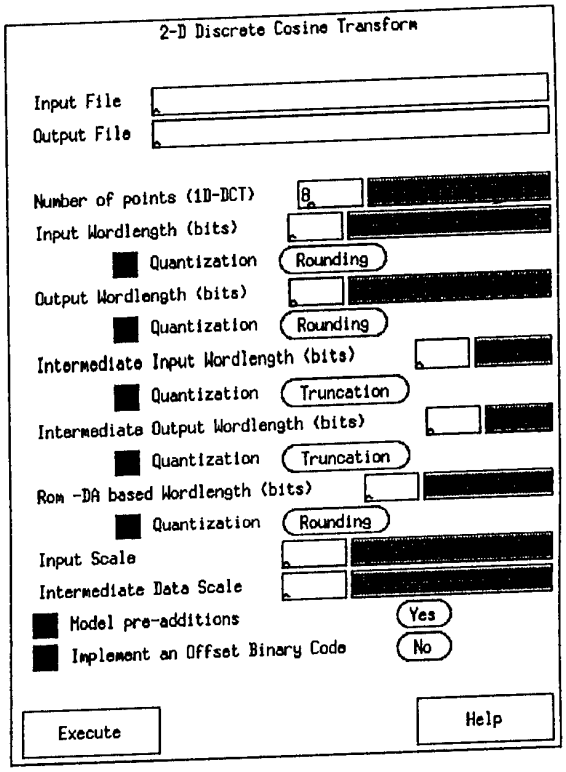


Figure 5: GUI of the 2-D DCT bit-true level model

4. BIT-LEVEL IMPLEMENTATIONS

Based on the programming framework of the high level implementation, a bit-true level model of the JPEG algorithm was built. Each process of the algorithm was again implemented with C code and converted into a Khoros routine. However this time, the arithmetic operations are all carried out in binary arithmetic, modeling accurately the same processing and dataflow as they would be executed by the corresponding processors and architectures described in section 3. The graphical user interface (GUI) corresponding to the bit-true level model of the 2-D DCT is shown in Figure 5. Both rounding and truncation can be modeled. The scale factors are required to fix the location of the binary point for each register. By using offset binary coding the number of words of the DAP's ROM can be reduced by a factor of two. Intermediate data refers to the data between the two 1-D DCTs. To analyze the quantization effects of a particular process, one simply substitutes the high-level implementation of that process in Figure 2, by its corresponding bit-true level model. Exhaustive simulations can then be executed to determine the optimum wordlength for the different signals and coefficients. The effects on the final compression ratio and reconstruction quality can also

be easily evaluated. This bit-true level reconfiguration and subsequent simulation is easily done by a few clicks on the Khoros visual language interface (Cantata).

For the 2-D DCT circuit, the optimum DAP's data- and ROM-wordlength found is 12 and 10 bits respectively. Simulations also have shown no difference of the results between using rounding or truncation, thus truncation mode was used. For the quantizer circuit, a ROM wordlength value of 9 bits and an input/output data wordlength of 12 bits were found. For these values, we obtained for several test images practically the same compression ratio and reconstructed image quality (Signal-to-Noise Ratio) as those obtained by using the floating point modules. The quantization table proposed in Annex K.1 of reference [1] was retained for our circuit.

5. DATAPATH DEVELOPMENT

For datapath development we use the tools of the Compass environment [8]. When a part of an algorithm is highly regular, the Layout editor tool is used to create full custom modules. This does not penalize the time for development since just a few cells must be designed, on the other hand, minimum silicon area and high speed are achieved. For less regular structures, the standard cell approach is used. The Cell Compiler tool is used to generate Random Access Memory modules. The circuit is built with the Logic Assistant tool. Intensive simulations are then carried out with both the Mixed-Mode and SPICE simulators.

6. RESULTS AND CONCLUSIONS

A semi-custom methodology for the VLSI implementation of image compression systems was reported. The software environments were described along with the an example of the implementation of a JPEG coder circuit. The area of the resulting chip is 4.6 x 3.1 mm^2 $\approx$ 14.5 mm^2 without pads. It was implemented in the 1.2μm CMN12 process from VLSI Technology Inc.. At a clock frequency of 36 MHz this circuit is able to process 25 CIF (Common Intermediate Format: 352 x 288 pixels) images per second. Thus it is suitable for motion JPEG (MJPEG) or for the non-recursive path of the H.261 low-bit rate video coder. Several improvements have been studied for this circuit, one being a power saving technique that trades image quality for power consumption [9].

Though image compression was addressed in this paper, the methodology could equally be applied to other kind of image processing applications. Intermediate results can also be used to develop solutions for other kind of technologies (DSP, FPGA, etc.). Current work involves the extension of the current methodology for the development of low-power image compression circuits for portable applications.

7. ACKNOWLEDGEMENTS

This work was supported by the Swiss National Science Foundation under Grant FN 2000-40'627.94, and by the Laboratory of Microtechnology (LMT EPFL). The latter entity is common to the Swiss Federal Institute of Technology, Lausanne, and the Institute of Microtechnology, University of Neuchâtel.

8. REFERENCES

[1] ITU-T Recommendation T.81 *Digital Compression and coding of continuous-tone still images.* September, 1992.

[2] K. R. Rao, and P. Yip, *"Discrete Cosine Transform: Algorithms, Advantages, Applications"*, Academic Press, Boston, MA, USA, 1990.

[3] D. Rasure, D. Arguiro, T. Sauer and C. William, "A visual language and software development environment for image processing", *Int'l J. of Imaging Systems and Technology*, Vol. 2, 1990, pp. 183-199.

[4] J. Bracamonte, *"A high and bit-true level implementation of the baseline JPEG image compression algorithm"*. Internal Report, Institute of Microtechnology, University of Neuchâtel, 1996.

[5] P. Denyer and D. Renshaw. *"VLSI signal processing: A bit-serial approach"*, Addisson-Wesley, VLSI System Series, 1985.

[6] U. Sjöström, *"On the design and implementation of DSP algorithms: An approach using wave digital state-space filters and distributed arithmetic"*. Ph.D Thesis, University of Neuchâtel, 1993.

[7] C. Henny, *"A VLSI implementation of a Huffman Coder"*, Diploma Project, University of Neuchâtel, August 1995.

[8] COMPASS Design Automation, Inc.: Manuals, COMPASS, San Jose, CA, USA, 1993.

[9] J. Bracamonte, M. Ansorge and F. Pellandini. "VLSI systems for image compression. A power-consumption/image-resolution trade-off approach". Accepted for publication. *Conference of Digital Compression Technologies & Systems for Video Communications.* Berlin, Germany, Oct. 7-11, 1996.

INVESTIGATION OF MODIFIED GOERTZEL ALGORITHM WITH APPLICATION TO DETECTION OF DTMF SIGNALS

Adam Dąbrowski and Tomasz Marciniak

Institute of Electronics and Telecommunications
Poznań University of Technology
ul. Piotrowo 3a
60-965 Poznań, POLAND
e-mail: dabrow@et.put.poznan.pl

ABSTRACT

This paper presents DTMF (dual tone multi-frequency signaling) detection using nonuniform digital filter bank. An algorithm for the detection uses a modified Goertzel algorithm [1]. It has been implemented in digital signal processor TMS320C50. To reduce the size of the analyzed block of samples, the varying block size is proposed, different for each of the DTMF frequency.

