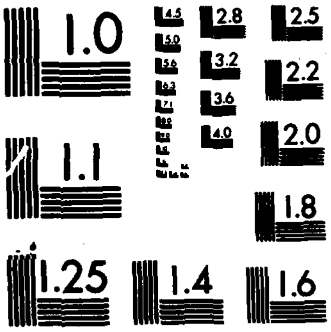
AD-A141 194   OBJECTIVE SPEECH QUALITY EVALUATION OF REAL-TIME SPEECH   1/2
CODERS(U) BOLT BERANEK AND NEWMAN INC CAMBRIDGE MA
V R VISWANATHAN ET AL. FEB 84 BBN-5504 DCA100-82-C-0005

UNCLASSIFIED   F/G 9/4   NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Bolt Beranek and Newman Inc.

(12)

bbn

Report No. 5594

# Objective Speech Quality Evaluation of Real-Time Speech Coders

## Final Report

Febuary 1984

Prepared for:
Defense Communications Agency

Report No. 5504

# OBJECTIVE SPEECH QUALITY EVALUATION OF REAL-TIME SPEECH CODERS

Final Report

Authors: V.R. Viswanathan, W.H. Russell, and A.W.F. Huggins

February 1984

Prepared by:

Bolt Beranek and Newman Inc.
10 Moulton Street
Cambridge, MA 02238

Prepared for:

Defense Communications Agency

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

# REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

| | | |
|---|---|---|
| 1. REPORT NUMBER<br>BBN Report No. 5504 | 2. GOVT ACCESSION NO.<br>A141 194 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>OBJECTIVE SPEECH QUALITY EVALUATION OF REAL-TIME SPEECH CODERS | | 5. TYPE OF REPORT & PERIOD COVERED<br>Final Report<br>Jan. 1982 - Jan. 1984 |
| | | 6. PERFORMING ORG. REPORT NUMBER<br>BBN Report No. 5504 |
| 7. AUTHOR(s)<br>V.R. Viswanathan, W.H. Russell, and A.W.F. Huggins | | 8. CONTRACT OR GRANT NUMBER(s)<br>DCA100-82-C-0005 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Bolt Beranek and Newman Inc.<br>10 Moulton Street<br>Cambridge, MA 02238 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Defense Communications Agency<br>Contract Management Division, Code 680<br>Washington, D.C. 20305 | | 12. REPORT DATE<br>February 1984 |
| | | 13. NUMBER OF PAGES<br>126 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) | | 15. SECURITY CLASS. (of this report)<br>Unclassified |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech quality evaluation, objective speech quality measures, real-time speech coders, estimation of speech coder time delay.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report describes the work performed in two areas: subjective testing of a real-time 16 kbit/s adaptive predictive coder (APC) and objective speech quality evaluation of real-time coders. The speech intelligibility of the APC coder was tested using the Diagnostic Rhyme Test (DRT), and the speech quality was tested using the Diagnostic Acceptability Measure (DAM) test, under eight operating conditions involving

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

channel error, acoustic background noise, and tandem link with two other coders. The test results showed that the DRT and DAM scores of the APC coder equalled or exceeded the corresponding test scores of the 32 kbit/s CVSD coder.

In the area of objective speech quality evaluation, the report describes the development, testing, and validation of a procedure for automatically computing several objective speech quality measures, given only the tape-recordings of the input speech and the corresponding output speech of a real-time speech coder. Five real-time speech coders, each operating under several single-link and tandem conditions, were used as test bed in this investigation. A procedure was developed for automatically synchronizing in time the coder output speech with the coder input speech. An input-speech database of 12 sentences was designed for objective speech quality measurements, by selecting a subset of the 72 sentences used in the DAM tests. For a given coder condition, objective measures were computed by comparing the synchronized coder output speech with the coder input, over the foregoing 12-sentence database. The test bed of coder conditions were divided into three classes: waveform coders, vocoders, and all coders. An optimized objective measure was developed for each of the three classes. Correlation of the objective measures with the subjective six-speaker DAM scores was computed in two ways: over individual speaker scores and over scores averaged over all six speakers. The resulting individual-speaker and all-speaker correlations of the best objective measure for the real-time coders were, respectively, 0.93 and 0.97 for waveform coders, 0.72 and 0.92 for vocoders, and 0.9 and 0.96 for all coders. Finally, the objective measures were evaluated over a test set of coder conditions, which was different from the set used for training or optimizing the measures. The changes in correlations over the training and test sets were small when both sets were sufficiently large and reasonably similar to each other.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# ACKNOWLEDGMENT

The authors wish to acknowledge the help of the following individuals: T. Barnwell, III, of Georgia Institute of Technology; A. Derr, K. Field, K. Karnofsky, and O. Kimball of BBN Laboratories; G. Moran of Defense Communications Agency; and W. Voiers of Dynastat.

# 1. INTRODUCTION

The two major objectives of this project were: (1) to formally test and refine, if necessary, the real-time 16 kbit/s adaptive predictive coder with noise shaping (denoted as APC-NS), which was developed at Bolt Beranek and Newman Inc. (BBN) as part of a Defense Communications Agency (DCA) sponsored contract [1 - 4] and (2) to develop, test, and validate a technique for objective speech quality evaluation of real-time speech coders. In this chapter, we state the specific goals of this work (Section 1.1), present the highlights of this work (Section 1.2), and provide an overview of the rest of this report (Section 1.3).

## 1.1 Goals of the Project

### 1.1.1 Test and Refinement of the APC-NS Coder

The project goals were as follows: 1) Evaluate the speech intelligibility of the real-time APC-NS coder using the six-speaker (3 males and 3 females) Diagnostic Rhyme Test (DRT) [5] and the speech quality using the six-speaker Diagnostic Acceptability Measure (DAM) test [6]; and 2) refine and re-evaluate the APC-NS coder if it does not meet the goal DRT and DAM scores given below and in Table 1. The goal scores are intended to correspond to the 32 kbit/s Continuously Variable Slope Delta modulation (CVSD) coder. In other words, the required 16 kbit/s APC-NS coder must provide the speech quality, the speech intelligibility, and the robustness of the 32 kbit/s CVSD coder, under the various operating conditions as follows.

| Test Condition | | Goal | Goal |
|---|---|---|---|
| No. | Description | DRT Score | DAM Score |
| 1 | Quiet Condition | 93 | 64 |
| 2 | 1% bit-errors | 91 | 55 |
| 3 | ABCP environment | 88 | 46 |
| 4 | PROC→PROC Tandem | 90 | 48 |
| 5 | PROC→CVSD Tandem | 90 | 46 |
| 6 | CVSD→PROC Tandem | 87 | 46 |
| 7 | PROC→LPC Tandem | 81 | 42 |
| 8 | LPC→PROC Tandem | 83 | 44 |

TABLE 1. Goal DRT and DAM test scores for the mediumband speech coder. (PROC refers to processor, namely, the APC-NS coder.)

- o Quiet Condition (Noise-free Input and Channel): Produce high speech intelligibility and quality, with a DRT score of 93 and a DAM score of 64.
- o Noisy Channel: Degrade only minimally (relative to the quiet condition) in the presence of 1% random channel bit-errors, with a DRT score of 91 and a DAM score of 55.
- o Acoustic Background Noise: Produce a DRT score of 88 and a DAM score of 46, when the coder's input speech is corrupted by the acoustic background noise typical in Air-Borne Command Post (ABCP) environment.
- o Tandem Operation: Produce satisfactory performance (in both directions) when the coder operates in tandem with itself, 2.4 kbit/s Linear Predictive Coder LPC-10, and 16 kbit/s CVSD. The goal DRT and DAM scores for these tandem operating conditions are given in Table 1.

### 1.1.2 Objective Speech Quality Evaluation

The overall project goal was to develop, test, and validate a procedure for automatically computing several objective speech quality measures, given only the tape-recordings of the input speech and the corresponding output speech of a real-time speech coder. The specific requirements of our work included the following:

- o Use, as test bed, the five real-time speech coders, all running on the CSP, Inc., MAP-300 array processor: 2.4 kbit/s linear predictive coder LPC-10 [7, 8]; 9.6 kbit/s adaptive predictive coder APC-SQ [7, 8]; 9.6 kbit/s residual-excited linear prediction coder RELP [9, 10]; 16 kbit/s APC-NS; and 16 kbit/s CVSD.
- o Use the set of objective measures recommended by Barnwell on the basis of the results of an extensive study sponsored by DCA [11, 12].
- o The computed objective measures must correlate highly

with subjective quality judgments generated by the DAM test, with the goal on the correlation coefficient being 0.9 or better.

- Develop a fixed, limited input-speech database for speech quality evaluations. Since the six-speaker DAM test uses 72 different sentences (12 per speaker) [6], the database design problem reduces to selecting a smaller input-speech database of, say, 12 sentences, which will be representative of the full set of 72 sentences.

- From the final version of the objective speech quality measurement FORTRAN software package, develop, deliver, and demonstrate another package that runs on the PDP-11/34 minicomputer (under RSX-11M) located at the Defense Communications Engineering Center (DCEC), Reston, VA.

An important issue treated in this work is the problem of synchronizing in time the real-time coder output speech with the input speech. We note that the synchronization problem has not been treated in previous work on objective speech quality evaluation as the objective measures have been computed as part of the coder simulation [11 - 18].

## 1.2 Highlights of the Work

The results of the DRT and DAM tests of the APC-NS coder show that the coder meets or exceeds the goal scores under all but one of the conditions given in Table 1. The DRT score for the APC-CVSD tandem condition is 85.9, about 4 points below the goal. However, we have identified a problem with the CVSD coder implementation on the MAP-300, which might be partially responsible for the inferior performance of the above tandem. The APC-NS coder is substantially more robust than the goal scores given in Table 1 indicate, under the conditions of channel error, ABCP noise, and self tandem. We, therefore, decided that refinement of the APC-NS coder was not necessary, and devoted all the remaining effort to the objective speech quality evaluation task. The highlights of our work on this latter task are presented below.

For the five real-time speech coders, we considered a total of 31 coder conditions including the quiet, channel error, and tandem operating conditions. We divided the coder conditions into two classes: _waveform coders_ and _vocoders_. The third class, called _all coders_, included no such preclassification, and it includes all coder conditions. Following the work of Barnwell [11, 12], we developed separate objective measures for each of the three classes. We computed the correlations of the objective measures with the subjective DAM scores in two ways: over individual speaker scores and over scores averaged over all six speakers.

In the initial phase of our work, we used the so-called file-to-file implementations of the real-time coders, which allow the output speech to be perfectly synchronized in time with the input speech. The best objective measure computed over the file-to-file coder conditions produces a correlation of 0.95 over individual speakers and 0.99 over all speakers, for waveform coders. For the other two classes, the best individual-speaker and all-speaker correlations are, respectively, 0.63 and 0.68 for vocoders and 0.87 and 0.93 for all coders.

For the database design problem, we used a multidimensional scaling technique and selected a database of 12 DAM test sentences, 2 per speaker.

For the synchronization problem, we specially designed an input signal that allows us to estimate quite reliably the time delay introduced by a coder. The input signal is random noise and consists of several alternating regions of high and low energies. We selected two versions of the input signal, one for waveform coders and one for vocoders and all coders. The time delay estimation algorithm detects the transitions between the high and low energy regions in both the coder input and output, and compares the locations of the corresponding input and output transitions in determining the time delay estimate. For waveform coders, a very accurate estimate is obtained by cross-correlating the output signal with the input signal around each transition. The estimation error is less than a few samples for waveform coders and is on the order of 50 samples for vocoders. The resulting synchronization errors have a negligible effect on the computed objective measures.

The procedure for objective speech quality evaluation of real-time speech coders is briefly described as follows. Given the input tape containing the time delay estimation (TDE) signal and the 12-sentence database, we apply the input, using a tape recorder, to the real-time coder. The input and the coder output are recorded simultaneously on a two-channel tape recorder. The two-channel tape is digitized using a two-channel A/D facility. The digitized input and output versions of the TDE signal are used to compute the coder time delay, which in turn is used to synchronize the digitized 12-sentence output speech with the corresponding input. Finally, several objective measures are computed from the input and the synchronized output of the 12 sentences.

The individual-speaker and all-speaker correlations of the best objective measure for the real-time coders are, respectively, 0.93 and 0.97 for waveform coders, 0.72 and 0.92 for vocoders, and 0.9 and 0.96 for all coders. The all-speaker correlation exceeds our goal of 0.9 in each case.

The correlations reported above are all maximum values in that they were computed for the same set of coder conditions over which the objective measures were optimized or trained. Said another way, the training and test conditions were the same. If, however, the test conditions are different from the training conditions, the test-set correlations will in general be lower than the training-set correlations. In our experiments, we observed negligible to modest changes in the correlations. It is important to note that if the test set is quite different from the training set, the test-set correlations could be drastically lower than the training-set correlations. We consider a test coder to be "quite different" from the coders in the training set, if it produces distortions in the output speech that are not covered by the training-set coders. As an example, we point out the case where the training set has coders all operating over noise-free channels and the test set has coders all operating over fairly noisy channels.

## 1.3 Overview of the Report

Chapter 2 treats the subjective evaluation of the APC-NS coder and gives the DRT and DAM scores for this coder under a number of different conditions. In Chapter 3, we discuss the methodology of objective speech quality evaluation of speech coders, review past work, and point out the major issues considered in this research. In Chapter 4, we describe the objective measures considered in this work and the method used for optimizing the objective measures. Chapter 5 describes the test bed of real-time coder conditions. In Chapter 6, we present our objective speech quality evaluation work using file-to-file coder implementations and perfect input-output synchronization. The topics treated in this chapter include database design, performance of the optimized measures, and performance comparisons over training and test sets. The problem of time delay estimation and synchronization is treated in Chapter 7. In Chapter 8, we consider the real-time coder conditions, describe how the objective speech quality measurement is performed, and present the correlation results. In Chapter 9, we describe briefly the objective speech quality measurement FORTRAN software package that we developed for the sponsor's PDP-11/34. Finally, in Chapter 10, we present a summary of this work and discuss some topics that warrant further research.

## 2. SUBJECTIVE EVALUATION OF THE REAL-TIME 16 KBIT/S APC-NS CODER

For a detailed description of the 16 kbit/s APC-NS coder and its real-time implementation, the reader is referred to the publications [1 - 4]. As a brief summary of the APC-NS algorithm, we mention that the analog input speech is lowpass filtered at 3.2 kHz, sampled at 6.621 kHz, and divided into non-overlapping frames of 32.625 ms duration (216 samples). The APC encoder employs (1) 3-tap pitch prediction and 6-pole spectral prediction to obtain the residual, (2) forward-adaptive quantization of the residual samples using 2 bits/sample, and (3) pole-zero spectral shaping of the quantization noise to reduce its perception at the coder output. The real-time, full-duplex implementation of the APC-NS algorithm on the MAP-300 requires an execution time of a little over 30 ms per frame of speech, which leaves about 2.5 ms of unused processing time per frame.

For formal subjective evaluation of the APC-NS coder, we used the six-speaker (3 males and 3 females) DRT and DAM tests. The DRT is a speech intelligibility test consisting of a set of 192 words, with a different randomized word list used for each speaker [5]. The DAM is a speech quality test consisting of a different set of 12 sentences for each speaker [7]. Below, we discuss the test conditions, describe how we generated the test tapes, and present the test results.

### 2.1 Test Conditions

We considered a total of 8 conditions under which we evaluated the APC-NS coder: quiet condition (no acoustic background noise and no channel bit-errors), 1% channel bit-errors, input in Air Borne Command Post or ABCP noise environment, and 5 tandem conditions with the APC-NS coder operating in cascade with another coder. The 5 tandem conditions we considered are self-tandem (APC-NS operating in tandem with itself) and tandem of APC-NS in both directions with each of the 2 coders: the 16 kbit/s CVSD coder and the 2.4 kbit/s government-standard coder LPC-10 (APC-CVSD, CVSD-APC, APC-LPC, and LPC-APC tandems).

We obtained from the sponsor the source DRT and DAM test tapes for the quiet and ABCP noise environments, and the LPC-10 and CVSD processed DRT and DAM test tapes for use in the LPC-APC and CVSD-APC tandem conditions. We used the channel error simulator supplied by the sponsor to introduce random bit-errors in the transmitted bitstream, for the channel-error condition. For the APC-LPC tandem condition, we used the MAP-300 real-time implementation of LPC-10 (Version 44), which was developed at BBN as part of another DCA contract [8].

For use in the APC-CVSD tandem condition the sponsor provided us with real-time MAP-300 software for the 16 kbit/s CVSD coder. We installed this software on our MAP-300 system and compared the output CVSD speech with an analog tape copy of the corresponding speech from the DCEC MAP-300 CVSD coder. The CVSD speech from our implementation was found to be more noisy. From a detailed examination, we discovered that the DC level of the analog input to the MAP at BBN was approximately twice the level experienced at DCEC (about -.0083 versus -.0035, when expressed fractionally with range -1.0 to +1.0). The increased DC level caused the background "noise floor" to be large as compared to that on the tape. We altered the CVSD implementation on our MAP to subtract out a prespecified constant from the digitized speech samples. By trial and error, we selected a value for this constant (+0.007), which rendered a noise floor that sounded most similar to the one on the sponsor-supplied CVSD tape.

For both the CVSD and the APC-NS coders, we found that switching in the 100-Hz highpass filters at the MAP Speech Processor Interface did not change the output speech quality. Upon consultation with the COTR, George Moran, we decided to switch in the highpass filters for the CVSD and the APC-NS coders. (The LPC-10 implementation already uses the highpass filter.) As for lowpass filters, we used 3.2 kHz filters for both CVSD and APC-NS and 3.8 kHz filters for LPC-10.

## 2.2 Generation of Test Tapes

To ensure that the taping set-up could be reproduced exactly from one day to another, we began by driving the signal path (amplifier, lowpass and highpass filters, MAP, and lowpass filter) with a tone generator, and measuring signal levels and clipping levels at several points, using a scope, an RMS meter, and the facility in APC-NS for reading the maximum level encountered by the MAP A/D. An earlier set-up could then be replicated by re-establishing the same sequence of levels. During the taping sessions themselves, several precautions were taken to ensure that the input signal used as much of the A/D's dynamic range as possible without saturating, and also that the MAP algorithm worked as expected. The signal input to the MAP and the off-the-tape monitor of the output tape were monitored simultaneously on stereo headphones, one in each ear. The system delays were such that the two signals did not overlap each other, at least during the DRT tapes. The input signal was also monitored on a scope, and the maximum level encountered by the MAP A/D was read at the end of each speaker's material. Occasionally the gap between the preceding announcement and the test materials was too short, or the MAP algorithm generated a glitch caused by loss of synchronization when the coder parameters were accessed, and this necessitated backing up the tapes a bit. Otherwise, each tape was played through without stops.

Code letters A to H were randomly assigned to the eight experimental conditions, and these identification codes were used in the headers recorded at the start of each tape, to ensure a blind test.

During the course of making the tapes, we noticed a malfunction in the real-time implementation of CVSD. The malfunction of CVSD appeared after the coder had been running for about 40 minutes, and sounded like a repeated click or burst of noise, with a period of about a tenth of a second. Restarting CVSD solved the problem: the noise went away. During the preparation of the DRT and DAM tapes for the APC-CVSD tandem, we simply restarted the CVSD coder whenever the foregoing malfunction occurred.

## 2.3 Test Results

The DRT and DAM test tapes were sent to Dynastat for scoring and evaluation. We received from Dynastat a compete set of the computer printouts containing the detailed scoring for six-speaker DRT and DAM tests of the APC-NS coder under each of the 8 conditions. The DRT and DAM test scores averaged over 3 male speakers, 3 female speakers, and all 6 speakers are given in Table 2 for the 8 test conditions. (The DAM scores given are the Composite Acceptability Measure [6].) From the results for the conditions 1, 2, and 4, we see that the APC-NS coder has an impressively robust performance in 1% channel error and in self tandem. The six-speaker DRT and DAM scores (taken from Table 2) are given in Table 3 along with the standard error of listener means and the goal scores (taken from Table 1, Section 1.1.1). We note that eight listeners were used in all the DRT tests, and eleven listeners were used in all the DAM tests. From the DAM scores given in the table, we observe that the test score is higher than the goal score in all but the first condition; the difference between the goal and test scores in the first condition is not statistically significant ($P \sim 0.25$).

Considering the DRT results given in Table 3, we note that the APC-NS coder exceeds the goal for the conditions 2,3,4, and 6. For the remaining 4 conditions, the difference between the goal and test scores is statistically significant only for the APC-CVSD tandem condition ($P<0.01$). In fact, this is the only case in which the APC-NS coder clearly failed to achieve the goal score. We note here that while the CVSD speech used in the CVSD-APC tandem (condition 6) was obtained from the sponsor, the CVSD coder we used in preparing the test tapes for the APC-CVSD tandem is a MAP-300 implementation supplied by the sponsor. The low DRT score for the APC-CVSD tandem is perhaps due to the differences between the two CVSD coder implementations. We reported above a DC bias problem and a malfunction of the MAP-300 implementation of the CVSD coder.

| TEST CONDITION | | DRT SCORE | | | DAM SCORE | | |
|---|---|---|---|---|---|---|---|
| No. | DESCRIPTION | COMBINED | MALE | FEMALE | COMBINED | MALE | FEMALE |
| 1 | IDEAL | 92.9 | 94.9 | 90.8 | 62.8 | 65.2 | 60.4 |
| 2 | 1% BIT-ERROR | 92.6 | 94.2 | 90.9 | 58.9 | 62.3 | 55.7 |
| 3 | ABCP NOISE | 88.9 | 90.6 | 87.1 | 51.5 | 55.3 | 47.7 |
| 4 | APC-APC TANDEM | 92.4 | 93.9 | 91.0 | 60.8 | 63.1 | 58.6 |
| 5 | APC-CVSD TANDEM | 85.9 | 86.8 | 84.9 | 48.2 | 51.3 | 45.2 |
| 6 | CVSD-APC TANDEM | 89.0 | 90.9 | 87.1 | 47.0 | 49.6 | 44.6 |
| 7 | APC-LPC TANDEM | 80.9 | 83.1 | 78.7 | 43.0 | 46.4 | 40.0 |
| 8 | LPC-APC TANDEM | 81.9 | 83.3 | 80.4 | 47.7 | 50.1 | 45.4 |

TABLE 2. DRT and DAM scores for the APC-NS coder, averaged over 3 male speakers, over 3 female speakers, and over all six speakers.

| Test Condition | | DRT Score | | | DAM Score | | |
|---|---|---|---|---|---|---|---|
| | | | Test | | | Test | |
| No. | Description | Goal | Mean | Std. Error | Goal | Mean | Std. Error |
| 1 | Ideal | 93 | 92.9 | 0.67 | 64 | 62.8 | 0.9 |
| 2 | 1% Bit-error | 91 | 92.6 | 0.95 | 55 | 58.9 | 1.2 |
| 3 | ABCP Noise | 88 | 88.9 | 0.86 | 46 | 51.5 | 0.6 |
| 4 | APC-APC Tandem | 90 | 92.4 | 0.60 | 48 | 60.8 | 1.1 |
| 5 | APC-CVSD Tandem | 90 | 85.9 | 0.94 | 46 | 48.2 | 0.9 |
| 6 | CVSD-APC Tandem | 87 | 89.0 | 0.89 | 46 | 47.0 | 0.7 |
| 7 | APC-LPC Tandem | 81 | 80.9 | 1.42 | 42 | 43.0 | 1.2 |
| 8 | LPC-APC Tandem | 83 | 81.9 | 1.33 | 44 | 47.7 | 1.1 |

TABLE 3. Goal and test (six-speaker) DRT and DAM scores for the APC-NS coder. Standard error of listener means for each score is also given.

A detailed breakdown of the DRT scores showed that the only category of error in which there was a major difference between the two tandem conditions involving the CVSD coder was Sustension. Sustension is the feature that distinguished between f/p, th/t, sh/ch, v/b, dh/d, zj/jh, as in the word pairs sheet/cheat and von/bon, for example. The sustension scores were 14% lower when CVSD was the second member of the tandem than when it was the first member, and this difference was slightly larger for male speakers (16%) than for females (12%). An analysis of the error scores on individual word pairs (obtained from the lists of the 10 most frequent errors presented under each condition) showed a moderate incidence of errors on the word pairs sheet/cheat and shoes/choose in the CVSD tandem conditions, but not under any other conditions. Both pairs involve the distinction between sh and ch before high front vowels. Spectrograms show that most of the affricative energy in "cheat" lies above 3.2 kHz (the lowpass filter cutoff used for CVSD and APC-NS) but that the fricative energy in "sheet" extends well down into the passband, to perhaps 2 kHz. This means that the short, intense high-frequency burst of frication energy that is the hallmark of the ch is simply absent, and the distinction between the sh and the ch depends on the more steady-state frication energy of the sh, which is at a low level because most of its energy, also, falls outside the passband. If anything, this situation should be even more pronounced for female speech, since the low-frequency limit of the frication energy would be higher for female voices. If, in addition, there was reason to believe that a particular coder might add low-level background noise, this might further increase the similarity of the words in the pairs sheet/cheat and shoes/choose. In view of the problems reported in Sections 2.1 and 2.2, we believe that our implementation of the CVSD coder might have had a slightly higher background noise level, which would have further reduced the sh vs. ch discriminability.

From the results reported above, we conclude that the performance of the APC-NS coder meets or exceeds the goal scores. Also, the APC-NS coder is substantially more robust than the goal scores indicate, under the conditions of channel error, ABCP noise, and self tandem. We, therefore, decided that refinement of the APC-NS coder was not necessary, and devoted all the remaining effort to the objective speech quality evaluation task.

## 3. OBJECTIVE QUALITY EVALUATION OF SPEECH CODERS

Quality assessment of speech coders is often performed to determine the user acceptance of a coder, or to compare the performance of competing coder types, or to evaluate the different choices of a given coder's design parameters. Procedures used for speech quality measurement are either subjective or objective, depending upon whether or not they make use of subjective judgments from human listeners. Subjective procedures require extensive testing with human listeners, which is expensive in terms of both time and money. On the other hand, objective measures would enable evaluation to be done by computer as well as ensure uniformity in speech quality evaluation. Also, objective measures can be incorporated into the design of better quality coders. Of course, the validity of any objective procedure must first be established by comparing its results against subjective judgments.

Below, we first describe a general methodology for objective quality evaluation of speech and then briefly review past work. Following that, we point out the specific topics considered in this research.

### 3.1 Review of Methodology and Past Work

Figure 1 illustrates the problem of objective speech quality measurement considered in this research. Given the input speech and output speech of a speech coder, we first synchronize the output speech with the input speech by compensating for the time delay introduced by the speech coder and then compute one or more

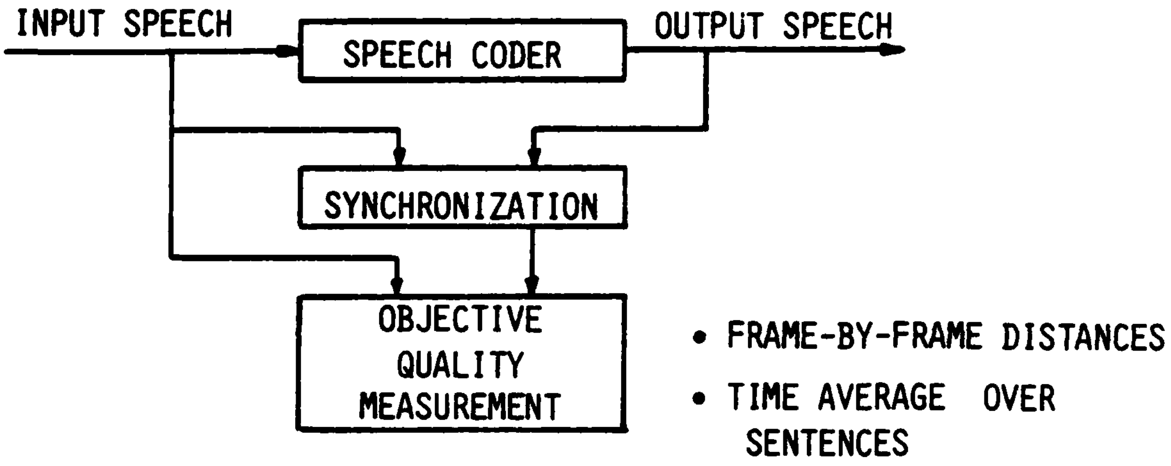

INPUT SPEECH
SPEECH CODER
OUTPUT SPEECH
SYNCHRONIZATION
OBJECTIVE QUALITY MEASUREMENT
• FRAME-BY-FRAME DISTANCES
• TIME AVERAGE OVER SENTENCES

Figure 1. A schematic diagram that illustrates the procedure for objective speech quality measurement.

objective speech quality measures from the input speech and the synchronized output speech, as follows: Compare the input speech with the synchronized output speech on a frame-by-frame basis (a frame may be 10-30 ms long), compute several spectral and parametric distance measures, and average the frame-by-frame measures in time over different sentences from a predetermined speech database. The goodness of the computed objective measures is evaluated by correlating the objective scores with the corresponding subjective judgments over a database of speech coders. By proper selection of the frame-by-frame distance measures and of the method used for time averaging them, we may maximize the correlation of the objective measures with the subjective judgments.

The foregoing methodology for objective speech quality measurement was originally proposed for narrowband LPC vocoders as part of an earlier project at BBN [13]. By assuming that the vocoder represents pitch and voicing accurately, the authors of this paper computed the frame-by-frame distance measures as the distance between the input speech and output speech LPC model parameters or spectra. A number of frequency-weighted spectral measures and parametric measures were proposed in [14, 15] for computing the frame-by-frame distances. Combining the frame distances into a single speech quality score involves first weighting the frame errors with a suitable time-weighting function to reflect the relative importance of the individual frames to perceived speech quality and then averaging the weighted frame errors. Two time-weighting functions were found to produce good results in [15]. The first one, called energy weighting, is a linear or a piecewise linear function of the frame speech energy expressed in decibels. The second one, called dynamic fidelity weighting, in effect changes the reference with respect to which the frame errors are computed. Rather than using the parameters extracted every frame, this approach uses the extracted parameters only for perceptually significant frames as determined by a variable-frame-rate scheme [19], and it uses linearly interpolated parameter values for the remaining frames. For averaging the weighted frame errors over time, reference [15] used a two-term composite measure in which the first term is the usual r-th norm over all frame errors (e.g., r=2 for the rms value), and the second term is the r-th norm over only the top 10% frame errors. The authors of this paper considered several composite measures involving different frame-by-frame measures and different weighting functions, and compared them against a database of subjective speech-quality judgments [15]. The correlation coefficient values computed over individual sentence-length utterances ranged between 0.9 and 0.95 for males and between 0.69 and 0.93 for females.

In the work of Barnwell at Georgia Tech [11, 12], the above described objective evaluation approach was investigated (1) over an extensive set (264 cases) of LPC, waveform, and voice-excited coders and controlled distortions such as lowpass filtered speech and speech corrupted by additive random noise; (2) using an extensive set of objective measures (a total of 500) including simple parametric, spectral, and signal-to-noise ratio (SNR) measures, frequency-variant versions of these measures, and composite measures involving up to six individual measures; and (3) employing, for correlation studies, the DAM test results involving the overall DAM score (Composite Acceptability Measure) as well as several isometric and parametric component scores. For the composite measures, both linear and nonlinear regression techniques were used to compute the optimal weights associated with the individual measures. The best composite measure involves a binary preclassification of the coder under evaluation as a waveform coder or as an LPC-type vocoder, and it produced a correlation coefficient of 0.9 when the measure was compared against the overall DAM score. For each of the two classes of coders, the composite measure is a linear combination of a different set of six individual measures. The best composite measure obtained without preclassification produced a correlation coefficient of 0.86. It is important to note that the weights in the linear combination used by the aforementioned best composite measures were determined by maximizing the correlation coefficient over the database of DAM scores. Therefore, the above-reported high correlation coefficient values represent optimistic estimates in that these measures may, in general, produce lower correlations when they are compared against a different database (i.e., test set) of DAM scores.

In addition to the work reviewed above, there are various other studies reported in the literature, which deal with specific types of speech coders (e.g., [16, 17, 18]). We note that for the purpose of this work, we chose to use the set of objective measures recommended by Barnwell [11].

## 3.2 Topics Considered in this Research

In this research, we were concerned with objective speech quality evaluation of <u>real-time</u> speech coders. We have treated the important problem of synchronizing the real-time coder output speech with the input speech. We note that the synchronization problem has not been treated in previous work on objective speech quality evaluation as the objective measures have been computed as part of the coder simulation.

Since the six-speaker DAM test uses a total of 72 different sentences (12 per speaker), we have investigated the database design problem and developed a smaller database of 12 sentences that are representative of the full set of 72 sentences.

We have optimized the objective measures over one set of coder conditions (training set) and evaluated their correlation with the DAM scores on another set (test set), to examine the sensitivity of the measures to training-set versus test-set differences.

## 4. OBJECTIVE MEASURES OF SPEECH QUALITY

In this chapter, we describe the objective measures we chose to include in our investigation (Section 4.1); discuss the different ways of computing the correlation of the objective measures with the subjective (DAM) scores (Section 4.2); present the method we used for optimizing the objective measures (Section 4.3); and describe briefly the package of FORTRAN programs we developed for implementing the objective measure (Section 4.4).

### 4.1 Description of Objective Measures

For the purpose of this project, as mentioned earlier, we chose to use the set of objective measures recommended by Barnwell on the basis of the results of an extensive correlation analysis with DAM test results involving 500 objective measures and 264 speech coders and controlled distortions [11]. The chosen set includes 7 measures: log area ratio (LAR) measure, reflection coefficient (RFC) measure, feedback (or predictor) coefficient (FBC) measure, likelihood ratio (LHR) measure, linear spectral distance (LSD), frequency-variant spectral distance (FVSD), and short-time banded signal-to-noise ratio (STB-SNR). The first 5 measures are simple parametric or spectral measures, and the other 2 measures are composite measures, each of which is computed as a linear combination of simple spectral measures. For computing the first 6 of these measures, we perform linear prediction (10th order) analysis over a frame of input speech and over the corresponding frame of output speech, and then compute parametric and spectral distances as follows. For LAR, RFC, and FBC, we compute the $L_1$ norm between the respective frame parameters of the input speech and those of the output speech. If x(i) and y(i) denote the i-th parameter of the input speech and the output speech, respectively, then the $L_1$ norm between x and y is given as follows:

$$L_1\ (x,\ y) = \frac{1}{M} \sum_{i=1}^{M} |x(i) - y(i)|, \tag{1}$$

where M is the LPC order (M=10 in our case).

For LHR, we compute the likelihood ratio (or prediction error ratio) between the two sets of LPC parameters [20], and raise it to the power 0.25. If a and b are the predictor coefficient vectors of the input speech and the output speech, and R is the autocorrelation matrix of the input speech, then the error ratio is given by:

$$\text{Error ratio} = \frac{b^T R b}{a^T R a}\ , \tag{2}$$

where T denotes a transpose. The numerator of this ratio is the energy of the linear prediction error signal if the predictor coefficient vector b is used for the input speech, and the denominator is the error energy if a is used for the input speech. Since a is obtained by minimizing the prediction error over the input speech frame, the above error ratio will be equal to or greater than 1.

For LSD, we compute the LPC all-pole model spectra for input and output speech, normalize them to have the same geometric mean, and calculate the $L_2$ norm between the normalized spectra. If a and b are the predictor coefficient vectors of the input speech and the output speech, then the LPC model spectra are given by

$$P_a(\omega) = \frac{G_a^2}{A(\omega)}\ , \tag{3}$$

$$P_b(\omega) = \frac{G_b^2}{B(\omega)} , \qquad (4)$$

where $\omega$ is the frequency in radians/s, $G_a^2$ and $G_b^2$ are the respective prediction error energies, and $A(\omega)$ and $B(\omega)$ are the respective inverse filter spectra:

$$A(\omega) = 1 + \sum_{k=1}^{M} a(k)\ e^{-j\omega k} , \qquad (5)$$

$$B(\omega) = 1 + \sum_{k=1}^{M} b(k)\ e^{-j\omega k} . \qquad (6)$$

The geometric-mean normalized LPC spectra are $1/A(\omega)$ and $1/B(\omega)$, respectively. The required $L_2$ norm is therefore given by

$$L_2 \text{ norm} = \left[ \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{1}{A(\omega_i)} - \frac{1}{B(\omega_i)} \right]^2 \right]^{1/2} , \qquad (7)$$

where N is the number of frequency points considered.

For FVSD, we compute the input and output LPC model spectra; divide each spectrum into 6 bands (0-400 Hz, 400-800 Hz, 800-1300 Hz, 1300-1900 Hz, 1900-2600 Hz, and 2600-3400 Hz); normalize the bands separately so that the average spectral amplitude over each band is unity; compute, for each band, a weighted $L_4$ norm between the normalized spectra, with the weighting function being the input LPC spectrum; and form a linear combination of norms of the 6 bands. If we denote the normalized spectra as $P_a'$ and $P_b'$, then the required weighted $L_4$ norm over the k-th band is:

$$L_4 \text{ norm} = \left[ \frac{\sum_{i=1}^{N_k} P_a(\omega_i)\,[P_a'(\omega_i) - P_b'(\omega_i)]^4}{\sum_{i=1}^{N_k} P_a(\omega_i)} \right]^{1/4}, \qquad (8)$$

where $N_k$ is the number of frequency points considered in the k-th band.