1. INTRODUCTION

The DTMF system for push-button telephone sets is a CCITT standard that appears as Recommendation Q.24 in CCITT Blue Book [2]. In DTMF signaling, each signal consists of a couple of sinusoidal signals with proper frequencies. These couples are allocated to the various digits and symbols of a touch-tone keypad. These frequencies belong to two mutually exclusive groups (the low frequency group and the high frequency group) of four frequencies each. The major requirements of DTMF system are as follows:

- frequencies of receiving signals:
  low frequency group: 697, 770, 852, 941 Hz,
  high frequency group: 1209, 1336, 1477, 1633 Hz,
  frequency tolerance: ± 1.8%.
- level of receiving signals for which the receiver works correctly: 0 ÷ -30 dBm,
- twist: ± 5 dB.
- time parameters:
  duration of a generated signal: min. 60 ms,
  duration of a break between signals of two consecutive digits: min. 60 ms
  maximum speed of signaling: one digit per 120ms,
  duration of signal recognition: max. 40 ms,
  duration of break recognition: max. 40 ms.

This work was supported partly by grant KBN-44-452 and partly by DPB-44-443/II

2. SPECTRAL ANALYSIS OF DTMF SIGNALS

Detection of DTMF signals is provided using discrete Fourier transform (DFT). Because only a subset of DFT output samples is needed, FFT (fast Fourier transform) algorithms are not in this case optimally effective.

Better approach would be a Goertzel type DFT algorithm which allows serial processing and requires smaller memory space and smaller number of computations. Goertzel algorithm is given by the graph in Fig.1.

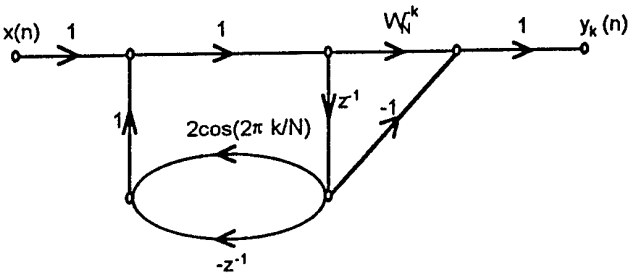


Figure 1. Goertzel algorithm

To reduce errors of allocation of center frequencies [4], the coefficients $\cos(2\pi k/N)$ can be replaced by $\cos(2\pi f_k/f_s)$, where f_k - DTMF frequency, f_s- sampling frequency, but the effect of this correction is not of primary importance.

Because the procedure of spectral analysis is provided using energies of eight DTMF sinusoidal components and their second harmonics, the graph of Goertzel algorithm reduces to the form in Fig.2. The basic algoritm step is then given by

$$w_k(n) = x(n) + 2\cos(2\pi f_k / f_s) w_k(n-1) - w_k(n-2) \quad (1)$$

On the end of the block of N samples, the energy can be computed as

$$\varepsilon_{kN} = |y_k(N-1)|^2 \qquad (2)$$

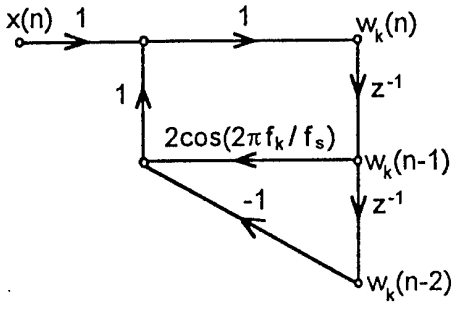


Figure 2. Goertzel algorithm for DTMF signal detection

3. NUMBER OF ANALYZED SAMPLES

Proper selection of the number N of analyzed samples is very important [3,5]. The number should be as small as possible for a correct analysis of duration a signal (or break). At 8 kHz sampling rate and 40 ms signal duration, the number N would be 160. But decreasing the number of samples, signal to noise ratio (SNR) decreases too. Second problem is the degradation of receiver frequency resolution and icreasing of the error of center frequency of filter passbands, provided that theoretical coefficients like in equation (1) are used. Fig.3 presents the maximum of the frequency error defined as

$$\delta_k = \frac{f_{kN} - f_k}{f_k} 100\% \qquad . \qquad (3)$$

where f_{kN} - frequency computed for the given number N,
f_k - nominal DTMF frequency.

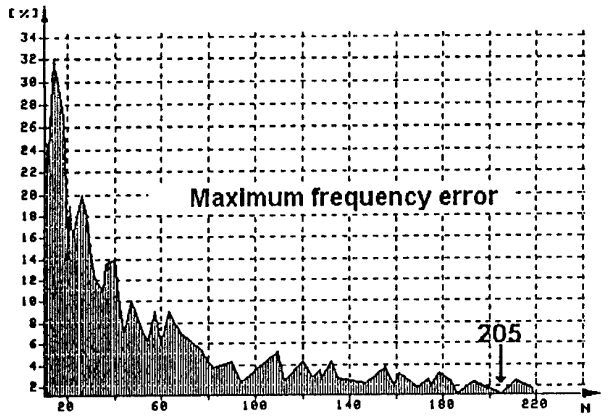


Figure 3. Maximum frequency error for all DTMF frequencies

At constant number N for all DTMF frequencies, the number N must be set to at least 205 (cf. Fig.3) . For N=205 the maximum frequency error for all DTMF frequencies equals 1.3%, thus is tolerably small.

To reduce the frequency error at smaller number N, selection of different numbers N depending on detected frequencies can be used. Error values for $80 \le N \le 85$ corresponding to the theoretical coefficients $\cos(2\pi k/N)$ computed according to equation (1) presents Table 1. Fig.4 also shows the DFT corresponding to digit "5" for different numbers N.

Table 1. Errors for varying block sizes N chosen optimally for individual frequencies in the neighbourhood of N=80

DTMF frequency [Hz]	N	δ_k [%]
697	80	0.43
770	83	0.14
852	85	-0.58
941	85	0.02
1209	80	-0.74
1336	84	-0.20
1477	81	0.30
1633	83	0.34