For STB-SNR, we use the input speech as "signal" and the difference between the input speech and the output speech as noise; "filter" the signal and noise into 6 bands (band definitions same as in FVSD); compute the short-time SNR for each band as the ratio in dB of the band signal energy and the band noise energy over the frame under consideration; and form a linear combination of the short-time SNR's of the 6 bands. In our implementation, instead of using filters to divide the signal and the noise into 6 bands, we computed the power spectra of the signal and the noise via DFT and calculated the band energies by summing the spectral values over individual bands.

For time averaging the above-described frame measures over frames, we use simple, unweighted average for all measures except LSD; for LSD, we use the frame energy in dB of the input speech as the weighting function. For the band short-time SNR's used in STB-SNR, we average over frames the quantity (1+SNR) in dB and compute the required SNR in dB from this average (this average in dB is denoted below as SNR´) as

$$\text{SNR(dB)} = 10 \log\,[10^{\text{SNR}'/10} - 1]. \qquad (9)$$

This method of averaging the short-time SNR essentially ignores the frames for which the SNR in dB is negative, which may happen during pauses in the speech signal [18].

To obtain consistent and reliable estimates, we found through experiments that an FFT order of 8 (i.e., 256-point FFT) was necessary for the LSD measure and an FFT order of 11 was necessary for both the STB-SNR and the FVSD measures.

The constants required for the linear combination used in FVSD and STB-SNR measures are computed via linear regression analysis with DAM test scores over a training set, as described below in Section 4.3. From the foregoing seven measures, we compute three overall composite measures (OCM), one for each of the three coder classes: waveform coders, LPC-type vocoders, and all coders. The OCM is a linear combination of 6 of the foregoing seven measures: The unused seventh measure is FBC for waveform coders and STB-SNR for vocoders and for all coders. As before, the constants of each of the linear combinations are determined via linear regression analysis with the DAM test scores over an appropriate training set.

## 4.2 Correlations with Subjective Judgments

We compared the objective speech quality scores with subjective judgments collected via six-speaker DAM tests and computed correlations over different databases of speech coder conditions. We computed two types of correlation between the objective and subjective data: (1) regular, or Pearson's product-moment, correlation (we shall call this simply correlation); and (2) rank order, or Spearman's rank correlation. For the second type, two sets of ranks are first assigned to the speech coders under study using, respectively, objective and subjective data, and then regular correlation is computed between the two sets of ranks.

Since the DAM test provides subjective scores for individual speakers, averaged over male speakers, averaged over female speakers, and averaged over all speakers, we computed the correlation between the objective and subjective data for each of the 4 cases. To minimize possible confusion, we have chosen to present in this report only the regular correlation and only the individual-speaker and all-speaker correlations.

## 4.3 Least Squares Optimization of Objective Measures

As mentioned in Section 4.1, each of the composite measures FVSD, STB-SNR, and OCM uses coefficients to weight and linearly combine individual component measures. The problem considered in this section is to determine the values of these coefficients by optimizing the composite measure over a given database of speech coder conditions. Given that we have, say, p component objective measures $M_1$, $M_2$, ..., $M_p$, we wish to compute a linear composite measure $\hat{S}$ as a prediction of the subjective (DAM test) score S, as follows:

$$\hat{S}_i = \bar{S} + \sum_{j=1}^{p} (M_{ij} - \bar{M}_j)\, c_j \,, \qquad (10)$$

where the subscript i is the item number, which refers to a given coder running under a given condition; $M_{ij}$ is the computed value of the j-th objective measure $M_j$ for the i-th item; $\bar{M}_j$ is the mean of $M_j$ over the items from a given database; $\bar{S}$ is similarly the mean of the subjective scores; and $c_j$ are the coefficients used to weight the individual objective measures. As an estimate of the subjective score, $\hat{S}_i$ is <u>unbiased</u> over the "training" database, since its mean is equal to $\bar{S}$. The above expression for $\hat{S}$ can be rewritten as:

$$\hat{S}_i = c_o + \sum_{j=1}^{p} c_j M_{ij}, \tag{11}$$

where the additive constant $c_o$ includes the means of subjective and objective scores.

We compute the coefficients $c_j$ by minimizing the error measure:

$$E = \sum_{i=1}^{N} (S_i - \hat{S}_i)^2 , \tag{12}$$

where N is the number of items in the database. The least squares method results in a set of linear normal equations given below:

$$Q c = r, \tag{13}$$

where $Q = [Q_{ij}]$ is a p x p symmetric matrix, with

$$Q_{ij} = \sum_{k=1}^{N} (M_{ki} - \bar{M}_i)(M_{kj} - \bar{M}_j); \tag{14}$$

$c = (c_1, c_2, \ldots, c_p)^T$ is the column vector of the coefficients; and $r = (r_1, r_2, \ldots, r_p)^T$ is a column vector, with

$$r_j = \sum_{k=1}^{N} (M_{kj} - \bar{M}_j)(S_k - \bar{S}). \tag{15}$$

The coefficients $c_j$ are obtained by solving the above normal equations. We note that publicly available statistical packages have subroutines for solving the least squares optimization problem. In our work, we used the subroutine LLSQF from the International Mathematical and Statistical Library.

We point out that the foregoing least squares method also maximizes, as a desirable byproduct, the correlation coefficient (regular correlation, not rank-order correlation), computed over the training database, between the composite measure $\hat{S}$ and the subjective score S. This maximum correlation coefficient can be computed easily, as shown below. We caution that the correlation coefficient computed over a different ("test") database will in general be lower than this maximum value.

It can be shown that the minimized value E* of the error E given above is related to the variance $\sigma_S^2$ of the subjective score as follows:

$$E^* = (1 - \rho^2)\, \sigma_S^2 \quad . \tag{16}$$

where $\rho$ is the correlation coefficient for the optimized composite measure. From this last expression, $\rho$ can be easily determined, since both $\sigma_S$ and E* are known or can be computed.

For the simple case p=1, it can be shown that

$$c_1 = \rho\sigma_S/\sigma_M \, , \tag{17}$$

$$\hat{S}_i = \bar{S} + \frac{\rho\sigma_S}{\sigma_M} \; (M_i - \bar{M}), \tag{18}$$

where $\sigma_M$ is the standard deviation of the objective scores over the database of speech coder conditions.

We point out that a composite measure may be optimized to predict the all-speaker subjective score or the individual-speaker subjective score, which will result in maximizing, respectively, the all-speaker correlation or the individual-speaker correlation. Notice that the number of items (N) in the database is the number of coder conditions for the all-speaker case, and it is six times the number of coder conditions for the individual-speaker case as there are six subjective scores (one per speaker) for every coder condition. The number N must be sufficiently large compared to the dimension p of the composite measure, to ensure the validity of the optimized measure for use over other (test) coder conditions. This last issue will be treated later in Chapter 6.

Another issue related to the optimization is what we call combinatorial analysis. The purpose of this analysis is to determine, for a given set of component measures, the best single measure, the best two-element composite measure, the best three-element composite measure, and so on, along with their respective optimized correlations. From these data, we can determine the incremental benefit each additional component measure offers, which in turn can be used, if necessary, in our decision to include or exclude one or more component measures.

## 4.4 Software Implementation of the Objective Measures

We developed the software for objective speech quality measurement as two separate programs on our VAX-11/780 computer running under the VMS operating system. Each program is organized around a human-engineered, interactive command system that allows the user to ask the program to reset key parameters or execute one of several functions the program has been designed to perform. Given the digitized input and output speech of a coder to be evaluated, the first program synchronizes the output speech with the input speech, performs linear prediction and spectral analysis of the two speech signals at a prespecified frame rate, and computes and stores in disk files frame-by-frame parameter data and a number of parameter and spectral measures. One parameter file is created for each sentence from the speech database. The second program performs several functions: it reads one or more frame-by-frame parameter files, computes simple and composite objective measures, and writes them out into disk files (one objective measures file per coder condition); it reads two or more objective measures files, compares the objective scores with subjective scores, and computes the different types of correlations; it performs least squares optimization to determine the coefficients of a given composite measure; and it performs the combinatorial analysis for a given composite measure.

## 5. TEST BED OF REAL-TIME SPEECH CODER CONDITIONS

As test bed for our objective speech quality evaluation work, we used five real-time coders, all implemented on the CSP, Inc. MAP-300 array processor. As mentioned in Chapter 1, these five coder are: 16 kbit/s APC-NS [1-4], 16 kbit/s CVSD, 9.6 kbit/s APC-SQ [7, 8], 9.6 kbit/s RELP [9, 10], and 2.4 kbit/s LPC-10 [7, 8]. We obtained from the sponsor six-speaker (three males and three females) DAM test scores for each of the five coders, operating under several single-link and tandem-link conditions. Table 4 gives the various coder conditions we have chosen to consider in this study and for which DAM test scores are available, along with the all-speaker DAM test scores. We note that DAM test scores are also available for individual speakers. For the purpose of this work, we used only the composite acceptability measure of the DAM test [6]. The DAM tests of the various coder conditions given in Table 4 were performed over a period of about 2 years. This raises some concern about the reliability of the DAM test scores, which we have to keep in mind in sorting out the results of the correlation analyses of the objective and subjective data. For example, we note from Table 4 that the tandems LPC10-APCSQ and LPC10-APCNS produced higher DAM scores than did LPC-10 in the quiet condition! Reference [21] gives some examples that also support our concern regarding the reliability of the DAM test results gathered at different times. Nonetheless, we note that within the Department of Defense, the DAM test is a proven method of assessing speech quality and is widely used.

The MAP-300 real-time implementation of the CVSD coder supplied to us by the sponsor was not set up for simulating

| | CODERS | | | | |
|---|---|---|---|---|---|
| Test Condition | APC-NS | CVSD | APC-SQ | RELP | LPC-10 |
| Quiet | 62.8 | 51.1 | 53.4 | 50.4 | 46.4 |
| 1% Error | 58.9 | 46.1 | 43.8 | 45.5 | 39.9 |
| 5% Error | - | 43.1 | 30.3 | 35.6 | 33.6 |
| Proc->Proc | 60.8 | 41.3 | 46.0 | 43.7 | 43.1 |
| CVSD->Proc | 47.0 | - | 45.0 | 42.3 | - |
| Proc->CVSD | 48.2 | - | 45.0 | 44.2 | - |
| LPC->Proc | 47.7 | 41.2 | 47.2 | 41.3 | - |
| Proc->LPC | 43.0 | 37.4 | 42.1 | 37.2 | - |

TABLE 4. The all-speaker DAM test scores for the coder conditions considered. Proc stands for Processor, and the score in the row Proc->CVSD under the column RELP is the score for the tandem RELP->CVSD.

channel error. Therefore, we did not use the 1% and 5% error conditions for the CVSD coder. The total number of coder conditions we used is 31. Table 5 shows our classification of the coder conditions into waveform coders (WFC) and vocoders (VCD). A tandem-link is classified as WFC only if both coders in the tandem are waveform coders. The 6 RELP coder conditions marked HYB are hybrid in that they are hybrid between waveform coders and vocoders. The question of whether to classify these hybrid conditions as WFC or as VCD will be treated in the next chapter. Excluding the 6 hybrid conditions, we have 13 waveform coder conditions and 12 vocoder conditions. Therefore, for correlation analysis of objective measures with subjective judgments, we have 13 data items for waveform coders, 12 for vocoders, and 31 for all coders, for the all-speaker case; these numbers for the individual-speaker case are, respectively, 78, 72, and 186. Since the composite measures FVSD and STB-SNR have each 7 coefficients (1 constant term and 6 weights), it does not make sense to optimize these measures for the all-speaker case over the vocoder and waveform coder classes. Optimization in all other cases should be reasonable. We shall discuss this issue further in Chapter 6.

| Test Condition | CODERS<br>APC-NS | <br>CVSD | <br>APC-SQ | <br>RELP | <br>LPC-10 |
|---|---|---|---|---|---|
| Quiet | WFC | WFC | WFC | HYB | VCD |
| 1% Error | WFC | - | WFC | HYB | VCD |
| 5% Error | - | - | WFC | HYB | VCD |
| Proc->Proc | WFC | WFC | WFC | HYB | VCD |
| CVSD->Proc | WFC | - | WFC | HYB | - |
| Proc->CVSD | WFC | - | WFC | HYB | - |
| LPC->Proc | VCD | VCD | VCD | VCD | - |
| Proc->LPC | VCD | VCD | VCD | VCD | - |

TABLE 5. Classification of the coder conditions for which the DAM test scores are available (WFC = waveform coders, VCD = vocoders, and HYB = hybrid). We have not marked the 1% and 5% error conditions for CVSD as our CVSD implementation did not allow the channel error simulation.

## 6. OBJECTIVE QUALITY EVALUATION OF FILE-TO-FILE CODERS

In the first phase of our work, we considered the so-called file-to-file implementation of the real-time coders, which allows the coder output speech to be perfectly synchronized in time with the input speech. Using the file-to-file coders, we developed, optimized, and tested the objective speech quality measures, and designed the input-speech database. Below, we explain what we mean by file-to-file speech coders (Section 6.1); describe the results from several initial experiments on objective speech quality evaluation and propose an approach to deal effectively with the limited size of the available database of speech coders (Section 6.2); present our solution to the input-speech database design problem (Section 6.3); present and discuss the correlation performance of optimized measures for waveform coders, vocoders, and all coders (Section 6.4); compare the correlations of the objective measures over training and test sets of coder conditions (Section 6.5); and finally discuss the issue of which way to classify the RELP coder (as a waveform coder or as a vocoder), for the purpose of objective speech quality evaluation (Section 6.6).

### 6.1 File-to-File Coder Implementation

As mentioned above, to guarantee perfect time synchronization of the coder output speech with the input speech, we developed a file-to-file version for each of the five MAP-300 real-time coders; this modified MAP-300 implementation accepts digitized input speech files and produces digitized output files. Except for the analog I/O modules and the modules for acquiring and maintaining synchronization between the receiver and the transmitter, the file-to-file version uses all other modules of the real-time coder. We added software to the file-to-file versions of all but the CVSD coder to simulate the random channel bit-errors.

Table 6 gives, for each of the five coders considered, the speech sampling rate of the coder, the fixed processing delay (in ms and in number of speech samples) introduced by the real-time coder, and the fixed processing delay introduced by the corresponding file-to-file coder. We wrote a program on our VAX to digitize the coder input speech using the A/D on the MAP-300; this program, therefore, allows us to digitize the input speech using the exact sampling rate of any of the five coders.

| Coder | Sampling Rate (kHz) | Real-Time Coder Delay ms (Samples) | | File-to-File Coder Delay ms (Samples) | |
|---|---|---|---|---|---|
| LPC-10 | 8.0 | 292.5 | (2340) | 157.5 | (1260) |
| APC-SQ | 7.68 | 275.0 | (2112) | 75.0 | (576) |
| RELP | 6.621 | 353.4375 | (2340) | 163.125 | (1080) |
| CVSD | 16.0 | 125.0 | (2000) | 0.0 | (0) |
| APC-NS | 6.621 | 358.875 | (2376) | 130.5 | (864) |

TABLE 6. Time delays introduced by the five MAP-300 coders.

To achieve perfect time synchronization of coder output with input, we used the file-to-file coder implementation and the sampling rate of the coder under consideration, and shifted back the output (digitized) speech relative to the input speech by a known integer number of samples equal to the processing delay introduced by the file-to-file coder. (We later extended this approach to the case of a common sampling rate for all five coders, as described in Section 6.2.) For the simulation of the tandem-link conditions, we used linear-phase interpolation-decimation filtering for digitally resampling the output of the first coder in the tandem to match the sampling rate required by the second coder and then compensated for the filtering delay by shifting back the resampled speech. We similarly resampled, if necessary, the output of the tandem to match the sampling rate of the overall input of the tandem, so that the two waveforms could be compared for computing the objective measures. Before we resampled the output of a coder, we of course compensated for the delay of that coder by shifting back the output speech as mentioned above. With the above file-to-file method of simulating the tandem links, we found that the CVSD-CVSD self-tandem behaved exactly like the single-link CVSD coder. As a result of this, we had available only 12 waveform coder conditions (see Table 5 in Chapter 5).

Since the CVSD coder uses a 3.2 kHz lowpass filter and 16 kHz sampling rate, the spectrum of its input and output speech will contain the filter rolloff at 3.2 kHz and no useful information over the range 4-8 kHz. Since we use linear prediction analysis in computing all but the STB-SNR measure, we resampled the CVSD input and output speech at 8 kHz before computing the objective measures, to improve the effectiveness of linear prediction modeling.

From Table 6, we see that the time delay will be maximum at about 353 ms for the APCNS-APCNS tandem. Since time synchronization of the output speech with the input requires shifting the output backwards, we must include silence longer than this maximum delay at the end of individual sentences of digitized input speech. We included about 175 ms silence at the beginning and about 375 ms silence at the end of each sentence.

## 6.2 Results From Initial Experiments

### 6.2.1 Initial Input-Speech Database

Considerably fewer than 72 DAM test sentences would suffice for computation of the objective measures. For our initial work of developing and testing the objective quality measures and as a first cut at the database design problem, we selected a set of 12 DAM test sentences (2 per speaker) through a phonemic analysis of the full set of 72 sentences. (This initial set was later slightly modified as a result of more formal investigation. See Section 6.3.) We counted the number of each category of speecn sounds in a hypothetical spoken version of each sentence. (We did not transcribe the 72 sentences as spoken in the DAM test materials.) The categories we considered for vowels were diphthongs and unstressed vowel, and those for consonants were voicing transitions, fricatives and affricates, stops, nasals, and glides. Then, from the 12 sentences for each speaker, we selected one sentence that had a high number of voicing transitions, fricatives, or stops, and a second second sentence that was low on these categories but high on nasals, glides, or diphthongs. The selected 12 sentences are listed in Table 7.

| Speaker | Sentence Number | Sentence | Phoneme Characteristics |
|---|---|---|---|
| JE | 8 | His clothes have some false cuffs. | unvoiced, fricatives |
| JE | 11 | That flower is in bloom. | voiced, glides, unstressed vowels |
| CH | 4 | Let's talk after his show. | unvoiced, fricatives |
| CH | 10 | I read the news today. | voiced |
| RH | 2 | Frank's neighbor mowed his lawn. | nasals |
| RH | 9 | He sprayed our house for bugs. | unvoiced, diphthongs |
| VW | 3 | These shoes were black and brown. | voiced, glides |
| VW | 11 | I had toast for breakfast. | unvoiced, fricatives, stops |
| KS | 11 | Some diphthongs have no sound. | unvoiced, fricatives, nasals |
| KS | 12 | The book was green and white. | voiced, glides, unstressed vowels |
| MP | 6 | Don't throw trash on the street. | unvoiced, fricatives, stops, glides. |
| MP | 12 | Invest your money now. | voiced, nasals |

TABLE 7. A list of 12 DAM test sentences selected as the initial input-speech database.