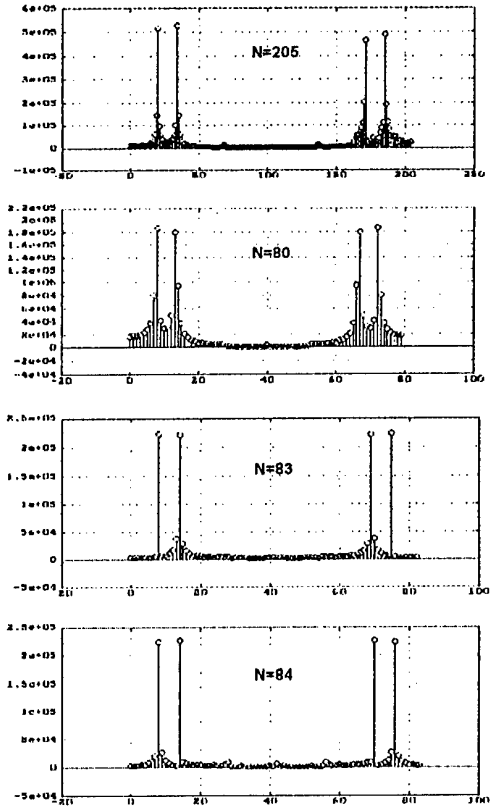


Figure 4. DFT of digit "5" for different numbers N

Number N=80 is very interesting because the time of analysis amounts 10 ms at 8 kHz sampling rate or 20 ms at 4 kHz sampling rate, i.e. one fourth or a half of the maximum recognition time, respectively. From Fig.4 follows that the magnitude of DFT for N=80 is much worse than for N=205. Energy levels between detection frequencies (770 Hz and 1336 Hz) and neighbourhood signals are 8 dB and 3 dB for N=205 and N=80 respectively. Therefore using the constant length N=80 is not possible. For interesting frequency we can find new better spectrum DFT, e.g. if N=83 for 770 Hz frequency or N=84 for 1336 Hz frequency. But this advantage can not be directly used because these frequencies are disturbed by high energies of DTMF frequencies in neighbourhood, if the optimal numbers of N from Table 1 are chosen for them. For the analysis of spectrum of DTMF signals energies for paricular Ns, e.g. $80 \le N \le 85$ (cf. Table 1) are summed and the analysis is provided for total energies.

4. PROGRAM FOR DTMF RECEIVER

The DTMF receiver program has been prepared in the assembler language of the TMS320C5x signal processor. The program for detection of DTMF signals starts with processor and analog interface initialization. After initialization, the program analyzes blocks of 85 samples according to equation (1). For each DTMF frequency, program saves six DFT samples for different numbers N in the range $80 \le N \le 85$. After taking samples the energies of DFT are computed and summed to get the overall energy ε_k given by

$$\varepsilon_k = \sum_{N=80}^{85} \varepsilon_{kN} \qquad (4)$$

where k is the index of the kth DTMF frequency and ε_{kN} is given by equation (2).

Next, the two frequencies (from low and high frequency group) with the highest energies are found. Then, checked if these energies are above the treshold and if the twist complies with requirements. Next, the energies of strongest signals are compared to the energies of the rest of tones in each group. Then, we compute energies of second harmonics corresponding to the strongest signals. These energies should be smaller than the energies of the first harmonics.

After all tests the program gives a digit which was recognized and goes to taking other samples from the next block.

5. CONCLUSIONS

We described an algoritm based on the proposed improved method for the detection of DTMF signals. Using a nonuniform Goertzel filter bank gives the posibility to reduce the number of blocks samples up to 85. The time of usage of the signal processor TMS320C50 for the mentioned algorithm is less than about 10%. Therefore, a simple processor may be used to receiving DTMF signals from multiple PCM channels.

6.REFERENCES

[1] A.Dąbrowski, T.Marciniak "Digital detection of DTMF signals"(in Polish), *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne* 9/1995.

[2] "Recommendation Q24", *CCITT Blue Book*, Genewa 1989.

[3] A.Dąbrowski, "Modified Goertzel algorithm for Detection of DTMF signals", *Proc. Int. Conf. on Signal and Systems, ICSS'94*, pp. IV.1-IV.3, Algiers, Algieria, Sept. 1994.

[4] B.Zheng, "Arbitrary allocation of center frequencies in signal detectors using the Goertzel algorithm", *Proc. Int. Conf. Signal Process. Appl. & Technology, ICSPAT'93*, pp.308-310, Santa Clara, USA, Sept. 1993.

[5] S.Bagchi, S.K. Mitra,"An efficient algorithm for DTMF decoding using the subband NDFT", *Proc. of the ISCAS '95*, pp.1936-1939, Seattle, USA, May 1995.

UVI_WAVE, THE ULTIMATE TOOLBOX FOR WAVELET TRANSFORMS AND FILTER BANKS

Nuria González Prelcic, Oscar W. Márquez and Santiago González

E.T.S.E. Telecomunicación, Universidade de Vigo
36200-VIGO (Spain)
email: nuria@tsc.uvigo.es

ABSTRACT

A wavelet software package named *Uvi_Wave* has been developed by the Signal Theory Group, in the University of Vigo, to provide a simple way to work with wavelets. So, the toolbox may be useful to make easier the understanding of theoretical concepts. Moreover, it provides an experimentation platform for wavelet applications. This paper describes the contents and main features of *Uvi_Wave*, that has been implemented within the powerful Matlab environment and is freely distributed.

1. INTRODUCTION

The availability of software tools make possible for students and researchers to explore and test quickly the possibilities of new methods. This is very important when dealing with emergent techniques, like wavelets, in the digital signal processing area. Moreover, this allows the user to concentrate on the applications, since it provides the basic algorithms and utilities for experimentation and further wavelet based developments.

Wavelets can be used to solve very different problems that appear in many areas of electrical engineering, such as speech, audio and image coding, computer graphics, communications, numerical analysis, statistics, etc [1, 2]. The enthusiasm with which wavelets have been accepted in so many fields, makes interesting the development of a complete set of tools. With this aim, we present here a wavelet toolbox that can be used as general package for research and educational projects.

2. UVI_WAVE WAVELET TOOLBOX

Uvi_Wave provides a set of Matlab command line functions and demonstrations to analyze and synthesize signals by means of wavelets and wavelet packets. It also includes tools for the design and test of two-channel filter banks. The different routines allow an exhaustive and advanced treatment of wavelet based techniques in a simple and powerful enviroment.

This work was partially supported by the University of Vigo.

2.1. IMPLEMENTATION

We have used Matlab as implementation platform of the tools described in the next section. Matlab has been chosen because it is well known for researchers, engineers and students, and the routines can run on many platforms with minimum changes.

The current version of the toolbox includes 121 matlab functions and 12 data files with sample signals and filters. The Matlab code is compatible with versions up to 4.0, and all the routines have been tested on Unix and Windows platforms.

All the functions offer on-line help, with descriptions of the algorithms and hints to use them. Moreover, a complete reference manual is available. The functions are indexed, cross-referenced, and their input/output arguments are explained in detail. To illustrate the usage and application of the diferent functions the manual introduces some examples, together with an algorithm description if required.

3. STRUCTURE OF THE TOOLBOX

In the next sections, a short description of the main functions of the *Uvi_Wave* Toolbox is presented.

3.1. DISCRETE WAVELET TRANSFORM

Calculation and displaying of the Discrete Wavelet Transform (DWT) and Inverse Discrete Wavelet Transform (IDWT) for unidimensional or bidimensional signals [3]. Table 1 briefly references these functions. There is no restrictions on the signal length. For avoiding border effects when reconstructing finite-length signals, a periodic extension method is used [4]. Figure 1 shows

the 2-stage wavelet transform of a square with an inscribed cross image.