### 6.2.1.1 Waveform Coders

For the foregoing database of 12 DAM test sentences, we generated the digitized input and output speech files, for each of the 12 waveform coder conditions, and then computed the various objective measures for the all-speaker and individual-speaker cases. The two banded measures, FVSD and STB-SNR, and the overall composite measure were optimized using all available data items as the training set. We performed correlation analysis of the computed objective measures with the DAM scores, separately for the all-speaker and individual-speaker cases. The resulting correlation coefficients are given in Table 8. Several comments are in order. First, in the all-speaker case, as we correlate averages of objective scores over speakers with averages of DAM scores over speakers, the correlation coefficients should be generally larger for the all-speaker case than for the individual-speaker case. This is also evident in Table 8. Second, the correlation coefficients for the simple measures LAR, RFC, and LHR are high, and those (maximum over the training set) for FVSD, STB-SNR, and the overall composite measure are all excellent for all speakers and quite high for individual speakers. Third, the drop in correlation going from all speakers to individual speakers is more for STB-SNR than for FVSD and OCM. Fourth, the correlations for the simple measures are negative as these measures are error measures (the lower the error, the higher the quality). The composite measures, on the other hand, are computed, as noted in Section 4.3, as prediction of the DAM score, which explains why their correlations are positive.

| Measure | Individual Speakers | All Speakers |
|---|---|---|
| LAR | -0.821 | -0.871 |
| RFC | -0.811 | -0.864 |
| LHR | -0.776 | -0.817 |
| LSD | -0.556 | -0.605 |
| FVSD | 0.934 | 0.995 |
| STB-SNR | 0.891 | 0.986 |
| OCM | 0.936 | 0.996 |

TABLE 8. Correlation coefficients for the simple and composite objective measures over the 12 waveform coder conditions.

### 6.2.1.2 Common Sampling Frequency

For objective speech quality evaluation, ideally we should use a single sampling frequency for all coders; this is one step closer to the approach of not requiring any prior knowledge of the details of the coder under evaluation. To see what should be the common sampling frequency, let us consider the LPC10-APCNS tandem as an example. For objective measure computations, we lowpass filter the output of this tandem at 3.2 kHz (bandwidth of APC-NS) and resample at 8 kHz, to be able to compare with the input of the tandem. Thus, the filter's sharp rolloff is within half the sampling frequency of 8 kHz. Since, as mentioned earlier, we use linear prediction analysis in computing all but the STB-SNR measure, and since the effectiveness of LPC modeling is lowered by the sharp rolloff in the passband, the objective distances come out to be larger than they should be. This suggests the use of the lowest of the coder sampling rates, namely 6.621 kHz, as the common sampling frequency[1].

We digitally resampled at 6.621 kHz (with the lowpass filter cutoff at 3.2 kHz) the input and the output speech files for all coder conditions that required such resampling. We computed the various objective measures and evaluated their correlation with the DAM scores. The resulting correlation coefficients for waveform coders are given in Table 9. Comparing the results in Table 9 with those in Table 8, we see that all simple parametric and spectral measures produce substantially higher correlation with a common sampling frequency. The correlation coefficients for FVSD and STB-SNR show only small changes, as these 6-band measures, with their highest band going up to only 3.4 kHz, already have a built-in bandlimiting aspect. We also noted modest improvements in correlation over the vocoder conditions. Encouraged by these results, we decided to use the common sampling frequency of 6.621 kHz for the remainder of the file-to-file coder work. (For real-time coders, we used 6.67 kHz. See Chapter 8.)

### 6.2.2 Enlarged Input-Speech Database

We enlarged our initial database of 12 selected DAM test sentences by adding the first 6 unused sentences in the DAM sentence list of every speaker. We believe that the enlarged database of 48 sentences (8 per speaker) is adequate for the database design problem (see Section 6.3).

[1] However, this would eliminate any advantage a coder would have from a good high-frequency response beyond about 3.3 kHz.

| Measure | Individual Speakers | All Speakers |
|---|---|---|
| LAR | -0.872 | -0.936 |
| RFC | -0.849 | -0.913 |
| LHR | -0.867 | -0.922 |
| LSD | -0.820 | -0.947 |
| FVSD | 0.935 | 0.991 |
| STB-SNR | 0.893 | 0.993 |
| OCM | 0.941 | 0.995 |

TABLE 9. Correlation coefficients for the simple and composite measures over the 12 waveform coder conditions, for the case using a common sampling frequency of 6.621 kHz.

### 6.2.2.1 Waveform Coders

For the enlarged database, we computed the correlation coefficients for the various measures over the 12 waveform coder conditions. The results are given in Table 10. A comparison of the results in Table 10 with those in Table 9 for the 12-sentence database shows only minor differences. We may interpret this to mean that the initial set of 12 sentences is indeed quite representative of the full set of 48 sentences (see Section 6.3 on database design).

From a computational viewpoint, it is tempting to infer from the results given in Table 10 that we need to use just FVSD and not compute the other five measures required to evaluate the overall composite measure. To pursue this last point further, we performed combinatorial analysis (see Section 4.3), and the resulting best composite measures of various lengths and their correlation coefficients are given in Table 11 for all speakers

| Measure | Individual Speakers | All Speakers |
|---|---|---|
| LAR | -0.879 | -0.934 |
| RFC | -0.859 | -0.915 |
| LHR | -0.864 | -0.913 |
| LSD | -0.831 | -0.935 |
| FVSD | 0.949 | 0.997 |
| STB-SNR | 0.895 | 0.994 |
| OCM | 0.954 | 0.997 |

TABLE 10. Correlation coefficients for the simple and composite measures over the 48 sentences and the 12 waveform coder conditions.

and in Table 12 for individual speakers. These results suggest the use of the two-element composite measure (FVSD,LSD) for the purpose of reducing the computation. Notice that both the measures FVSD and LSD require computing the LPC all-pole model spectrum.

### 6.2.2.2 Vocoders

Since the DAM test scores for the RELP tandem conditions were not initially available, we considered only the 10 vocoder conditions given in Table 5. We optimized the two composite measures FVSD and OCM, and evaluated the correlation coefficients for the two composite measures and the five simple measures over the 10 vocoder conditions. The results are given in Table 13.

As we compare the results given in Table 13 with those in Table 10 for 12 waveform coder conditions, we make two observations. First, the correlation coefficients of the simple

| Measures | Correlation |
|---|---|
| FVSD | 0.99677 |
| FVSD,LSD | 0.99680 |
| FVSD,LSD,RFC | 0.99681 |
| FVSD,RFC,LAR,LHR | 0.99700 |
| FVSD,RFC,LAR,LHR,STB-SNR | 0.99704 |
| All six measures | 0.99709 |

TABLE 11. Best composite measures of various lengths, for all speakers and over the 12 waveform coder conditions.

| Measures | Correlation |
|---|---|
| FVSD | 0.94867 |
| FVSD,LSD | 0.94955 |
| FVSD,STB-SNR,LHR | 0.95119 |
| FVSD,STB-SNR,RFC,LHR | 0.95294 |
| FVSD,STB-SNR,RFC,LHR,LAR | 0.95344 |
| All six measures | 0.95390 |

TABLE 12. Best composite measures of various lengths, for individual speakers and over the 12 waveform coder conditions.

measures are substantially smaller for the vocoder class than for the waveform coder class, and the correlation coefficients of the two composite measures are not as bad, but are still significantly less for vocoders than for waveform coders. Second, the variability of the correlation in going from

| Measure | Individual Speakers | All Speakers |
|---|---|---|
| LAR | -0.125* | -0.311* |
| RFC | -0.113* | -0.288* |
| LHR | -0.046* | -0.271* |
| FBC | -0.187* | -0.268* |
| LSD | -0.254* | -0.430* |
| FVSD | 0.496 | 0.840 |
| OCM | 0.560 | 0.983 |

TABLE 13. Correlation coefficients for the simple and composite measures over the 10 vocoder conditions. The asterisks mark the correlations that fail to achieve statistical significance at P=0.01 (see text).

individual speakers to all speakers is substantially larger for vocoders than for waveform coders. Since the correlations of the simple measures given in Table 13 are quite small, it is reasonable to ask if they are statistically significant. If we set the significance level at 0.01, then the correlation values marked by asterisks in Table 13 are not significant. We note that the correlations given in the table for the two composite measures are all highly significant (at a level better than 0.001).

There are at least two reasons for the poor performance of the objective measures over the vocoder conditions. First, as noted in Chapter 5, the DAM scores for the LPC10-APCNS and LPC10-APCSQ tandems are inconsistent in that they are higher than the DAM score for LPC-10 in the quiet condition (see Table 4). Second, as we explored using the LAR measure, we found that the LAR scores for the 1% and 5% error conditions were much smaller than the corresponding DAM scores indicate. The reason for this, we believe, is that whereas occasional bit-errors cause a noticeable perceptual degradation and hence a low DAM score, the effects of such errors, even if they are large, are 'diluted' by the process of averaging over the database, which is performed in computing the LAR measure. One possible remedy for this problem is to average over only the top, say, 20% of the frame-by-frame LAR distances - an idea we had previously developed as part of another study [15].

We then optimized the measures and evaluated their correlations over a subset of 6 vocoder conditions, obtained by excluding the above two tandems and the two channel error conditions. The results are given in Table 14. Since there are only 6 coder conditions, we did not consider the all-speaker case for FVSD and OCM. (Since the composite measures each have 7 coefficients, we can select them to predict perfectly the DAM scores for the 6 conditions.) The correlations given in Table 14 are, as expected, clearly larger than those in Table 13, although they are still far inferior to those for the waveform coder class. We have used asterisks in Table 14 to indicate the correlations that fail to achieve statistical significance at P=0.01.

#### 6.2.2.3 All-Coder Class

For this class, we considered a total of 25 coder conditions: 12 waveform coder conditions, 10 vocoder conditions, and 3 RELP coder single-link conditions (see Table 5). We optimized the 2 composite measures, FVSD and OCM, over the 25 conditions and the 48 DAM test sentences. The correlation

| Measure | Individual Speakers | All Speakers |
|---|---|---|
| LAR | -0.386* | -0.784* |
| RFC | -0.420 | -0.822* |
| LHR | -0.199* | -0.611* |
| FBC | -0.580 | -0.928 |
| LSD | -0.312* | -0.572* |
| FVSD | 0.747 | - |
| OCM | 0.800 | - |

TABLE 14. Correlation coefficients for the simple and composite measures over a subset of 6 vocoder conditions.

coefficients of the optimized composite measures and the simple measures are given in Table 15. We point out that both FVSD and OCM measures produce high correlations with DAM scores, and that the variability in correlation between the individual-speaker and all-speaker cases is modest.

| Measure | Individual Speakers | All Speakers |
|---|---|---|
| LAR | -0.706 | -0.776 |
| RFC | -0.677 | -0.745 |
| LHR | -0.706 | -0.772 |
| FBC | -0.402 | -0.496 |
| LSD | -0.698 | -0.794 |
| FVSD | 0.882 | 0.943 |
| OCM | 0.886 | 0.958 |

TABLE 15. Correlation coefficients for the simple and composite measures over the 25 all-coder conditions.

The results of combinatorial analysis are given in Table 16 for all speakers and in Table 17 for individual speakers. These results seem to indicate that a composite measure consisting of FVSD, LHR, LSD, and FBC may be adequate.

| Measures | Correlation |
|---|---|
| FVSD | 0.94250 |
| FVSD, LHR | 0.94440 |
| FVSD, LHR, LSD | 0.95067 |
| FVSD, LHR, LSD, FBC | 0.95482 |
| FVSD, LHR, LSD, RFC, LAR | 0.95765 |
| All six measures | 0.95770 |

TABLE 16. Best measures of various lengths, for the all-coder class and all speakers.

| Measures | Correlation |
|---|---|
| FVSD | 0.88210 |
| FVSD, LHR | 0.88549 |
| FVSD, LHR, LSD | 0.88565 |
| FVSD, LHR, LSD, FBC | 0.88600 |
| FVSD, LHR, FBC, RFC, LAR | 0.88606 |
| All six measures | 0.88608 |

TABLE 17. Best measures of various lengths, for the all-coder class and individual speakers.

### 6.2.3 Limited Database of Coder Conditions

The database of coder conditions available to us is rather limited, especially for waveform coder and vocoder classes. This raises two issues of concern. First, with each composite measure requiring 7 coefficients, does it make sense to optimize the composite measures for the all-speaker case? With the full database being already limited, how should it be divided into training and test sets? We discuss these two problems below and indicate practical solutions.

Let us examine the size of the training set used to optimize the composite measures. One of the criteria we used on the size is that it should be large enough so that the evaluated correlation coefficient is adequately significant in a statistical sense. For example, if we demand a significance level of 0.001 or better and if we expect the correlation coefficient r of the composite measure to be about 0.8, then we need at least N=11 data items (i.e., 11 coder conditions for the all-speaker case and 2 coder conditions times 6 speakers for the individual-speaker case). If, however, r=0.6, we need $N \geq 25$; $N \geq 7$, if r=0.9. The foregoing criterion is clearly important also in deciding the test-set size. For the training set, however, satisfying this criterion alone is not sufficient. In the pattern recognition literature, one finds a rule of thumb for the training-set size as 3 to 10 times the number k of (scalar) coefficients that are optimized [22]. For FVSD and STB-SNR measures, k=7. For the OCM measure, it is not clear what k should be. We, however, know that $7 \leq k \leq 17$ ( = 1 overall constant + 6 weights for FVSD + 6 for STB-SNR + 4 weights, 1 each for 4 simple measures) for waveform coder OCM and $7 \leq k \leq 12$ for vocoder or all-coder OCM. Applying this second criterion, we need at least

N=21 data points. Thus, the all-speaker composite measures can be "reasonably" trained only for the all-coder class with 25 coder conditions.

How do we deal with the training problem for the all-speaker case? For example, for the important class of waveform coders, we have a total of only N=12 data points for the all-speaker case. We believe that a reasonable solution to this problem is as follows. Since we have a sufficiently large N for the individual-speaker case, we first train the measure over the individual-speaker data. We then average the resulting objective scores separately over sets of six speakers to obtain the all-speaker objective scores. Correlation is then evaluated between these all-speaker scores and the all-speaker DAM test scores. Since the actual coefficients of the composite measure we use in this approach will have been trained over a sufficiently large number of data points, we have essentially solved our problem. All-speaker correlations given in the rest of this report were computed using this approach.

Next, let us take up the problem of dividing an already limited database into training and test sets. There is no need to divide the full data set into equal training and test sets. Recall from the foregoing discussion that the test set needs to be just large enough to yield the desired level of significance for the test-set correlation. Thus, we must choose the size of the training set to be much larger than that of the test set. This idea is useful for carrying out the training-set versus test-set performance comparisons, especially for the individual-speaker case (see Sections 6.5 and 8.4).

## 6.3 Database Design

In Section 6.2.1, we described our initial selection of 12 sentences out of the full set of 72 DAM test sentences. We now describe a more formal approach for selecting the best subset of 12 sentences. In this approach, we followed a procedure based on Multi-Dimensional Scaling (MDS), developed in earlier work [23]. The MDS-based procedure is less appropriate to the problem of reducing the DAM sentence database than it was in the earlier work, because the finest level at which DAM scores were available was that of the individual speaker, whereas in our earlier work, subjective scores were available for each individual sentence. Since the only scores available at the latter level were the various objective measures, we decided to use the best of these. Since the overall composite measure trained over 48 sentences and 12 waveform coder conditions yielded a high correlation of 0.954 with the DAM scores for the individual speakers (see Table 10), we used this measure in place of the DAM scores. To further justify this substitution, we submitted the individual-speaker DAM data and the individual-speaker objective measure (OCM) data computed over the 48 sentences to analysis by an MDS program called ALSCAL [24]. ALSCAL models rectangular matrix data as two sets of points, one for the rows of the matrix and one for the columns. In the present case, the coder conditions correspond to the rows and the speakers to the columns, and ALSCAL places the points in a multidimentional space, so that the distance from each coder point to each sentence point agrees as well as possible with the corresponding objective score. The data were treated as dissimilarities (large scores represent large distances), which had the effect of placing each point representing a sentence close to the points representing the coders that performed worst on that sentence. This seemed reasonable, since there are many perceptually distinct ways in which speech can be degraded, but only one way in which it can be undegraded. Two-dimensional solutions gave an excellent fit to the data, the values of Stress and RSQ (proportion of variance accounted for) being 0.087 and 99.8% for the DAM scores, and 0.023 and 99.9% for the composite objective measure. Agreement between the two solutions was measured by inspection and by giving each solution as an initial configuration for the other set of data, and running ALSCAL for one iteration to yield the Stress and RSQ values. Although the plots of the solutions were not identical when superimposed, they were similar. The formal comparisons yielded good agreement, with Stress being 0.175 and 0.179, and RSQ being 97.3% and 97.2%, respectively. This result shows that DAM scores and objective measure data appear to be interchangeable, at least within the context of these tests, and therefore justifies our using the objective measure as the basis for selecting the reduced subset of sentences.

Next, an ALSCAL analysis was performed on the objective measure data for the 48 individual sentences, using the same 12 waveform coder conditions. The two-dimensional solution gave an excellent fit to the data, the values of Stress and RSQ being 0.087 and 99.2%. The two-dimensional solution for the 48-sentence data is plotted in Figure 2. All but one of the coder points (identified by the coder mnemonics[2]) lie to the left of the Y-axis, and all but two of the sentence points (plotted using

[2] SQ in Figure 2 refers to APC-SQ in the quiet condition; SQ1 is 1% error; SQ5 is 5% error; SQSQ is self-tandem; SQCV is APCSQ-CVSD tandem; etc.