Function	Purpose
wt	1D Discrete Wavelet Transform
iwt	1D Inverse Wavelet Transform
wt2d	2D Discrete Wavelet Transform
iwt2d	2D Inverse Wavelet Transform

Table 1: *Discrete Wavelet Transform*

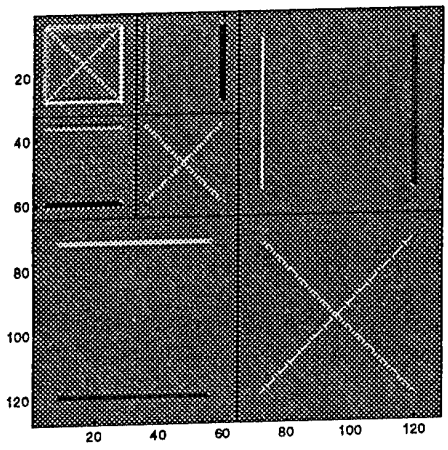


Figure 1: Wavelet decomposition with 2 levels.

3.2. SCALE AND WAVELET FUNCTION

Calculation and displaying of the discrete approximation to the scale and wavelet functions. Their computation is based on the cascade algorithm [5]. Same manner, any basis function in the wavelet packet basis library can be computed. Table 2 lists the functions.

Function	Purpose
wavelet	Wavelet and Scale functions
wavepack	Wavelet Packet functions calculation

Table 2: *Continuous wavelet functions*

3.3. SCALOGRAM

Computation of the Continous Wavelet Transform coefficients can be performed using the functions in table 3. The representation of its modulus above a time-scale plane (scalogram) is a well suited tool for signal analysis. The Morlet wavelet is used as prototype function.

Figure 2 presents the scalogram of a rectangular pulse obtained using scalog.

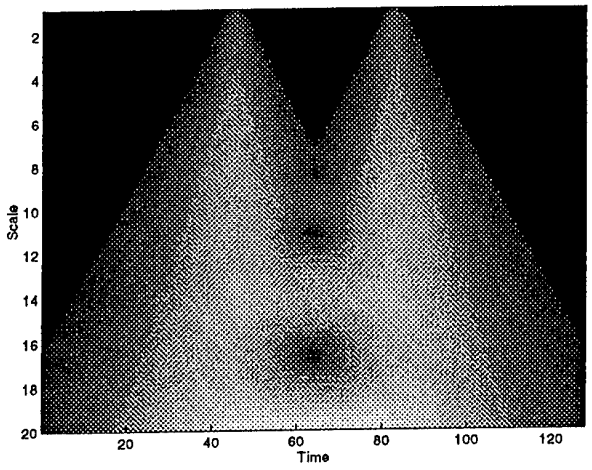


Figure 2: Scalogram of a rectangular pulse.

Function	Purpose
morletw	Morlet Wavelet calculation
scalog	Scalogram computation
srf	3D plot of the scalogram

Table 3: *Continous Wavelet Transform*

3.4. MULTIRESOLUTION ANALYSIS

The objective of these routines is to obtain a representation of the signal at different resolutions. So, approximation and detail signals at different scales are computed. There are functions for both unidimensional and bidimensional signals, as table 4 presents.

Function	Purpose
aprox	1D approximation signals
detail	1D detail signals
multires	Complete 1D multiresolution analysis
mres2d	2D approximation/detail image
nssffb	Non sub-sampled filter bank
inssffb	Inverse non sub-sampled filter bank
nss2d	2D non sub-sampled filter bank
inss2d	Inverse 2D non sub-sampled filter bank

Table 4: *Multiresolution analysis*

Other routines in the table implement the FIR filter bank structure performing the wavelet transform, without decimation. It can be mainly used for time or spatial singularity detection. They have been implemented for 1-D and 2-D signals too.

Figure 3 displays the original and the approximation signals (for 2^1 to 2^4 scale) obtained with aprox.

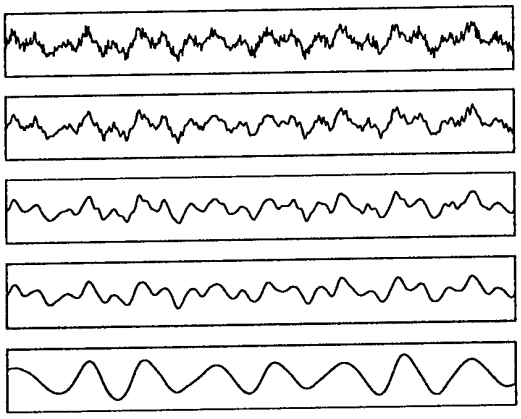


Figure 3: Original and approximation signals.

3.5. WAVELET PACKETS

Discrete-time algorithms for the Wavelet Packet Transform are implemented, with the same capabilities as the Wavelet Transform. Two formats are allowed to encode the basis (i.e., the filter bank tree) for unidimensional transforms: natural and frequency order [6]. Some tools to manage the different formats have been included. Table 5 lists most of these routines, together with those of the next section.

Moreover, the viewing utilities allow the user to plot the time-frequency plane tiling performed by a certain wavelet packet basis, or the corresponding filter bank tree scheme.

3.6. BASIS SELECTION ALGORITHMS

In addition to the wavelet packet transform for unidimensional or bidimensional signals, different basis selection algorithms have been included, with additive [6, 7] and non-additive [8, 7] cost functionals.

Basis selection algorithms for 1-D signals comprise pruning [7] and growth [9] types. Both of them have a different implementation according to the additiveness of the used cost function. For 2-D signals, only pruning algorithm with additive costs (*best basis* algorithm [7]) has been implemented. All these functions are listed in table 5.

3.7. FILTER BANK DESIGN

A set of algorithms that yield several filter families have been implemented. The selected design algorithms include orthogonal [10, 5, 11, 12] and biorthogonal families [5], as presented in table 6. Some functions to compute filter regularity estimates are also included.

Function	Purpose
`wpk`	Wavelet Packet Transform
`iwpk`	Inverse Wavelet Packet Transform
`wpk2d`	2D Wavelet Packet Transform
`iwpk2d`	2D Inverse Wavelet Packet Transform
`pruneadd`	Pruning algorithm for additive costs
`prunenon`	Pruning for non-additive costs
`growadd`	Growth algorithm for additive costs
`grownon`	Growth algorithm for non-additive costs
`prune2d`	Quadtree pruning for additive costs
`lpenerg`	Energy with l^p norm
`shanent`	Shannon entropy
`logenerg`	'Log energy' functional
`cwent`	Coifman-Wickerhauser entropy
`weaklp`	Weak l^p norm
`cmparea`	Compression area
`cmpnum`	Compression number

Table 5: *Wavelet Packets*

Function	Purpose
`wspline`	Spline biorthogonal filters
`daub`	Daubechies orthogonal filters
`symlets`	Least-asymmetric orthogonal filters
`maxflat`	Maximally flat orthogonal filters
`lemarie`	Battle-Lemarié orthogonal filters
`remezflt`	Remez solution orthogonal filters
`tempreg`	Hölder regularity temporal estimate
`specreg`	Regularity spectral estimate
`regdaub`	Hölder regularity for Daubechies filters

Table 6: *Wavelet filters generation*

3.8. SUBBAND MANAGEMENT UTILITIES

Utilities for locating, extracting or inserting any subband are provided. These functions allows to process separately the content of a single subband, if desired. All of them works with wavelet and wavelet packet transforms, and with any signal size. For the wavelet case, maxima extraction, minima deletion or local extrema extraction can be performed, too. Table 7 lists most of these utilities.