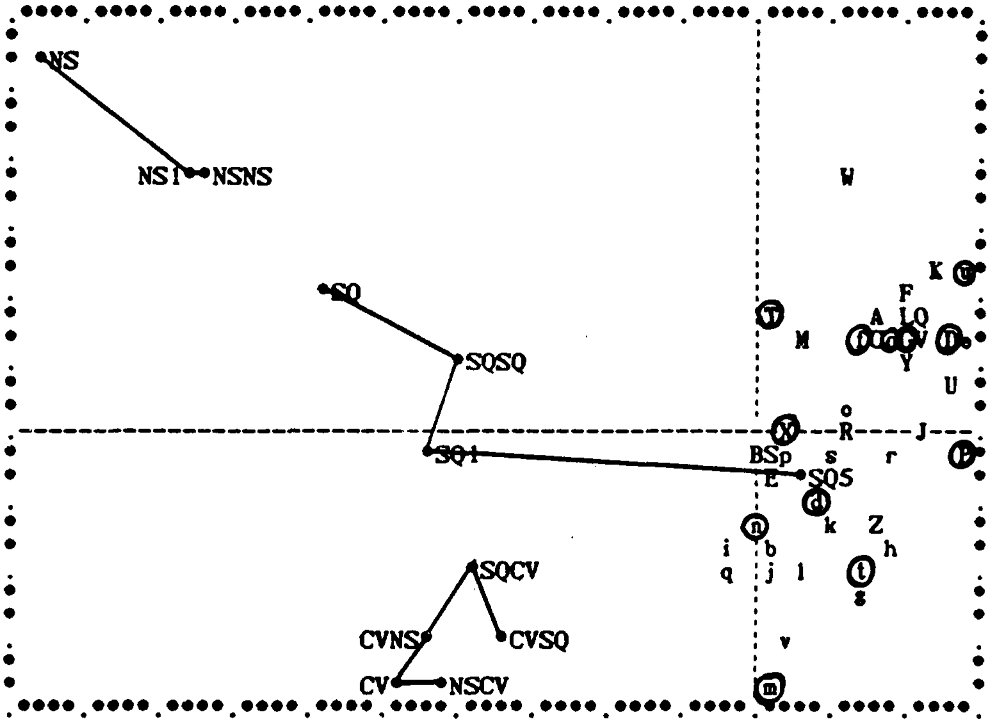


Figure 2. ALSCAL solution plot showing 12 waveform coder conditions and 48 DAM test sentences.

letters of the alphabet) lie to its right. We have identified in Figure 2 three clusters of coder points, by drawing a line through the points within a cluster.

We performed a new ALSCAL analysis, this time on the objective measures for the 25 all-coder conditions (see Section 6.2.2.3) across the 48-sentence database. The two-dimensional solution, plotted in Figure 3, again had an acceptably low value of Stress and of RSQ (0.136 and 98.2% respectively). Clusters of coder points are identified in Figure 3 by joining lines through points in each cluster.

Although the 12 phonetically chosen sentences from our initial database (see Section 6.2.1) were well distributed over the cluster of 48 sentence points in the solution plots in Figures 2 and 3, we discovered that the distribution could be considerably improved by substituting three new sentences for three of the original set. Each substitution, each for a different speaker, was made within the set of sentences for that speaker, so that the constraint that each speaker contribute two sentences was retained. Furthermore, each of the substitutions was perfectly acceptable on the basis of the informal phonetic analysis that led to the original selection. In the original phonemic analysis, often there were two or three candidates with very similar qualifications, and the choice between them was arbitrary. In each case, the substitutions we now made resulted from making a different choice from the same set of equivalent candidates. We replaced the three sentences RH2, VW3, and MP12 in Table 7 with the sentences RH4, VW7, and MP7:

| | |
|---|---|
| RH4 | That is the oldest wine. |
| VW7 | Music can calm the nerves. |
| MP7 | They want two red apples. |

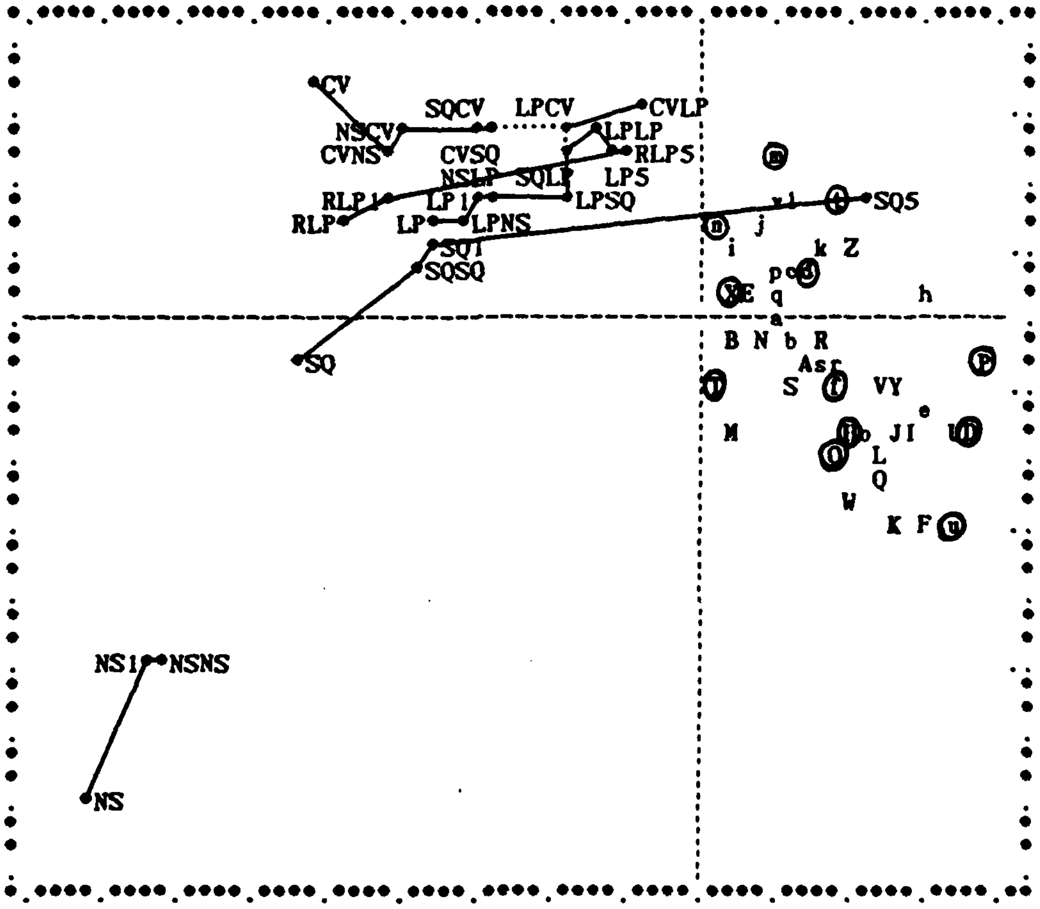


Figure 3. ALSCAL solution plot showing 25 all coder conditions and 48 DAM test sentences.

The 12 new selected sentences are marked by rings in Figures 2 and 3. (Minor differences in the identities of the marked sentence points are due to superimposed data points.)

## 6.4 Performance of Optimized Measures

For the results reported below, we optimized each composite measure over individual speakers and computed the corresponding all-speaker measure by averaging the individual-speaker measure over the 6 speakers. Also, we used the new 12-sentence database and the larger 48-sentence database for training and testing the measures. The correlation results are reported separately for waveform coders, vocoders, and all coders.

### 6.4.1 Waveform Coders

Table 18 gives both individual-speaker and all-speaker correlations for three different cases. The entry 48/12 in the column labelled "Database" means that optimization of the (individual-speaker) measures was performed over the 48 sentences, and the correlations given in that row were computed over the 12 sentences. From a study of the results in Table 18, we make several observations. First, differences in correlations between the two cases 48/48 and 12/12 (row 1 vs row 3) are quite small, which reinforces the validity of our selection of the 12 sentences as representative of the full set. Second, comparing the 48/12 and 12/12 cases (row 2 vs row 3), we find that the all-speaker correlations for the two cases are about equal, and the differences in individual-speaker correlations are small to modest. Thus, if prediction of individual-speaker DAM scores is important, we must use the coefficients of the composite measures optimized over the 12 sentences. Third, correlations for the FVSD measure are higher than those for the STB-SNR measure, and they are in general only slightly lower than the correlations for the OCM measure. Fourth, because of the way we computed the all-speaker measures (by averaging individual-speaker measures), occasionally the correlation for FVSD came out to be higher than that for OCM (see row 1). Finally, we note that the FVSD and OCM correlations are all excellent.

| No. | Database | Individual Speakers | | | All Speakers | | |
|---|---|---|---|---|---|---|---|
| | | FVSD | SNR | OCM | FVSD | SNR | OCM |
| 1 | 48/48 | 0.949 | 0.895 | 0.954 | 0.992 | 0.970 | 0.990 |
| 2 | 48/12 | 0.922 | 0.877 | 0.912 | 0.992 | 0.973 | 0.991 |
| 3 | 12/12 | 0.947 | 0.889 | 0.947 | 0.991 | 0.963 | 0.991 |

TABLE 18. Individual-speaker and all-speaker correlations for the 12 waveform coder conditions.

### 6.4.2 Vocoders

Table 19 gives individual-speaker and all-speaker correlations for 5 different cases. Recall that we obtain the 8 vocoder conditions (rows 1, 3, and 4) from the earlier set of 10 by discarding the two tandems (LPC-APCSQ and LPC-APCNS) that gave inconsistent DAM test scores, and we obtain the 6 vocoder conditions (row 2) by discarding in addition the two channel error conditions of LPC-10 (1% and 5%). Row 5 in Table 19 corresponds to the above 8 vocoder conditions plus the two tandems RELP-LPC10 and LPC10-RELP. Several comments are in order. First, correlations for vocoders are in general substantially lower than those for waveform coders given in Table 18. While the individual-speaker correlations in Table 19 are all significant at the 0.001 level or better, many of the all-speaker correlation scores (those marked by asterisks) are not significant even at the 0.01 level. Second, as expected, correlations for 8 vocoders are lower than those for 6 vocoders. In fact for 6 vocoders, both FVSD and OCM measures produce reasonable correlations for individual speakers and very high correlations for all speakers. Third, the differences in correlations between the 48/48 and 12/12 cases (row 1 vs row 4) and between the 48/12 and 12/12 cases (row 3 vs row 4) are generally small for all speakers and are small to modest for individual speakers. Fourth, the OCM measure provides in general a modest improvement over FVSD. Finally, we note that adding the 2 RELP tandems to the 8 vocoders lowers the correlations slightly.

Next, we offer our reasoning as to why the correlations are low for vocoders. We have already mentioned above the two channel-error conditions. We believe that to be applicable to channel-error conditions, the objective measures should average over only the top 20 or 30% of the frame-by-frame distances [15], instead of averaging over all the frames. Thus, with this modification, we expect that the correlations for vocoders should be similar to those given in row 2 of Table 19. It is our conjecture that these correlations would increase further if we did not have any tandem conditions. (Five out of the six vocoder conditions are tandem links.) In effect, we are saying that the present objective measurement procedure must be modified to be

| No. | Coders | Database | Individual Speakers | | All Speakers | |
|---|---|---|---|---|---|---|
| | | | FVSD | OCM | FVSD | OCM |
| 1 | 8 vocoders | 48/48 | 0.520 | 0.663 | 0.603* | 0.782 |
| 2 | 6 vocoders | 48/48 | 0.747 | 0.800 | 0.949 | 0.939 |
| 3 | 8 vocoders | 48/12 | 0.479 | 0.475 | 0.593* | 0.804 |
| 4 | 8 vocoders | 12/12 | 0.576 | 0.635 | 0.612* | 0.747* |
| 5 | 10 vocoders | 12/12 | 0.561 | 0.628 | 0.604* | 0.682* |

TABLE 19. Individual-speaker and all-speaker correlations for the vocoder class.

applicable to tandem links (see Chapter 10). (We note that Barnwell's distorted database did not contain any channel-error or tandem-link conditions [11].) To see if removing the tandem-link waveform coder conditions improves the correlations, we optimized the 3 composite measures, for the individual-speaker case, over the six single-link waveform coder conditions and the 48 sentences. The resulting correlations are 0.957, 0.929, and 0.960, respectively, for FVSD, STB-SNR, and OCM. These are slightly higher than those given in row 1 of Table 18.

### 6.4.3 All Coders

Table 20 gives individual-speaker and all-speaker correlations for the all-coder class. As noted above, the differences in correlations between the 48/12 and 12/12 cases are significant only for the individual-speaker case. Also, we observe that the differences in correlation between FVSD and OCM are small.

| No. | Coders | Database | Individual Speakers | | All Speakers | |
|---|---|---|---|---|---|---|
| | | | FVSD | OCM | FVSD | OCM |
| 1 | 25 coders | 48/48 | 0.882 | 0.886 | 0.935 | 0.944 |
| 2 | 25 coders | 48/12 | 0.801 | 0.785 | 0.931 | 0.940 |
| 3 | 25 coders | 12/12 | 0.859 | 0.867 | 0.915 | 0.927 |

TABLE 20. Individual-speaker and all-speaker correlations for the all-coder class.

### 6.4.4 Performance Over Subsets of a Training Set

In practice, one is interested in finding out how a composite measure trained over several types of coders performs over a subset of these coders, especially if the application requires the objective evaluation of the coders from this subset. Performance comparisons between the subsets and the training set are somewhat related to the training versus test performance comparisons, which is the topic of the next section.

We considered the all-coder FVSD and OCM measures trained over the 25 coder conditions and 48 sentences (see Section 6.4.3), and evaluated over several subsets. The results are given in Table 21. Numbers in parentheses are maximum correlations over the respective subsets (which will be achieved if we train the measure over the respective subset). We wish to make three observations. First, for subsets involving waveform coders (Nos. 2 and 3), the subset performance is superior to the full-set performance. (This is also true for the all-speaker case of subset No. 7.) For all other subsets, the subset performance is inferior to the full-set performance. Second, for the same subsets involving waveform coders, the differences between the actual subset performance and the maximum possible subset performance (numbers in parentheses) are rather small. For all other subsets, such performance differences are large. Third, when the training set involves a variety of quite different types of coders (as it is the case in this example), the performance over individual subsets varies a lot (e.g., the correlation for the individual-speaker OCM varies from 0.464 to 0.943). This variability does not necessarily mean that the particular measures lack robustness.

| No. | Subset | Individual Speakers | | All Speakers | |
|---|---|---|---|---|---|
| | | FVSD | OCM | FVSD | OCM |
| 1. | Full set of 25 coder conditions | 0.882 | 0.886 | 0.935 | 0.944 |
| 2. | 12 waveform coder conditions | 0.945<br>(0.949) | 0.943<br>(0.954) | 0.988<br>(0.992) | 0.990<br>(0.990) |
| 3. | 12 waveform coder and 3 RELP coder conditions | 0.930<br>(0.936) | 0.933<br>(0.945) | 0.977<br>(0.982) | 0.985<br>(0.988) |
| 4. | 10 vocoder conditions | 0.446<br>(0.497) | 0.464<br>(0.560) | 0.514*<br>(0.528*) | 0.552*<br>(0.649*) |
| 5. | 6 vocoder conditions | 0.626<br>(0.747) | 0.609<br>(0.800) | 0.873<br>(0.949) | 0.871<br>(0.939) |
| 6. | 10 vocoder and 3 RELP coder conditions | 0.596<br>(0.610) | 0.621<br>(0.625) | 0.676 | 0.728 |
| 7. | 3 RELP coder conditions | 0.868 | 0.866 | 0.985 | 0.985 |

TABLE 21. Correlation coefficients of all-coder composite measures evaluated over subsets of coder conditions. Numbers in parentheses are maximum correlations over respective subsets. The numbers marked by asterisks fail to achieve significance at the 0.01 level.

## 6.5 Performance Comparisons Over Training and Test Sets

The results presented below show that the various objective measures we considered have relatively stable performance over training and test sets provided that the two sets are similar. Of course, if the test set is completely different from the training set, then we expect that the test-set performance will be severely degraded compared to the training-set performance. In the experiments described below, we chose the training and test sets to be reasonably similar.

### 6.5.1 Waveform Coders

We performed 3 experiments. First, we considered the 2 subsets, denoted by S1 and S2. Each set has all 12 waveform coder conditions but contains a different set of 3 speakers. We trained or optimized the composite measures on one set and tested or evaluated them on another. Table 22 gives the results. From the table, we observe that the all-speaker performance over training and test sets is more stable than the individual-speaker performance, and the performance degradation in the latter case is quite small for FVSD and OCM and is large but not drastic for STB-SNR.

In the second experiment, we considered the 12 waveform coder conditions, and divided them into a 9-coder training set and a 3-coder test set, each set containing all 6 speakers. The 3 conditions for the test set are: APCNS-APCNS, CVSD quiet, and APCSQ in 1% error. The training set has the remaining 9 conditions. Correlations over the training and test sets are given in Table 23. Considering the all-speaker case, the test-set performance is slightly better for FVSD and STB-SNR and is modestly worse for OCM, all compared to the training-set performance. The performance degradation for the individual-speaker case is large but not drastic.

| Test Set | Training Set | FVSD | STB-SNR | OCM |
|---|---|---|---|---|
| S1 | S1 | 0.991<br>(0.949) | 0.951<br>(0.887) | 0.993<br>(0.951) |
| | S2 | 0.981<br>(0.912) | 0.969<br>(0.806) | 0.979<br>(0.905) |
| S2 | S1 | 0.977<br>(0.943) | 0.922<br>(0.769) | 0.980<br>(0.943) |
| | S2 | 0.985<br>(0.976) | 0.967<br>(0.957) | 0.982<br>(0.978) |

TABLE 22. All-speaker and individual-speaker (numbers within parentheses) correlations over training and test sets, involving the sets S1 and S2 (waveform coders).

In the third experiment, we considered the 12 waveform coder and the 3 RELP coder single-link conditions, and divided them into a 10-coder training set and a 5-coder test set. The 5 conditions for the test set are: APCNS-APCNS, CVSD quiet, APCSQ in 1% error, RELP in 1% error, and CVSD-APCNS. The results are given in Table 24. The performance degradation of each measure is small for the all-speaker case. For the individual-speaker case, STB-SNR exhibits a large degradation, while the other two measures degrade only modestly.

| | Individual Speakers | | | All Speakers | | |
|---|---|---|---|---|---|---|
| | FVSD | SNR | QCM | FVSD | SNR | QCM |
| Training Set | 0.963 | 0.913 | 0.968 | 0.998 | 0.984 | 0.997 |
| Test Set | 0.887 | 0.750 | 0.861 | 0.9993 | 0.9995 | 0.927 |

TABLE 23. Training vs test performance comparison for the 12 waveform coder conditions.

| | Individual Speakers | | | All Speakers | | |
|---|---|---|---|---|---|---|
| | FVSD | SNR | QCM | FVSD | SNR | QCM |
| Training Set | 0.950 | 0.900 | 0.959 | 0.991 | 0.972 | 0.996 |
| Test Set | 0.885 | 0.747 | 0.913 | 0.980 | 0.963 | 0.988 |

TABLE 24. Training vs test performance comparison for the 12 waveform coder and 3 RELP coder conditions.

### 6.5.2 Vocoders

For this case, we considered the 8 vocoder conditions and divided them into a 6-coder training set and a 2-coder test set; the latter has the 2 conditions: APCNS-LPC10 and CVSD-LPC10. The results are given in Table 25. We did not compute the all-speaker correlations for the test set since there are only 2 items to correlate over. The test-set performance for the individual-speaker case was substantially better for FVSD and slightly worse for OCM than the training-set performance. Given that there are only 8 coder conditions and that these conditions are fairly different from each other, we cannot make meaningful statements regarding training-set vs test-set performance changes.

| | Individual Speakers | | All Speakers | |
|---|---|---|---|---|
| | FVSD | OCM | FVSD | OCM |
| Training set | 0.436 | 0.691 | 0.485 | 0.911 |
| Test set | 0.653 | 0.680 | - | - |

TABLE 25. Training vs test performance comparison for the 8 vocoder conditions.

### 6.5.3 All Coders

We considered the 25 coder conditions, and divided them into an 18-coder training set and a 7-coder test set. The test set has the following 7 conditions: APCNS-APCNS, APC-SQ quiet, RELP in 1% error, LPC10-CVSD, LPC-10 in 5% error, CVSD-APCNS, and APCSQ-LPC10. The results are given in Table 26. Because the database for the all-coder class is reasonably large, we have confidence in the results given in Table 26. It is clear from this table that the performance of both FVSD and OCM is quite stable over the training and test sets.

| | Individual Speakers | | All Speakers | |
|---|---|---|---|---|
| | FVSD | OCM | FVSD | OCM |
| Training set | 0.884 | 0.888 | 0.941 | 0.949 |
| Test set | 0.882 | 0.880 | 0.943 | 0.949 |

TABLE 26. Training vs test performance comparison for the 25 all-coder conditions.

## 6.6 Classification of the RELP Coder Conditions

The RELP coder is a hybrid between the waveform coder and the vocoder. Thus, we have the option of treating it as a waveform coder or as a vocoder. In deciding which of these two classifications is better, we should examine two issues. First, how much the inclusion of the RELP coder conditions into each class degrades the performance of the composite measures over the coder conditions that are already included in that class? Second, what is the performance of the composite measures over the RELP coder conditions? The classification that produces lower degradation (first issue) and higher performance (second issue) is the one we should adopt.

As given in Table 5 (Chapter 5), we have a total of eight RELP coder conditions, of which the two tandem conditions involving LPC-10 are clearly vocoder conditions. The question, then, is how should we classify the remaining six RELP coder conditions (which are marked as HYB in Table 5). We have two cases corresponding to the two ways of classifying these six RELP conditions:

CASE A: Treat the six RELP conditions as waveform coder conditions, and optimize the measures over the 12 waveform and six RELP coder conditions and over the 12 sentences. Correlations are given in Table 27.

CASE B: Treat the six RELP conditions as vocoder conditions, and optimize the measure over the 10 vocoder[3] and six RELP coder conditions and over the 12 sentences. Correlations are given in Table 28.

| Measure | Individual Speakers | All Speakers |
|---|---|---|
| FVSD | 0.903 (0.947) | 0.965 (0.991) |
| STB-SNR | 0.881 (0.889) | 0.955 (0.963) |
| OCM | 0.934 (0.947) | 0.985 (0.991) |

TABLE 27. Correlation results for the set of 12 waveform and 6 RELP coder conditions. Numbers in parentheses are for 12 waveform coder conditions alone.

The numbers in parentheses in Table 27 are the optimized correlations over the 12 waveform coder conditions alone. We see clearly that adding the six RELP coder conditions to the 12 waveform coder conditions changes the correlations only slightly. We then evaluated the Case A measures separately over the 12 waveform coder conditions and over the six RELP conditions. The results are given in Table 29. Similarly, we evaluated the Case

[3] We have excluded the LPC10-APCNS and LPC10-APCSQ tandems as their DAM scores are inconsistent.

B measures over the 10 vocoder conditions and over the six RELP coder conditions. The results are given in Table 30.

| Measure | Individual Speakers | All Speakers |
|---|---|---|
| FVSD | 0.608 (0.561) | 0.604 (0.604) |
| OCM | 0.639 (0.628) | 0.812 (0.682) |

TABLE 28. Correlation results for optimization over the 10 vocoder and the six RELP coder conditions. Numbers in parentheses are optimized correlations for vocoder conditions alone.

|  | 12 Waveform Coders |  | 6 RELP Coders |  |
|---|---|---|---|---|
| Measure | Individual Speakers | All Speakers | Individual Speakers | All Speakers |
| FVSD | 0.942 | 0.988 | 0.697 | 0.891 |
| STB-SNR | 0.870 | 0.948 | 0.843 | 0.984 |
| OCM | 0.942 | 0.989 | 0.822 | 0.955 |

TABLE 29. Correlation results of Case A measures over 12 waveform coders and over six RELP coders.

Comparing the results given in Table 30 with those in parentheses in Table 28, we find that the inclusion of the six RELP conditions into the vocoder class results in a substantial decrease in the correlations over the 10 vocoders: a reduction in the correlation of OCM from 0.628 to 0.412 for individual speakers and from 0.682 to 0.377 for all speakers. On the other

| | 10 Vocoders | | 6 RELP Coders | |
|---|---|---|---|---|
| Measure | Individual Speakers | All Speakers | Individual Speakers | All Speakers |
| FVSD | 0.458 | 0.473 | 0.692 | 0.886 |
| OCM | 0.412 | 0.377 | 0.508 | 0.652 |

TABLE 30. Correlation results of Case B measures over 10 vocoders and over six RELP coders.

hand, comparing the results given in Table 29 with the numbers given within parentheses in Table 27, we find that the inclusion of the six RELP conditions into the waveform coder class produces only a minor reduction in correlation over the 12 waveform coders. Considering the correlation over the six RELP coders, we see from Tables 29 and 30 that Cases A and B produce similar results for FVSD, and Case B produces a substantially lower correlation for OCM than does Case A. Thus, it is very clear that we should classify the six RELP conditions as belonging to the waveform coder class.

## 7. TIME DELAY ESTIMATION AND SYNCHRONIZATION

In the work reported thus far, we used the file-to-file version of each real-time coder, and achieved synchronization of the coder output speech with the input speech by compensating for the known processing delay of the coder. In this chapter, we report our work on the more general synchronization problem involving real-time speech coders and unknown coder processing delay. Below, we state the synchronization problem and outline our proposed approach (Section 7.1); present experimental results dealing with the sensitivity of objective measures to simulated synchronization errors (Section 7.2); describe our design of the coder input signal to be used for estimating the time delay introduced by the coder (Section 7.3); present the time delay estimation algorithm (Section 7.4); and present some experimental results we obtained using this algorithm on both file-to-file coders and real-time coders (Section 7.5).

### 7.1 Statement of Problem and Our Approach

The problem of synchronizing a real-time coder's output speech with its input speech is the same as the problem of estimating the time delay introduced by the coder and then compensating this delay by shifting back the output speech relative to the input speech. Here we have tacitly assumed that the time delay is fixed, i.e., not time-varying. This assumption is true for the five real-time coders considered in our investigation. The time delay for these coders is a sum total of delays introduced by the coding algorithm (e.g., filtering required in high-frequency regeneration in RELP [10]) and delays introduced by the particular implementation (e.g., A/D and D/A buffering, transmit and receive buffering, etc.).

The problem of estimating the time delay between two signals has been investigated extensively in the sonar signal processing area [25]. In that application, each signal is received at a different sensor, and the estimated delay is used in computing the bearing of the target source. Each sensor signal is modeled as a delayed version of the target source signal corrupted by an additive noise that is uncorrelated with the noise of the other sensor signal. The time-delay estimation approach is to cross-correlate the two sensor signals and declare the lag at which the cross-correlation function is maximum as the estimated delay.

In our case, the time-delay estimation problem is simpler in one sense and harder in another sense than the sonar problem. We have the freedom of designing the input signal in a way that makes the problem easier to deal with or leads to a more reliable estimate or does both. On the other hand, the cross-correlation approach may not be directly applicable for non-waveform coders like the LPC-10, since its output cannot be reasonably modeled as its input plus noise. A modified cross-correlation approach for this case may be to cross-correlate time functions of short-term characteristics (energy, formant frequencies, etc.) of the input and the output signals, since the LPC-10 attempts to preserve these characteristics in the output.

## Our Proposed Approach to the Synchronization Problem:

In our approach, we apply as input to the coder under investigation a specially designed time delay estimation (TDE) signal and record simultaneously the coder input and output, using a two-channel tape recorder. We then digitize the input and output signals on the two-channel tape simultaneously, using a two-channel A/D facility. The TDE algorithm compares the digitized output signal with the digitized input signal, and provides an estimate of the coder time delay. Finally, the estimated time delay is used in synchronizing the coder output speech with its input by simply shifting the output backwards relative to the input.

We note that the input (or source) tape for objective speech quality evaluation consists of the TDE signal followed by the selected 12 DAM test sentences. The two-channel tape-recording and the two-channel digitization, mentioned above, are performed for the complete input tape (see Appendix for details). By simultaneously tape recording and simultaneously digitizing the coder input and output signals, we avoid the problems of the single-channel approach in which we tape or digitize one signal at a time; these problems include stretching of the tape, variations in tape recorder speed, and errors in locating the beginning of each signal. Re-recording the input signal, which is required in the two-channel approach, might slightly degrade the signal, but we believe that the effect, if any, of this re-recording on time delay estimation or on objective measure computation should be negligible.

Before we describe the design of the TDE signal and present the TDE algorithm, we present some experimental results dealing with the sensitivity of the composite objective measures to simulated errors in synchronization (i.e., errors in the estimated delay). Clearly, such sensitivity data will give an idea of the required accuracy of the estimated time delay.

## 7.2 Sensitivity of Objective Measures to Simulated Synchronization Errors

To examine the effect of errors in the estimated value of the coder time delay on the performance of objective measures, we introduced known errors, recomputed first the frame-by-frame measures and then the composite objective measures using the coefficients of the no-error case (perfect time-delay estimation), and evaluated the correlations with the DAM test scores. We used the file-to-file coders for this study. For waveform coders, we considered an error in the time delay of five samples, as we assumed we could estimate the time delay of these coders very accurately. The resulting correlations over 12 waveform coder conditions and 3 RELP coder single-link conditions are given in Table 31, along with those for the no-error case for comparison. The correlations of FVSD and OCM change only slightly under error, but, as expected, the correlations of STB-SNR decrease drastically. (We note that the dominant component in the OCM considered is FVSD and not STB-SNR; if the latter were dominant, we would have seen a larger sensitivity for OCM.)

| | Individual Speakers | | All Speakers | |
|---|---|---|---|---|
| Measure | No Error | Error | No Error | Error |
| FVSD | 0.909 | 0.911 | 0.965 | 0.966 |
| STB-SNR | 0.882 | 0.590 | 0.960 | 0.659 |
| OCM | 0.938 | 0.910 | 0.991 | 0.980 |

TABLE 31. Performance of objective measures over 12 waveform and three RELP coder conditions, under a simulated five-sample error in the coder time delay.

For the vocoder class, we considered the eight vocoders (we excluded the tandems LPC10-APCNS, LPC10-APCSQ, LPC10-RELP, and RELP-LPC10), simulated a time-delay error of 100 samples (12.5 ms at 8 kHz and about 15.1 ms at 6.621 kHz), and, as above, computed the correlations. The results are given in Table 32. Again, the performance change caused by the error in the delay is only small. (That the correlations of FVSD increase slightly in error is not statistically significant.)

| | Individual Speakers | | All Speakers | |
|---|---|---|---|---|
| Measure | No Error | Error | No Error | Error |
| FVSD | 0.576 | 0.577 | 0.612 | 0.631 |
| OCM | 0.635 | 0.617 | 0.747 | 0.728 |

TABLE 32. Performance of objective measures over 8 vocoders, under a simulated 100-sample error in the coder time delay.

We then considered the all-coder class of 25 conditions: 12 waveform coders, 3 (single-link) RELP coders, and 10 vocoders. We assumed a time-delay error of 5 samples for waveform and RELP coders and of 100 samples for vocoders. The correlation results are given in Table 33. We note that both FVSD and OCM produced higher correlation in error than in no error; the changes, however, are not significant.

From the results reported above, we infer that both FVSD and OCM are well behaved and produce only a small drop in correlation, under "reasonable" errors in the estimated coder time delay; but, STB-SNR is highly sensitive to such errors.

| | Individual Speakers | | All Speakers | |
|---|---|---|---|---|
| Measure | No Error | Error | No Error | Error |
| FVSD | 0.859 | 0.873 | 0.915 | 0.928 |
| OCM | 0.867 | 0.882 | 0.927 | 0.940 |

TABLE 33. Performance of objective measures over 25 all-coder conditions, under a simulated time-delay error of 5 samples for waveform and RELP coders and of 100 samples for vocoders.

## 7.3 Design of the Input Signal for Time Delay Estimation

It is intuitively clear that the input signal should be speech-like and not be "foreign" to the coder, since this should avoid any possible pathological situations and since the coders have been designed specifically for speech.

Should we generate an input signal that is completely voiced or is completely unvoiced or has both types of segments as in speech? A careful analysis indicates that we should use unvoiced signals. There are several reasons for this decision. First, frame-oriented coders that use a pitch-excited source (e.g., LPC-10) will use an average pitch over the frame, and hence the pitch pulses in the output speech will not, in general, be located the same way as in the input speech. Second, coders like LPC-10 make voicing errors. Third, for a correctly identified voiced frame, estimated pitch values can be in error.

Having decided to generate unvoiced signals, we then investigated the use of multiple-segment or bursty signals. Such a signal has several alternating segments of low and high energies. Durations of individual segments are different and are chosen randomly. Each transition between segments can be used to estimate the time delay (by comparing that transition in the coder output with the transition in the coder input). With the resulting multiple time-delay estimates, we can select the most robust one or obtain a more reliable estimate through some averaging technique. The averaging idea may be important for non-waveform coders like LPC-10, since the individual estimates are prone to error because of the frame-oriented nature of these coders.

We then developed an interactive program for generating the input signal. The program has a user-specified basic frame size, and it determines the lengths of the low-energy and high-energy regions as randomly chosen integer multiples of this frame size. Random region lengths should allow coder framing and processing to influence each transition differently. The user also specifies the minimum and maximum lengths of the low-energy and high-energy regions and the sampling rate. The lower limit on the region length must be large enough to adequately separate one transition from another (which is necessary to minimize confusion in locating transitions) and to provide a sufficient number of samples around each transition for the TDE algorithm processing (see Section 7.4). The upper limit on the region length helps to minimize the computational effort required by the TDE algorithm. The individual samples of the two types of regions are generated using a zero-mean uniformly distributed random number generator. The generated samples can be optionally filtered through an all-pole filter to obtain a speech-like signal. The user has the option of specifying a fixed all-pole filter and a fixed mean-square value in dB, one for each region; or he can vary one or more of these quantities as a function of the frame number.

Using the above-mentioned interactive program, we generated a number of TDE signals and tested them on a number of file-to-file coder conditions. We used a basic frame size of 12 ms, a minimum region length of 24 frames, a maximum region length of 30 frames, and a sampling rate of 6.67 kHz. For all-pole filtered TDE signals, we used a fixed two-pole filter (complex pole at 500 Hz with a bandwidth of 200 Hz). For the purpose of these experiments, we manually located the transitions in the coder output, chose a region around each transition, and cross-correlated this region with the corresponding one in the input. The cross-correlation approach worked well for waveform and RELP coders, but was not as good for vocoders, as expected. For vocoders, the time delay was estimated by directly comparing the located output transitions with the input transitions. The results from these experiments helped us in the design of the TDE signal and also provided insights for developing the TDE algorithm. The results relating to the design of the TDE signal are summarized below.

All-pole filtering of the random signal caused or increased confusion in locating the energy transitions. Therefore, we decided against using an all-pole filter. Let E1 denote the mean-square value in dB of the low-energy region and E2 denote the corresponding quantity of the high-energy region. With a value of E1=40 dB, we found that using E2=45 dB produced good delay estimates for vocoders; transitions could be identified more readily when we used E2=50 dB for waveform coders. We note that with E2=50 dB, LPC-10 synthesized the high-energy regions as voiced segments.

The TDE signal we finally chose is shown in Figure 4. It has an overall duration of about 5 seconds. The first second of

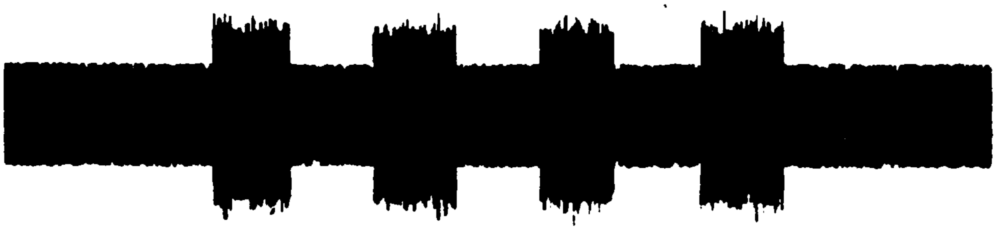


Figure 4. The time delay estimation signal containing alternating low-energy and high-energy random noise segments.

the signal contains a single low-energy region. This section of the signal is provided to collect all coder startup characteristics, such as acquiring sync. This section of the processed signal must be discarded so that any distortions caused by coder startup will not disrupt the TDE analysis. The last second of the signal also contains a single low-energy region, and is provided to assure that all energy transitions, when delayed by a coder, are present in the coder output signal.

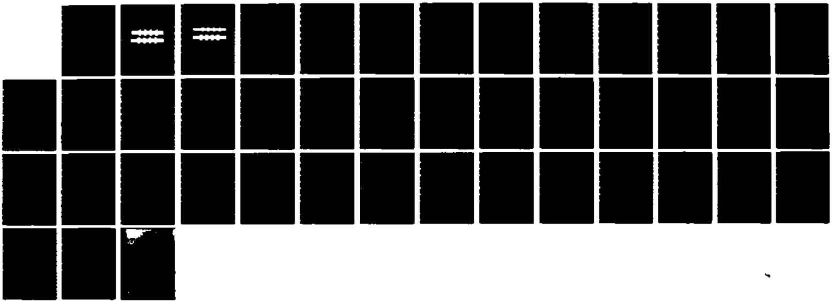
The middle three-second section of the signal contains alternating low and high energy regions, with a total of four high energy regions, which provides eight energy transitions for delay estimation. Tests showed that multiple time-delay estimates (one for each energy transition) can be used to extract one value that is robust. We had observed, for example, that errors in the location of several of the transitions occurred. Some of these errors occurred in the first transition, which indicates that a single energy transition would not be adequate.