3.9. DEMONSTRATION SCRIPTS

In addition, some demos and test signals have been included, illustrating the main features and capabilities of *Uvi_Wave*, how the functions are called and some applications. All the demonstration functions can be accessed via the user friendly menu shown in figure 4.

All the messages included have an educational goal, introducing short explanations of the main concepts shown in the demos. Figure 2 shows the scalogram of

Function	Purpose
`bandsite`	WT Subband localization
`bandext`	WT Subband extraction
`bandins`	WT Subband insertion
`bandmax`	WT Subband maxima extraction
`elmin`	WT Minima removal
`localext`	WT Local extrema extraction
`bandadj`	2D Subband normalization
`siteband`	Wavelet Packet subband localization
`extband`	Wavelet Packet subband extraction
`insband`	Wavelet Packet subband insertion

Table 7: *Subband management*

a rectangular pulse obtained during the demo. Besides these demos, some scripts help on 1-D basis formats and 2-D transform output format.

3.10. AVAILABILITY

Uvi_Wave has been written with an educational and research assistant goal. So, it is freely distributed, under GNU public license, through the Internet. It is available at the Communication Technologies Department ftp server (ftp.tsc.uvigo.es), and there is a WWW related page with information and links to the toolbox (http://www.tsc.uvigo.es/~wavelets/Uvi_Wave.html).

Around 600 copies of the latest version of *Uvi_Wave* have been downloaded by anonymous ftp. Furthermore, about 100 people are participating in a mailing list for general discussions on wavelets, reporting about the toolbox performance, and help on using it.

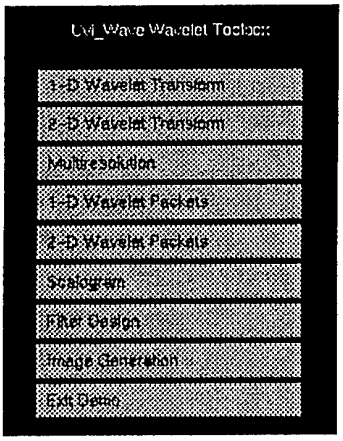


Figure 4: Menu for the main demo function.

4. CONCLUSIONS

A comprehensive and easy to use toolbox for understanding wavelets and researching on their applications has been presented. It can be used in a wide range of platforms since it has been implemented in Matlab. The experience along the last two years has shown its usefulness in a large variety of contexts. Further developments comprises Cosine Packets, additional wavelet filter types, speeding-up some functions by means of *.mex* routines, graphical interactive tools, etc.

5. REFERENCES

[1] M. Vetterli and J. Kovačević. *Wavelets and Subband Coding.* Prentice Hall, 1995.

[2] A. N. Akansu and M. J. Y Smith. *Subband and Wavelet Transforms: Design and Applications.* Kluwer Academic Publishers, 1996.

[3] Y. Meyer. *Wavelets, Algorithms and Applications.* SIAM, 1993.

[4] K. C. McGill and C. Taswell. Wavelet transform algorithms for finite-duration discrete-time signals. Technical report, Veterans Affairs Medical Center, Palo Alto, CA, 1991.

[5] I. Daubechies. *Ten Lectures on Wavelets.* SIAM, Philadelphia, 1992.

[6] M. V. Wickerhauser. Lectures on wavelet packets algorithms. Minicourse lecture notes, INRIA, Roquencourt, France, 1991.

[7] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithm for best basis selection. *IEEE Transactions on Information Theory,* 38(2):713–718, March 1992.

[8] C. Taswell. Near-best basis selection algorithms with non-additive information cost functions. In *Proceedings of IEEE International Symposium on Time-Frequency and Time-Scale Analysis,* 1994.

[9] C. Taswell. Top-down and bottom-up tree search algorithms for selecting bases in wavelet packet transforms. In *Wavelets and Statistics,* pages 345–359. Springer Verlag, 1995.

[10] G. Battle. A block spin construction of ondelettes. Part I: Lemarié functions. *Commun. Math. Phys.,* pages 601–615, 1987.

[11] O. Rioul and P. Duhamel. A remez exchange algorithm for orthonormal wavelets. *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing,* 41(8), August 1994.

[12] P. P. Vaidyanathan. *Multirate Systems and Filter Banks.* Prentice Hall, 1993.

Wavelet Networks for Multi-Scale Non-linear System Modeling

Aweke N. Lemma, Ed F. Deprettere
Department of Electrical Engineering
Delft University of Technology, 2628 CD Delft
emails: aweke@cas.et.tudelft.nl/ed@cas.et.tudelft.nl
Tel. +31-2786465, Fax. +31-2786190

Abstract— **Using wavelet networks, it is possible to capture the characteristics of non-linear dynamic systems in a multi-scale strategy. Starting from the coarsest approximation (coarsest scale) we go step-wise to the finer scales. At each step the error signal or the residue of the system is modeled. This procedure is repeated until the residual drops below some modeling error bound. The modeling is carried out using compactly supported biorthogonal wavelets. By choosing appropriate wavelet basis, it is possible to obtain a near optimal model.**

I. Introduction

In system modeling, it is required to build a mathematical model for an unknown system based on a finite set of observed data. This is the case in many applications such as channel equalization, plant identification, signal compression and so forth. The model uses the observed data to capture the features of the system, and when excited with a new set of input signals, it is expected to follow the system behavior closely. The fundamental issue is to keep the size of the model as small as possible.

Many solution methods are proposed in the literature. Recent developments are the use of neural networks and fuzzy systems [1], [2], [3]. The dominant problem in all of these approachs is the excessive size of the model which makes the update algorithms complex and slow.

In this paper we present networks that use wavelets as source of the non-linearity. The method uses the so-called multi-scale modeling. Owing to the existence of biorthogonal wavelet pairs, the weight vectors are determined by an inner product technique. This makes the update algorithm simple and fast.

The use of wavelet networks for non-linear system modeling is not a new concept, it has been discussed in [4], [5] and [6]. However, in these works, the multi-resolution aspect of the wavelet networks is not well considered. Here, on the other hand, we generate a more systematic modeling strateggy by making use of the extra behavior offered by the multi-resolution nature of the wavelet network. This is done through the so-called multi-scale modeling. In multi-scale modeling, the problem is subdivided into several sub-problems, and the modeling is carried out starting from the coursest approximation and then proceeding to the finest scale. The advantage of this approach over the usual fine to course analysis is that at each finer scale the model approximates the system at increasing resolution. This is equivalent to first capturing the most general feature of the system and systematically proceeding to its detailed characteristics.