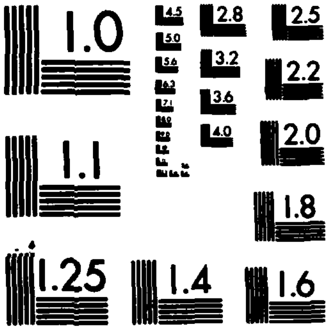
Figure 5 provides a comparison of the input TDE signal (lower plot) and the corresponding output signal (upper plot) from the APC-NS coder operating in the quiet condition. While some smearing of the transitions is evident in the output signal, it seems rather simple to locate the transitions (at least manually). Figure 6 provides a similar comparison for LPC-10 operating in 5% channel error. The transitions in the LPC-10 output signal are much more smeared than we see in Figure 5. Also, channel error has caused noticeable energy changes in the output signal, which might cause problems in automatically locating the transitions.

To be used with real-time coders, the TDE signal shown in

AD-A141 194 OBJECTIVE SPEECH QUALITY EVALUATION OF REAL-TIME SPEECH 2/2
CODERS(U) BOLT BERANEK AND NEWMAN INC CAMBRIDGE MA
V R VISWANATHAN ET AL. FEB 84 BBN-5504 DCA100-82-C-0005

UNCLASSIFIED F/G 9/4 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

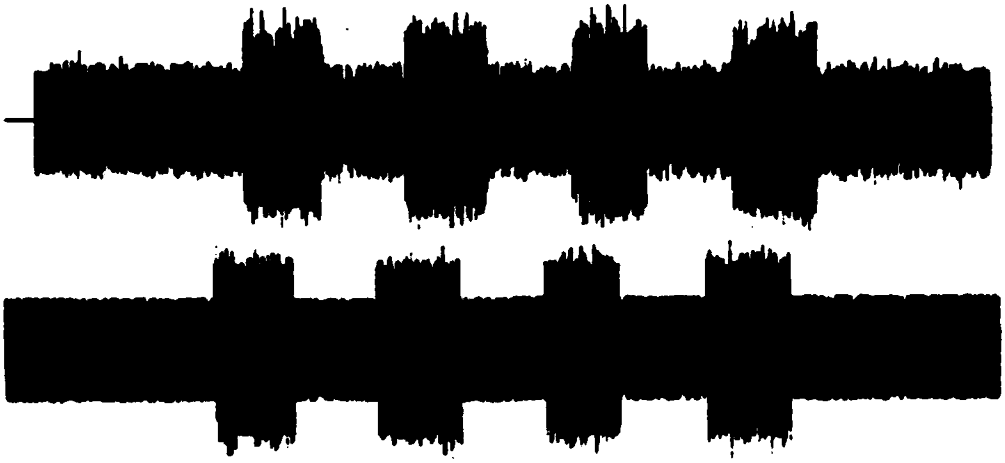


Figure 5. The input time delay estimation signal (lower plot) and the corresponding output (upper plot) obtained from the APC-NS coder operating in the quiet condition.

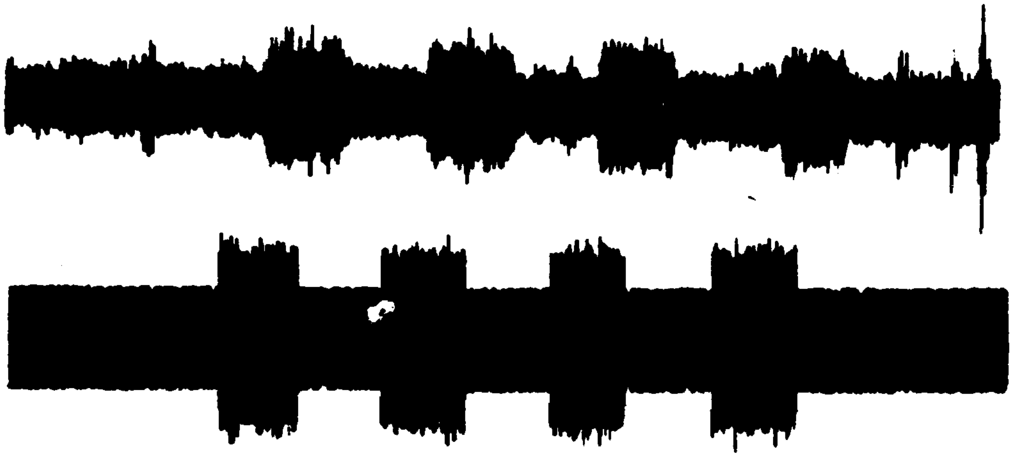


Figure 6. The input time delay estimation signal (lower plot) and the corresponding output (upper plot) obtained from the LPC-10 coder operating in 5% channel error.

Figure 4 must be preceded by an initialization signal, to help properly initialize any of the adaptive mechanisms of the coder. For example, LPC-10 uses adaptive or time-evolving thresholds in its pitch and voicing algorithm. Without such initialization signal, the real-time LPC-10 coder synthesized the output TDE signal as an all-voiced signal, which would surely produce incorrect time delay estimates. (For the file-to-file LPC coder, we applied the TDE signal file twice, treated the first application as initialization, and used the second output file for TDE processing.) Experimentally, we found that a random-noise initialization signal produced the desired results. We synthesized this signal digitally using the same program that generated the TDE signal (but without any transitions) and using a mean-square value of 60 dB.

Finally, we recommend the use of two TDE signals, one for waveform coders and another for vocoders and all coders. The first TDE signal has E1=40 dB and E2=50 dB, and the second one has E1=40 dB and E2=45 dB.

## 7.4 Time Delay Estimation Algorithm

Our algorithm for estimating the coder time delay consists of the three steps listed below. We assumed that the two-channel tape containing the input TDE signal and the corresponding coder output has been digitized using a two-channel A/D facility, to produce two disk files containing the digitized versions of the input and the output TDE signals. (For the file-to-file case, we need only to process the digitally synthesized TDE signal through the file-to-file coder to obtain the required output signal.)

### Steps in the TDE Algorithm

1. Determine the start point of the TDE signal on the input file.
2. Locate the eight transitions on the output signal.
3. Compare the transitions on the output signal with those on the input signal, compute eight time delay estimates, and produce one value as the desired time delay estimate.

Clearly, Step 1 is required only for real-time coders. (As we use the digitally synthesized input TDE signal for the file-to-file case, the start point and the location of the transitions are known exactly.) Because of analóg taping and cuing in the A/D process, the start point of the input TDE signal has to be estimated. We will explain how this is done at the end of this section, as we actually use the same procedure required for Step 2.

For Step 2, we first compute a frame-by-frame "Energy Contour Buffer" or ECB for the output signal, and from the ECB, we determine the best estimate of the frame at which each transition occurs. These two issues are described below.

The energy contour buffer gives a preliminary estimate, on a frame basis, of the high and low energy regions in the TDE signal. We use as frame size the same 12 ms value we used in generating the TDE signal (see Section 7.3), so that the ECB analysis frame boundaries and the energy transition edges of the TDE signal will approximately coincide. (Exact alignment may not be possible because of analog taping and A/D processing.) High energy region frames are indicated by a value of one in the ECB, and low-energy frames by zeros. The ECB is computed as follows.

We first discard the first second of the output TDE signal, as mentioned above in Section 7.3, to eliminate coder startup distortions. Then an average of the frame energies is computed for the first low-energy region. A threshold is then computed as a predetermined constant (we used 1.5 for vocoders and 2.0 for waveform coders and all coders) times this average energy. Zeros are stored in the ECB for the frames that were used in obtaining this average. For the rest of the TDE signal, each frame energy is compared with the threshold. We insert a one in the ECB if a frame energy is above the threshold and a zero if the energy is below the threshold.

From the ECB, we next determine the frame at which each transition occurs. Ideally, the ECB should contain contiguous sequences of zeros and ones indicating the region locations. However, because of local variations of energy caused by channel error and coder framing and processing, several frames may exist in the vicinity of each transition as candidate locations of an energy transition. Our algorithm collects together possible transition frames for each transition. For the occurrence of a transition the ECB is searched from the beginning for the first frame containing a one. If this frame is followed by at least 5 frames of zeros, it is discarded as a fluke, and the next frame in the buffer containing a one is located. As soon as the 5-frame test fails, all frames containing a one preceded by a zero are located for the next 10 consecutive frames. The search for these frames will be aborted before the 10-frame limit if five consecutive frames of ones are found, indicating a solid detection of the first high-energy region. The second transition can be approximately located from the data (group of possible transition frames) of the first transition since the first high-energy region length is known. The first frame in the group of possible transition frames of the first transition is used to locate the start frame of the second transition. A search over 10 frames around the start frame of the second transition is made to collect all possible frames containing zeros preceded by frames containing ones. (Remember that the second transition is from high to low energy.) The search is aborted prematurely if five consecutive frames of zeros are found, indicating a solid detection of the next low-energy region. This procedure is repeated for locating the rest of the six transitions. We then have a group of possible transition frames for each of the 8 transitions.

Next, we determine a single frame as the best location of each transition. First, we identify the three transitions with the smallest group of possible transition frames. A search is then made for the set of three frames, one from each of the three chosen transitions, which provides the least error. An error can be computed for each set since the correct durations between transitions are known. One of the frames from the chosen set is then used to find the best single frame for each of the remaining 5 transitions (best in the sense of least error).

Step 3 of our algorithm is to compute the 8 delay estimates and extract one value as the coder delay. This step is performed differently for waveform coders, vocoders, and all coders.

For waveform coders a cross-correlation of the input and output TDE signals is computed around each transition. A 1000-sample window is centered at the start of the chosen frame for each transition. We compute 501 lags (250 positive, 250 negative, and one at zero lag) for the 8 correlation functions, one for each transition. Each correlation function is normalized with the square-root of the product of the autocorrelations of the input and output TDE signals computed over the 1000-sample window. The location of the peak of each normalized correlation function gives the delay estimate for that transition. The delay estimate corresponding to the highest peak correlation is chosen as the final delay estimate.

For vocoders, the delay is determined directly by comparing each of the 8 chosen transition frames for the output signal with the corresponding one for the input signal. Eight delays are computed, one for each transition. The delay is defined as the difference in samples between the transition locations in the coder input and output TDE signals. The median of these 8 delay estimates is declared as the coder delay.

For the all-coder class, the cross-correlation functions are computed as in the waveform coder case. The maximum peak-correlation value is then compared with a threshold. If the peak correlation exceeds the threshold, the coder delay is taken as the value corresponding to this peak; if the peak is below the threshold, the delay is computed as in the vocoder case. The threshold for the peak correlation was empirically determined as 0.2. Our experiments showed that for vocoders the peak correlation obtained was approximately 0.1, and for waveform coders the minimum value of the highest peak correlation was approximately 0.25.

Now, let us return to Step 1 of the TDE algorithm, which is required for real-time coders, as we mentioned above. Once the start point of the input TDE signal is determined, the locations of the 8 transitions are easily computed as we know these relative to the start point. The procedure for estimating the start point is as follows. First, we apply on the input signal the same procedure we described above for Step 2 to locate the energy transitions approximately. We then correlate a section of the signal around the first transition with the corresponding section of the original, digitally synthesized input signal. (The latter is directly used in the file-to-file case as the input TDE signal, as mentioned above.) The peak of this correlation provides an estimate of the location of the first transition. An estimate of the start of the input signal is then obtained, as we know the duration of the first low-energy segment.

Notice that the two-channel tape-recording and the two-channel A/D steps may, in general, introduce a non-unity gain between the coder input and output. Even if one tries to make the gain unity, this may be accomplished only approximately. Clearly, any gain value other than unity will have a degrading impact on the effectiveness of the signal-to-noise ratio measure STB-SNR. All the other objective measures we use, fortunately, are not affected by the absolute signal levels. To compensate for the non-unity gain, we added a provision to the TDE program to compute and print out the ratio of the energies of the first low-energy segment of the TDE input and output signals. We modified the objective measurement program so that it will accept this input-to-output energy ratio and include it in computing the STB-SNR measure.

## 7.5 Experimental Results

We used the TDE signal presented in Section 7.3 and the TDE algorithm described in Section 7.4 for estimating the time delay introduced by both file-to-file and real-time coders operating under several single-link and tandem conditions. The results are reported below.

### 7.5.1 File-To-File Coders

We considered a total of 13 coder conditions: CVSD in the quiet, and each of the other 4 coders (APC-NS, APC-SQ, RELP, and LPC-10) in the quiet, in 1% channel error, and in 5% channel error. Thus, we considered 7 waveform coder conditions, 3 RELP coder conditions, and 3 vocoder conditions. The time delay was estimated without error for all waveform coder conditions. We obtained the best delay estimate for the RELP coder conditions when we treated them as waveform coders. The resulting error in the estimated time delay was 1 sample, for each of the 3 RELP conditions. The delay estimates for the 3 LPC coder conditions were in error by 20 samples. From the results we reported in Section 7.2 on simulated time-delay errors, it is clear that the errors we just reported should not affect the performance of the objective measures.

### 7.5.2 Real-Time Coders

Below, we present the results we obtained for the five real-time coders in the quiet condition. We note that the "true" coder delay estimate obtained from the knowledge of the coder algorithm and implementation is only approximate since the analog lowpass filters used at the coder input and output introduce frequency-dependent group delays into the signal path. (If these filters were "linear-phase filters," we could theoretically determine the exact additional delays they would introduce.) In an attempt to resolve this problem, we displayed the digitized coder input and output signals on our IMLAC display computer, visually compared similar events on the input and output signals, and manually computed the delay estimate. For each of four coders, we obtained two such estimates by visual inspection of two different events of the signals; for the RELP coder, we could get only one reasonable visual time-delay estimate.

Table 34 lists, for each of the five real-time coders, three groups of time-delay estimates: 1) estimate produced by the TDE program, 2) two estimates obtained by visual inspection, and 3) estimate that is an integer number of frames of the coder in question and that is nearest the visual estimates. The samples in terms of which the time-delay estimates are given in the table have a period of 150 microseconds (6.67 kHz sampling rate). It is clear from the table that our TDE program produces satisfactory results. Again, we mention that we cannot give the "true" error in the time-delay estimate as we do not know exactly what the correct delay is (because of the input and output filters). We, however, know that the error is nonzero as we are using a fixed sampling rate of 6.67 kHz, which is different from the coder's internal sampling rate.

| Coder | Time-Delay Estimate | | |
|---|---|---|---|
| | TDE Program (Samples) | Visual Inspection (Samples) | Integer Coder Frames(Samples) |
| APC-NS | 2400 | 2369,2388 | 11 (2393) |
| CVSD | 807 | 799,830 | 2 (833) |
| APC-SQ | 1841 | 1828,1842 | 11 (1837) |
| RELP | 2360 | 2347 | 13 (2356) |
| LPC-10 | 2040 | 2064,2034 | 13 (1950) |

TABLE 34. Estimated time delays for the five real-time coders.

## 8. OBJECTIVE QUALITY EVALUATION OF REAL-TIME CODERS

In this chapter, we report objective quality evaluation results we obtained from real-time speech coders, using the input-speech database we previously designed (Section 6.3) and the time delay estimation algorithm we described above in Chapter 7. To yield the best prediction of the subjective, DAM test scores, the composite objective measures must be optimized over the real-time coder conditions. However, we initially investigated the use of the file-to-file coder objective measures on real-time coders. The results of this investigation, reported below in Section 8.1, not only indicate, as expected, suboptimal prediction of the DAM scores, but also help us identify and correct several factors that cause the performance of the objective measures to degrade. The topic of optimization of objective measures over real-time coder conditions is treated in Section 8.2. In Section 8.3, we present the results obtained from performance comparisons of objective measures over training and test sets of real-time coder conditions. Finally, in Section 8.4, we make several recommendations for objective speech quality evaluation of real-time coders.

We note that the Appendix contains a detailed description of 1) the coder input tape containing the TDE signal and the 12 DAM test sentences; 2) the procedure to check and set various amplifier gains and to prepare the two-channel tape containing the coder input and output; and 3) the procedure to digitize the two-channel tape using a two-channel A/D converter. In the coder input tape, the two DAM test sentences from each of the 6 speakers has been recorded as one unit or block, which enables the user to digitize the coder input and output as 6 separate speaker files. Another point we note here is that the coder input tape has two repetitions of the 3 blocks for male speakers, followed by two repetitions of the 3 blocks for female speakers. Only the second set of blocks in each case will be digitized. This arrangement allows the real-time coder algorithm (e.g., the pitch and voicing extraction algorithm used by LPC-10) to adapt to the speakers during the first set of blocks.

## 8.1 Use of File-To-File Coder Measures on Real-Time Coders

In this investigation, for the most part we considered the real-time APC-NS coder operating in the quiet condition and the FVSD measure that was optimized over the 12 file-to-file waveform coder conditions (see Section 6.4). Below, we report the results of several experiments (Section 8.1.1), present the idea of excluding silences from speech data used for objective quality measurement (Section 8.1.2), and discuss the issue of "repeatability' of the objective measures, which is relevant to the real-time coder case as the two-channel tape recording, cuing of the analog tape, and digitization are all prone to variabilities from one session to another (Section 8.1.3).

### 8.1.1 Experimental Results

For the APC-NS coder in the quiet condition, we obtained a time delay estimate of 2396 samples (359.4 ms) and an input-to-output energy ratio of 0.457, which, we recall from Section 7.4, is the gain change between the digitized coder input and the digitized coder output and is used in computing the signal-to-noise ratio measure STB-SNR. The FVSD and OCM measures each produced a value that was about 9 points lower than the value we obtained from the file-to-file APC-NS coder in the quiet; the corresponding reduction for STB-SNR was nearly 22 points as this measure is quite sensitive to phase changes introduced by D/A and A/D filters, A/D channel noise, etc. For subsequent detailed investigation, we considered the FVSD measure evaluated over one male speaker.

First, we found that one of the two channels used in two-channel A/D had a lowpass filter to eliminate high-frequency noise picked up by the connecting cable and the other channel did not. We installed the lowpass filter on the second channel and repeated the experiment. The FVSD measure changed only slightly from 53.4 to 54.0, but the value for the earlier file-to-file coder case was 63.4.