II. Problem Definition

Let be given a collection of observed inputs $\mathbf{u}_t = \{u_t\}$ and the corresponding collection of outputs $\mathbf{y}_t = \{y_t\}$, $t = 0, 1, 2, \ldots$, of unknown non-linear discrete-time multivariable dynamical system of the form

$$\mathbf{y}_{t+1} = \mathbf{f}(\mathbf{u}_t, \mathbf{y}_t), \tag{1}$$

where $\mathbf{u}_t$ is an n-dimensional input vector and $\mathbf{y}_t$ an m-dimensional output vector;

$$\mathbf{u}_t = \left[\begin{array}{cccc} u_t & u_{t-1} & u_{t-2} & \cdots & u_{t-n+1} \end{array}\right],$$

$$\mathbf{y}_t = \left[\begin{array}{cccc} y_t & y_{t-1} & y_{t-2} & \cdots & y_{t-m+1} \end{array}\right].$$

Assume that $\mathbf{u}_t$ and $\mathbf{y}_t$ are confined within a compact region $\mathcal{C}$ of an arbitrary shape. We intend to construct a wavelet network model for (1) over $\mathcal{C}$ using the so-called multi-scale strategy to meet a prescribed modeling error bound. For the sake of convenience, we define an $l = (n+m)$-dimensional vector $\mathbf{x}_t = \left[\begin{array}{cc} \mathbf{u}_t & \mathbf{y}_t \end{array}\right]^T$ such that (1) becomes $\mathbf{y}_{t+1} = \mathbf{f}(\mathbf{x}_t)$. Note that for a static system $\mathbf{x}_t = u_t$.

It could be shown that, [7], the non-linear modeling problem of (1) is well-defined, if it is such that $\mathbf{f}(.) \in L^2(\mathcal{R}^n)$. We consider systems under ths category. Note that the above condition is easily met if $\mathbf{f}(.)$ is continuous over the compact region $\mathcal{C}$.

III. Multi-scale Modeling

The multi-scale modeling is inherent in the wavelet network, it utilizes the multi-resolution nature of signal expansion using wavelets. Multi-resolution analysis is well discussed in many papers and standard wavelet texts. In nearly all applications, the analysis is performed from fine to course resolution (scale). In system modeling, however,

where we have global and local features to be captured by the model, it is evident that a more natural way of learning would be to go from course to fine scale. In doing so, we can gradually localize global feature. This is equivalent to the well known process of zooming-in.

The advantage of the global/local approach is that, at each finer scale, we model the residual of the previous scales. We proceed with the course to fine approach, each time modeling the residual, until the modeling error falls below a certain error bound. If large modeling error bound is tollerable, then we only have to capture the global features and ignore the local ones.

A. The Modeling

Let φ and ψ represent the analysis and synthesis biorthogonal wavelet pairs, respectively. We want to model the system transfer $\mathbf{y}_{t+1} = \mathbf{f}(\mathbf{x}_t)$ of (1) using the linear combination of the scaled and translated version of the synthesis mother wavelet $\psi(\mathbf{x}_t)$[1] as

$$\mathbf{f}(\mathbf{x}_t) = \sum_{j,k_\mu} \mathbf{c}_{j,k_\mu} \psi_{j,k_\mu}(\mathbf{x}_t), \tag{2}$$

where $\psi_{j,k_\mu} = 2^{j/2}\psi(2^j\mathbf{x}_t - k_\mu)$, and the m-dimensional coefficient vector $\mathbf{c}_{j,k_\mu}$ is obtained from[2]

$$\mathbf{c}_{j,k_\mu} = \langle \mathbf{f}(\mathbf{x}_t), \varphi_{j,k_\mu}(\hat{\mathbf{x}}_t) \rangle, \tag{3}$$

with the scaling factor $j = 0, 1, 2, \ldots$ and the translation factor $k_\mu \in \{k_1, k_2, \ldots, k_{n_j}\}$, n_j denoting the number of wavelet nodes at scale j.

A.1 More about $\psi(x)$

Given an excitation signal $x(t)$ and a wavelet basis ψ, $\psi(x)$ is a reordered version of part of the samples of ψ and, therefore, can be expressed as

$$\psi(x) = \psi P_x, \tag{4}$$

where P_x represents a sort of "permutation" matrix. It may niether be full rank nor square. If it is full rank, then we say $x(t)$ visits all the samples of ψ, and we can perfectly reconstruct ψ from $\psi(x)$. On the other hand, if P_x is not full rank, then $x(t)$ does not visit all the samples of ψ. However, if the missing samples constitute a minor portion of the total energy, we can still recover ψ with a resonably good precition.

A.2 Determining the biorthogonal pair for $\psi(x)$

Consider the biorthogonal wavelet pair (φ, ψ). From biorthogonality, we know that the inner product $\varphi\psi^T = I$. this means that φ and ψ form a biorthogonal set. Assume that we excite ψ with a signal $x(t)$ such that, using (4), we obtain $\psi(x) = \psi P_x$. If P_x is full rank, then we can find $\hat{P}_x$ such that $\hat{P}_x P_x^T = I$. Let us now generate a secondary excitation signal $\hat{x}(t)$ that excites the analysis wavelet nodes such that $\varphi(\hat{x}) = \varphi\hat{P}_x$. Then we have $\varphi(\hat{x})\psi(x)^T = I$. This means that the biorthogonal wavelet pair φ and ψ remain that way even after they are excited with the signals $\hat{x}(t)$ and $x(t)$ respectively. Which means, we can use (2) to approximate $\mathbf{f}$, where the weight coeficients are given by (3), and $\hat{\mathbf{x}}_t$ is such that, with $\varphi(\hat{\mathbf{x}}_t) = \varphi\hat{P}_{\mathbf{x}_t}$ and $\psi(\mathbf{x}_t) = \psi P_{\mathbf{x}_t}$, $\hat{P}_{\mathbf{x}_t} P_{\mathbf{x}_t}^T = I$[3].

B. The wavelet network

The wavelet network, as shown in Fig.1, is a three layer feed forward network. The first layer contains the wavelet nodes, the second layer the linear nodes and the third layer the summation nodes. To give a clearer picture, the detail of the wavelet network at scale j is presented in Fig.2.

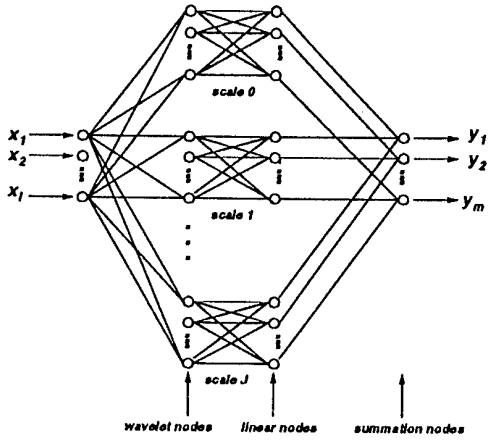


Fig. 1. *The structure of the wavelet network*

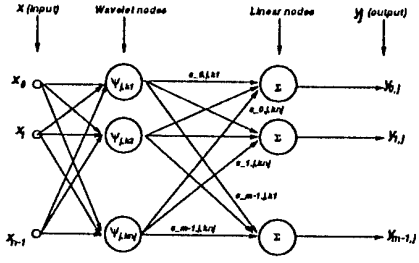


Fig. 2. *The wavelet network at scale j*

B.1 The wavelet nodes

The $\mathcal{R}^n \rightarrow \mathcal{R}$ activation function $\psi_{j,k_\mu}(\mathbf{x}_t)$ of the wavelet nodes in the first layer are generated by dilating

[1] To be defined soon, in the next sub section

[2] Note that we use a different excitation signal $\hat{\mathbf{x}}_t$ for the analysis wavelet. This will be clear soon.

[3] This relation is properly shown in proposition III-B.1

and translating a predetermined mother wavelet $\psi(\mathbf{x}_t)$. With $\mathbf{x}_{p,t}$ representing the p^{th} row of $\mathbf{x}_t$, $\psi(\mathbf{x}_t)$ is defined as

$$\psi(\mathbf{x}_t) = \prod_{p=0}^{l-1} \psi(\mathbf{x}_{p,t}), \tag{5}$$

and its biorthogonal complement $\varphi(\hat{\mathbf{x}}_t)$ as

$$\varphi(\hat{\mathbf{x}}_t) = \prod_{p=0}^{l-1} \varphi(\hat{\mathbf{x}}_{p,t}). \tag{6}$$

Proposition *III-B.1: The wavelet functions $\psi(\mathbf{x}_t)$ and $\varphi(\hat{\mathbf{x}}_t)$ defined in (5) and (6) form a biorthogonal wavelet pair in $L^2(\mathcal{R}^l)$ if*

1. $\mathbf{x}_{p,t}$, the p^{th} row of $\mathbf{x}_t$, visits all the samples of ψ for all $p = 0, 1, \ldots, l-1$, and

2. The wavelet families

$$\mathbf{S}_j = [\Psi_0(\mathbf{x}_t) \;\; \Psi_1(\mathbf{x}_t) \;\; \ldots \;\; \Psi_j(\mathbf{x}_t)]^T$$

and

$$\mathbf{A}_j = [\Phi_0(\hat{\mathbf{x}}_t) \;\; \Phi_1(\hat{\mathbf{x}}_t) \;\; \ldots \;\; \Phi_j(\hat{\mathbf{x}}_t)]^T$$

are such that

$$\mathbf{S}_j^T \mathbf{A}_j = I \tag{7}$$

for $j = 0, 1, \ldots$, where $\Psi_j(\mathbf{x}_t)$ and $\Phi_j(\hat{\mathbf{x}}_t)$ are defined as

$$\Psi_j(\mathbf{x}_t) = \left[\psi_{j,k_1}(\mathbf{x}_t) \;\; \psi_{j,k_2}(\mathbf{x}_t) \;\; \ldots \;\; \psi_{j,k_{n_j}}(\mathbf{x}_t)\right]^T \tag{8}$$

$$\Phi_j(\mathbf{x}_t) = \left[\varphi_{j,k_1}(\mathbf{x}_t) \;\; \varphi_{j,k_2}(\mathbf{x}_t) \;\; \ldots \;\; \varphi_{j,k_{n_j}}(\mathbf{x}_t)\right]^T \tag{9}$$

Note that condition (7) in the proposition implies that for $j = 0, 1, 2, \ldots$ the wavelet famly pair $\mathbf{S}_j$ and $\mathbf{A}_j$ form a biorthogonal basis in a subspace $V_j \subseteq V$, where V is the space in which the signal $\mathbf{f}(\mathbf{x}_t)$ is defined.

Proof: We avoid detailed mathematics and try to outline the proof on an abstract level. Let $\psi(\mathbf{t})$ and $\varphi(\mathbf{t})$ represent $\mathcal{R}^l \rightarrow \mathcal{R}$ multivariable biorthogonal wavelet pair. Note that $\psi(\mathbf{t})$ and $\varphi(\mathbf{t})$ could be be obtained using (5) and (6), respectively. Then using (4), $\psi(\mathbf{x}_t)$ and $\varphi(\hat{\mathbf{x}}_t)$ can be expressed as

$$\begin{aligned} \psi(\mathbf{x}_t) &= \psi(\mathbf{t}) P_{\mathbf{x}_t} \\ \varphi(\hat{\mathbf{x}}_t) &= \varphi(\mathbf{t}) \hat{P}_{\mathbf{x}_t} \end{aligned}$$

where $P_{\mathbf{x}_t}, \hat{P}_{\mathbf{x}_t}$ are permutation operators in appropriate spaces. If $\mathbf{x}_{p,t}$ visits all the samples of ψ for all $p = 0, 1, \ldots, l-1$, then $P_{\mathbf{x}_t}$ is full rank, and we can find $\hat{P}_{\mathbf{x}_t}$ such that $P_{\mathbf{x}_t} \hat{P}_{\mathbf{x}_t}^T = I$. With this choice of $\hat{P}_{\mathbf{x}_t}$, we have

$$\begin{aligned} \psi(\mathbf{x}_t)\varphi((\hat{\mathbf{x}}_t)^T &= \psi(\mathbf{t}) P_{\mathbf{x}_t} \hat{P}_{\mathbf{x}_t}^T \varphi(\mathbf{t}) \\ &= \psi(\mathbf{t})\varphi(\mathbf{t})^T = I \end{aligned} \tag{10}$$

■

B.2 The linear node

The second layer in the wavelet network is a linear node, it performs the computation[4]

$$y_{i,j} = \sum_{\mu} c_{i,j,k_\mu} \psi_{j,k_\mu}(\mathbf{x}_t), \;\; i = 0, 1, \ldots, m-1 \tag{11}$$

Or, letting $\mathbf{y}_j = [y_{0,j} \;\; y_{1,j} \;\; \ldots \;\; y_{m-1,j}]^T$ (see Fig. 2), we can re-write the last equation as

$$\mathbf{y}_j = C_j \Psi_j(\mathbf{x}_t), \tag{12}$$

where the $m \times n_j$ weight matrix C_j is

$$C_j = \begin{bmatrix} c_{0,j,k_1} & c_{0,j,k_2} & \cdots & c_{0,j,k_{n_j}} \\ c_{1,j,k_1} & c_{1,j,k_2} & \cdots & c_{1,j,k_{n_j}} \\ \vdots & \vdots & \vdots & \vdots \\ c_{m-1,j,k_1} & c_{m-1,j,k_2} & \cdots & c_{m-1,j,k_{n_j}} \end{bmatrix} \tag{13}$$

and $\Psi_j(\mathbf{x}_t)$ is the n_j-wavelet vector given in (8).

B.3 The summation node

The transfer function of the third and final layer is given by

$$\tilde{y}_i = \sum_j y_{i,j}, \tag{14}$$

where $\tilde{y}_i, \;\; i = 0, 1, \ldots, m-1$ is the i^{th} entry of $\tilde{\mathbf{y}}_{t+1}$, which in turn is the model approximation for $\mathbf{y}_{t+1}$. Thus, combining (12) and (14), we generate the transfer function of the wavelet network model for (1).

$$\tilde{\mathbf{y}}_{t+1} = \tilde{\mathbf{f}}(\mathbf{x}_t) = \sum_{j=0}^{J} C_j \Psi_j(\mathbf{x}_t). \tag{15}$$

The wavelet network described by (15) is a universal approximator. This is the consequence of the so-called multi-resolution behavior of wavelet expansion of signals, which is discussed in [8] and stated in the following lemma.

Lemma III-B.1: Any non-linear system of (1) can be approximated arbitrarily close by (15) for some large enough scale J. i.e. Given an arbitrary error bound ϵ, there exists a scale J such that

$$\|\mathbf{f}(\mathbf{x}_t) - \tilde{\mathbf{f}}(\mathbf{x}_t)\| < \epsilon.$$

C. The procedure

Let s input-output data points $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_{t+1}\}$, $t = 0, 1, 2, \ldots, s-1$ of (1) be represented by $\mathbf{X}$ and $\mathbf{Y}$, respectively. $\mathbf{X}$ and $\mathbf{Y}$ are $l \times s$ and $m \times s$ matrices, respectively. Let laso $\mathbf{y}_j$ represent the system approximation at scale j

[4] For the notations refere to Fig.2

such that $\mathbf{Y} = \sum_j \mathbf{y}_j$. Note that $\mathbf{y}_j$ is also an $m \times s$ matrix. Then, the multi-scale modeling procedure at scale j is obtained by decomposing (1) into (refer to Fig. 3)

$$\Delta \mathbf{y}_{j-1} = \mathbf{y}_j + \Delta \mathbf{y}_j \quad (16)$$
$$= C_j \Psi_j(\mathbf{X}) + \Delta \mathbf{y}_j \quad (17)$$

$\mathbf{y}_j = C_j \Psi_j(\mathbf{X})$ is the approximation of $\Delta \mathbf{y}_{j-1}$ (the residue at the previous scale), and $\Delta \mathbf{y}_j$ is the residue at the current scale. C_j is a $m \times n_j$ weight matrix defined in (13) and $\Psi_j(\mathbf{X}) = [\Psi_j(\mathbf{x}_0)\ \Psi_j(\mathbf{x}_1)\ \ldots\ \Psi_j(\mathbf{x}_{s-1})]$ is an $n_j \times s$ matrix. n_j, is the number of translation wavelet nodes at scale j.

For $j = 0$, $\Delta y_{j-1} = \Delta \mathbf{y}_{-1} = \mathbf{y}$. Thus, at the coursest scale we model the system output and for all other scales ($j > 0$) we model the system residue $\Delta \mathbf{y}_{j-1}$.

D. The updating algorithm

In (17), each row of $C_j \Psi_j(\mathbf{X})$ is the linear combination of the rows of $\Psi_j(\mathbf{X})$. The weight matrix C_j can be determined using an LMS type of algorithm. On the other hand, if we start from a biorthogonal wavelet pair (ψ, φ), such that $\psi\varphi^T = I$, the identity, we can use simple inner product rule to determine the weight vector as discussed in section III-A.

Consider the block diagram representation of the wavelet network structure show in Fig.3. At the start,

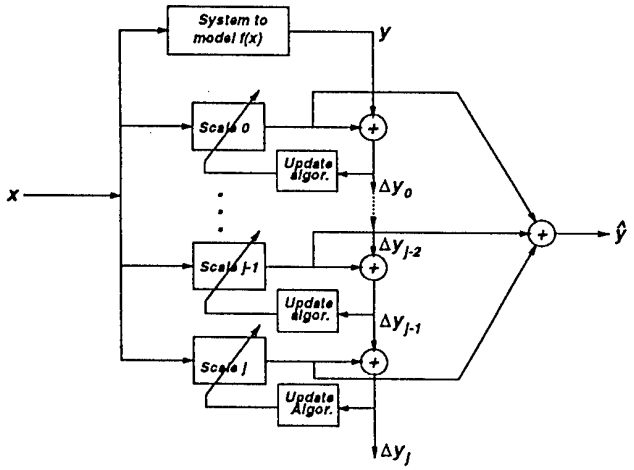


Fig. 3. *Modeling at scale j*

we assume that the initial value of the output $C_j \Psi_j(\mathbf{X})$ at scale-j is zero. From (17), this means that the initial value of the weight matrix is zero, and that of the residue is $\Delta \mathbf{y}_j = \Delta \mathbf{y}_{j-1}$. Let $C_j^{(0)}$ and $\Delta \mathbf{y}_j^{(0)}$ denote these initial values, respectively. Then, we recursively calculate the weights $C_j^{(n)}$ and the resudues $\Delta \mathbf{y}_j^{(n)}$, $n = 1, 2, \ldots$, using the two relations,

$$C_j^{(n+1)} = C_j^{(n)} + \langle \Delta \mathbf{y}_j^{(n)}, \Phi_j(\hat{\mathbf{X}}) \rangle \quad \text{and} \quad (18)$$
$$\Delta \mathbf{y}_j^{(n+1)} = \Delta \mathbf{y}_{j-1} - C_j^{(n+1)} \Psi_j(\mathbf{X}), \quad (19)$$

where $\Phi_j(\hat{\mathbf{X}}) = [\Phi_j(\hat{\mathbf{x}}_0)\ \Phi_j(\hat{\mathbf{x}}_1)\ \ldots\ \Psi_j(\hat{\mathbf{x}}_{s-1})]$ and $\hat{\mathbf{X}}$ is such that $\Psi_j(\mathbf{X})\Phi_j(\hat{\mathbf{X}})^T = I$. Note that if all the rows $\mathbf{X}_p$ of $\mathbf{X}$, $p = 1, 2, \ldots, l$ visit all the samples of the wavelet function ψ, then the above itteration is not necessary as $\langle \Delta \mathbf{y}_j^{(n)}, \Phi_j(\mathbf{X}) \rangle = 0$ for all $n > 0$.

REFERENCES

[1] J. Park and I.W. Sandberg, "Universal Approximation Using Radial Basis FunctionNetworks," *Neural Computation*, Vol.3:pp.246–257, 1990.

[2] R.M. Sanner and J.E. Slotine, "Gaussian Network for Direct Adaptive Control," *IEEE Tran. Neural Net.*, Vol.3, 1992.

[3] J.A. Dickerson and B. Kosko, "Fuzzy Function Approximation with Supervised Ellipsoidal Learning," *Proc. World Congress on Neural Nets*, Vol.2, 1993.

[4] Y.C. Pati and P.S. Krishnaprasad, "Analysis and Synthesis of Feedforward Neural Networks Using Affine Wavelet Transformations," *IEEE Tran. Neural Net.*, Vol.4, 1993.

[5] Q. Zhang and A. Benveniste, "Wavelet Networks," *IEEE Tran. Neural Net.*, Vol.3, 1992.

[6] J. Zhang, G.G. Walter, Y.Miao, and W.N.W.Lee, "Wavelet Neural Networks for Function Learning," *IEEE Tran. SP*, Vol.43, 1995.

[7] S. Tan, Y. Yu, and J. Vandewalle, "On-line Approach to Nonlinear System Identification Using Structure-adaptive Neural Networks," *Submitted to IEEE Tran. CAS Part I.*

[8] S. Mallat, "Multiresolution Approximations and Wavelet Orthonormal Basis of $L^2(R)$," *Trans. Amer. Math. Society*, Vol.315, 1989.