Second, we compared the individual band contributions to the FVSD measure for the real-time and file-to-file cases. The major differences were seen for Band 1 (0-400 Hz) and Band 5 (1900-2600 Hz). This motivated us to examine these frequency regions in detail. We found that the input speech used for the objective measure computation had been highpass filtered at 100 Hz for the file-to-file case but not for the real-time case. We then highpass filtered the input from the two-channel tape before digitizing, and the resulting FVSD measure increased to 56.9 as shown in Table 35. Another way of incorporating the highpass filter effect is to redefine Band 1 as 100-450 Hz; of course, this requires retraining or reoptimizing the FVSD measure. We decided to redefine Band 1 before we optimized the composite measures over the real-time coder conditions.

| Item | Input | Output | FVSD |
|---|---|---|---|
| 1 | (old file to file coder) | | 63.4 |
| 2 | No HPF | No HPF | 54.0 |
| 3 | HPF | No HPF | 56.9 |
| 4 | HPF | VAX D/A, No HPF | 62.4 |
| 5 | HPF | VAX D/A, HPF | 63.4 |
| 6 | (New File-to-File Coder) | | 65.7 |

TABLE 35. FVSD scores for the APC-NS coder, under six different ways of obtaining the coder input and output data for objective speech quality measurement. (HPF = Highpass filtering at 100 Hz.)

Third, we note that the output speech used in computing the objective measures for the real-time coder case involves the additional steps of D/A, 2 lowpass filters (one at the MAP output and another before A/D), and A/D; these steps are not performed in the file-to-file case. The D/A process introduces sin x/x distortion, and the lowpass filters introduce additional attenuation around the cutoff frequency; these two signal distortions will primarily affect high-frequency regions, which might have caused the problem in Band 5 noted above. However, as we listened to the speech at various points in the real-time coder and compared with the file-to-file coder output, we found that the real-time coder output sounded noticeably more noisy than the file-to-file coder output did. To examine this issue further, we applied the file-to-file coder output to the D/A converter on our VAX computer, and the analog output sounded significantly less noisy. Although the MAP D/A converter performed up to its specifications, it was apparently inferior to the VAX D/A. Using the VAX D/A, we obtained a value of 62.4 for FVSD (see Table 35).

Fourth, as shown in Table 35, when we used 100 Hz highpass filtering for the coder output, the FVSD score increased to 63.4, which is equal to the value for the old file-to-file coder case. Since the two-sentence speaker blocks, used for the real-time coder case, have more silence at the ends (to allow for easy cuing and digitizing), we re-evaluated the FVSD measure for the file-to-file case using the new input file. The result is shown in Table 35. Comparing items 5 and 6 in Table 35, we see that the difference between the FVSD values for the real-time and file-to-file cases has been reduced to 2.3. This difference is small and is the result of our using the FVSD measure trained on the file-to-file coder conditions.

From our systematic and detailed investigation reported above, we make the following important conclusions. First, the major difference between the file-to-file and the real-time cases is caused by the MAP D/A, which adds a perceivable amount of noise at the output of the real-time coder. We have shown that this noise is responsible for a substantial part of the observed large differences between the objective scores of the two cases. Second, since we used the real-time coder for the DAM testing of the APC-NS coder, we believe that the objective measures trained on the real-time coder conditions should perform substantially better than the ones trained on the file-to-file coder conditions. To the extent that our implementations of the LPC-10, APC-SQ, and CVSD coders on the MAP-300 (including differences in the D/A) may be different from those used in the

DAM testing performed by the Government, we should still expect some suboptimal performance of the measures trained over the MAP-300 real-time coder conditions. Third, 100-Hz highpass filtering of the input and output before digitizing is recommended. Finally, the two channel A/D will be effective only if the two channels have approximately the same characteristics.

### 8.1.2 Silence Exclusion

Since the two-sentence speaker blocks of input speech can have significant amounts of silence depending upon how the user cues the tape for digitizing, we implemented a method for excluding silence segments from speech data used for objective measure computation. In this method, the presence of silence is determined from the level of frame energy of the coder input signal. Recall from Section 4.1 that we compute coder input frame energy in the objective measure program for weighting the linear spectral distance measure. Frames for which this energy drops below a threshold are declared as silence and excluded from the objective measure frame averaging process. A single threshold is used for all 6 two-sentence utterances (one per speaker) of a given coder condition. The threshold is updated for each new coder condition to reflect variations that can occur in the level settings of the amplifiers used for recording and digitization. The variations are determined from a measure of the coder input signal energy that is computed in the time delay estimation program. Recall from Section 7.4 that the TDE program outputs an input-to-output energy ratio for the first low-energy region of the TDE signal. We modified the TDE program to print out the computed coder input signal energy in dB, E, of this region. The energy E is used in computing a threshold value T, for the specified coder condition, as T = E-T0. T0 is a constant that we have determined experimentally as 27 dB.

To test the effect of silence exclusion on objective measures, we computed the measures for the APC-NS coder under the quiet condition, with and without silence exclusion. Results showed a three-point increase in the FVSD composite measure when we excluded the silence regions.

### 8.1.3 Repeatability of Objective Measures

Since the steps involved in generating speech data for objective quality measurement of real-time coders, such as two-channel tape recording, cuing of the analog tape, and two-channel A/D, are all prone to variabilities from one session to another (or from one user to another), we were concerned about the repeatability of objective measures. To examine this issue, we conducted the complete objective evaluation (from start to finish, including taping, digitizing, etc.) of the APC-NS coder in the quiet condition twice; the two tests were separated by several days. The same person performed both tests. The FVSD measure values from the two tests are given in Table 36.

We note that the all-speaker FVSD score of the two tests differed by approximately one point, and that the differences in individual-speaker scores vary from a minimum of 0.2 points to a maximum of 2.4 points. As one interpretation of these results, we may say that differences in the objective measures of two coders, which are of the same magnitude as given above (1 point for the all-speaker case and 2 points for the individual-speaker case), are not significant.

| SPEAKER | TEST #1 | TEST #2 |
|---|---|---|
| 1 | 60.20 | 59.65 |
| 2 | 59.65 | 59.03 |
| 3 | 59.23 | 59.42 |
| 4 | 61.72 | 59.72 |
| 5 | 62.83 | 61.53 |
| 6 | 61.10 | 58.67 |
| All Speakers | 60.76 | 59.67 |

TABLE 36. FVSD scores for individual speakers and all speakers from two separate tests of the APC-NS coder.

## 8.2 Optimization of Objective Measures Over Real-Time Coder Conditions

We considered a total of 23 real-time coder conditions: 13 waveform coder conditions and 10 vocoder conditions. Of the 31 coder conditions listed in Table 5 (Chapter 5), we did not include any of the 8 RELP coder conditions since the RELP coder real-time implementation on our MAP did not operate properly (it failed to keep up in real-time). For the composite measures FVSD and STB-SNR, we redefined Band 1 as 150-400 Hz (see Section 8.1.1). Also, we excluded silence segments from speech data used for computing objective measures (see Section 8.1.2).

We performed the two-channel recording of the coder input and output signals for all 23 coder conditions, and digitized, using two-channel A/D, the coder input and output signals for the 6 two-sentence DAM utterances and the TDE signal for each coder condition. The TDE signals were processed by the TDE program to obtain the estimated delays, the input-to-output energy ratios for the STB-SNR measure, and the average input energy for silence detection. Parameter files and objective measure files were made for each condition.

Three coder classes were considered for optimization, waveform coders (13 conditions), vocoders (10 conditions), and all-coders (23 conditions, 13 waveform and 10 vocoder). We optimized the three composite measures FVSD, STB-SNR, and OCM for the waveform coder class, and the FVSD and OCM measures for the vocoder class and for the all-coder class. The correlations for the individual-speaker and all-speaker cases for the three coder classes are given in Table 37. From Table 37, we note that the correlation for the all-speaker OCM measure for each of the three coder classes is larger than our goal of 0.9.

| | Individual Speakers | | | All Speakers | | |
|---|---|---|---|---|---|---|
| Coder Class | STB-SNR | FVSD | OCM | STB-SNR | FVSD | OCM |
| Waveform | 0.776 | 0.921 | 0.926 | 0.843 | 0.970 | 0.970 |
| Vocoders | -- | 0.595 | 0.724 | -- | 0.749 | 0.921 |
| All Coders | -- | 0.884 | 0.898 | -- | 0.942 | 0.957 |

TABLE 37. Correlation coefficients for the individual-speaker and all-speaker cases for the STB-SNR, FVSD, and OCM composite measures for the waveform, vocoder, and all-coder classes for real-time coder conditions.

For the vocoder class, the all-speaker OCM correlation of 0.921 represents a substantial improvement over the correlation of 0.631 that was obtained for the same ten vocoder conditions in the file-to-file case. One possible reason for this improvement may be the use of silence exclusion in the real-time coder case. When silence segments are not excluded, the output silence segments will in general have noise for the tandem conditions involving LPC and a waveform coder; this could cause fairly large frame distance values for these segments (since the short-term spectra of the input and output signals could be rather different and since no vocoder measures except LSD are affected by frame energy). However, we have not confirmed the above reasoning by recomputing the file-to-file measure, with silence exclusion. Had we excluded the two tandem links that gave inconsistent DAM scores, the correlations for the vocoder class would be higher than reported in Table 37.

As expected, the STB-SNR measure (being sensitive to phase changes) produced lower individual-speaker and all-speaker correlations of 0.776 and 0.843 than those we obtained in the file-to-file case for a similar set of waveform coder conditions (0.889 and 0.963, respectively, for 12 waveform coders, see Table 18). Because of this result, we examined a waveform coder OCM measure that did not include the STB-SNR measure as one of the components. Optimization of an OCM measure that contained the FVSD and 4 simple measures (LAR, RFC, LHR, and LSD) resulted in an individual-speaker correlation of 0.923. (The all-speaker correlation stayed at 0.97.) This result indicates that the STB-SNR measure does not adversely affect the OCM measure and may provide some limited benefit.

For the all-coder class, Table 37 shows that the OCM measure is only slightly better than the FVSD measure. The relatively lower performance of the individual-speaker FVSD and OCM measures over the vocoder conditions was responsible for limiting the individual-speaker performance over the all-coder conditions. This supports our belief that further work is warranted in developing new and improved objective measures for vocoders (see Chapter 10).

## 8.3 Performance Comparisons Over Training and Test Sets

Following the procedure we used in the file-to-file case (Section 6.5), we conducted experiments to evaluate and compare the performance of objective measures over training and test sets, separately for the three classes of coders we considered. The results of these experiments are reported below. We hasten to point out that only the all-coder class has a reasonably large number of coder conditions to assign credibility to the evaluated training-set and test-set correlations; the correlation results reported for the other two cases must be interpreted with caution because of the rather limited data used.

### 8.3.1 Waveform Coders

We divided the 13 waveform coder conditions into a training set of 9 coder conditions and a test set of 4 coder conditions. The 4 coder conditions in the test set are: CVSD quiet, APC-SQ in 1% channel error, APCNS-APCNS tandem, and the CVSD self-tandem. Individual-speaker and all-speaker correlations over the training and test sets are given in Table 38. Surprisingly, the correlations of the STB-SNR measure degrade the least, albeit they are the smallest over both training and test sets. The change in correlation over training and test sets is substantial for both FVSD and OCM. FVSD seems to have the best overall performance.

| | Individual Speakers | | | All Speakers | | |
|---|---|---|---|---|---|---|
| | FVSD | STB-SNR | OCM | FVSD | STB-SNR | OCM |
| Training Set | 0.958 | 0.791 | 0.963 | 0.998 | 0.847 | 0.997 |
| Test Set | 0.763 | 0.711 | 0.743 | 0.928 | 0.797 | 0.883 |

TABLE 38. Correlations over training and test sets for the real-time waveform coder class.

### 8.3.2 Vocoders

We considered only 8 of the 10 vocoder conditions. The LPC10-APCSQ and the LPC-APCNS tandem conditions were not included because of their inconsistent DAM scores. The 8 vocoder conditions were divided into a training set of 6 coders and a test set of 2 coders; the latter had the two conditions: APCNS-LPC and CVSD-LPC. The correlation results are given in Table 39. The all-speaker correlations for the test set were not computed since only 2 items existed for correlation. The individual-speaker correlation decreases over the test-set only slightly for FVSD and modestly for OCM.

| | Individual Speakers | | All Speakers | |
|---|---|---|---|---|
| | FVSD | OCM | FVSD | OCM |
| Training Set | 0.650 | 0.773 | 0.817 | 0.939 |
| Test Set | 0.643 | 0.657 | --- | --- |

TABLE 39. Correlations over the training and test sets for the real-time vocoder class.

### 8.3.3 All Coders

We divided the 23 coder conditions into a training set of 17 coder conditions and a test set of 6 coder conditions. The test set contained the following conditions: APCNS-APCNS tandem, APCSQ quiet, LPC10-CVSD tandem, LPC10 in 5% error, CVSD-APCNS tandem, and APCSQ-LPC10 tandem. The correlation results are given in Table 40. We note that correlations for the individual-speaker and all-speaker FVSD measures and the all-speaker OCM measure increased slightly for the test-set data. Clearly, the correlation changes are rather small.

| | Individual Speakers | | All-Speakers | |
|---|---|---|---|---|
| | FVSD | OCM | FVSD | OCM |
| Training Set | 0.885 | 0.893 | 0.942 | 0.954 |
| Test Set | 0.896 | 0.886 | 0.962 | 0.965 |

TABLE 40. Correlations over the training and test sets for the real-time all-coder class.

## 8.4 Recommendations

From the results presented above in Sections 8.2 and 8.3, we recommend the use of the FVSD measure for waveform coders, as it has the best overall performance over training and test conditions, and the use of the OCM measure for vocoders and for all coders. It is possible that the performance of the STB-SNR measure, which is a generalization of the traditionally used waveform coder measure of signal-to-noise ratio, could be improved to a point it becomes more attractive than FVSD currently is, if some phase equalization of coder output speech (relative to input) were incorporated.

The results presented in Section 8.1 should convince the reader to pay attention to the implementation details of a real-time coder (such as filters, A/D, D/A, etc.). The same coding algorithm implemented on one set of hardware may yield better or worse speech quality than on another set of hardware. For example, we mention that the all-speaker FVSD measure of the APC-NS coder in the quiet condition was about 4 points lower for the implementation on the DCEC System A MAP than for the implementation on the BBN MAP. The above DCEC MAP had been known to produce additional noise at the output.

On the other hand, we are certainly not implying that the recommended objective measures are infallible. Only if the coder under test is similar to the coders included in the training set of the objective measure, should we expect the measure to give a good prediction of the coder speech quality. If the test coder is very different from the coders in the training set, a rather poor prediction of speech quality may result, as hinted by the results reported in Sections 6.5 and 8.3.

Finally, we reiterate the need to exercise care, caution, and attention to details in collecting the speech data from the real-time coder under test, as described in the Appendix.

## 9. SOFTWARE FOR OBJECTIVE QUALITY EVALUATION OF REAL-TIME SPEECH CODERS

The general Fortran simulation software package that we developed on our VAX-11/VMS computer contained an interactive command structure and numerous options to facilitate our research and development work on objective speech quality evaluation of real-time coders. From this general software package, we developed a simpler package that simulates our final choice of the objective speech quality measurement system. This final system has three separate programs: time-delay estimation program, objective measurement program, and correlation program. Given the digitized input and output TDE signal files, the TDE program gives the coder time-delay estimate, the input-to-output energy ratio, and the average input energy for the low-energy region. Given the six 2-sentence speaker input and output files for one coder condition, the objective measurement program computes and prints out the various objective measure values for the individual-speaker and all-speaker cases. Given a table of objective and subjective scores, the correlation program produces the regular and rank-order correlations.

In preparing the objective measurement program, we integrated the frame-by-frame measures program and the objective measures program (see Section 4.4) into one program. Of course, we removed code dealing with least-squares optimization and combinatorial analysis of composite measures. The task of extracting the final simulation system from the general development package involved a number of items including the following: removing code that was not relevant to the final system, taking out the interactive command structure, placing the values of coefficients and parameters in DATA statements, moving all COMMON statements to the MAIN program to facilitate an overlay structure, if required, for running on the sponsor's DCEC PDP-11/34, modifications to use sequential file I/O, reducing all array dimensions to their required values, eliminating buffers that were not part of the final system, substituting calls to our speech library with calls to routines included in the simulation package, Fortran changes to accommodate what is available on the DCEC PDP/11, and providing diagnostic printouts that may be turned on or off by the user. We thoroughly tested the final system software on our VAX and ensured that we got the same results as with the original programs.

We then modified the final system software to make it run on the DCEC PDP-11/34. The available address space on the PDP-11 is 32K 16-bit words. Of this space, approximately 12K words are used by the Fortran run-time library, leaving only about 20K words for user programs and data. We first modified the waveform file input-output portions of our VAX/VMS programs to use the DCEC QIOLIB input/output subroutine package. We then evaluated the routine-by-routine memory requirements for the the system programs, and determined an appropriate overlay structure for each program. The correlation program did not require the use of overlays to fit into the available address space; the other two programs did require overlays.

We installed the objective speech quality measurement software on the sponsor's PDP-11 at DCEC, tested each of the three programs using the real-time coder data generated at BBN, and compared the results with those we had obtained at BBN. The test results indicated correct installation of the software. Finally, we note that the user's guide [26] provides information necessary to install and use the objective speech quality measurement system software.

## 10. SUMMARY AND FUTURE RESEARCH

The work we performed in this research falls into two major areas: subjective testing of the real-time 16 kbit/s APC-NS coder and objective speech quality evaluation of real-time coders. We tested the speech intelligibility of the APC-NS coder using the Diagnostic Rhyme Test and the speech quality using the Diagnostic Acceptability Measure test, under eight operating conditions involving channel error, acoustic background noise, and tandem link with two other coders. The test results show that the DRT and DAM scores of the APC-NS coder equal or exceed the goal scores in all but one tandem condition, with the goal scores being the test scores of the 32 kbit/s CVSD coder. For that one tandem condition, we have provided evidence to suggest possible malfunction of the second coder (the 16 kbit/s CVSD) in the tandem link.

In the area of objective speech quality evaluation, we developed, tested, and validated a procedure for automatically computing several objective speech quality measures, given only the tape-recordings of the input speech and the corresponding output speech of a real-time speech coder. As test bed, we used five real-time speech coders, each operating under several single-link and tandem conditions; in all, we considered 31 coder conditions for which the subjective six-speaker DAM test scores were available to us. A summary of our work in this area is given below.

- We developed a procedure for automatically synchronizing in time the coder output speech with the coder input speech. The procedure involves the use of a specially designed input signal that facilitates the estimation of the time delay introduced by the coder.

- We designed an input-speech database of 12 sentences (2 per speaker) for objective speech quality measurements, by selecting a subset of the 72 sentences used in the DAM test; the selection was accomplished through a multidimensional scaling procedure.

- For a given coder condition, we computed objective measures by comparing the synchronized coder output speech with the coder input, over the foregoing 12-sentence database. We divided the coder conditions into three classes: waveform coders, vocoders, and all coders, and we developed and optimized objective measures for each of the three classes. We computed the correlation of the objective measures with the subjective DAM scores in two ways: over individual speaker scores and over scores averaged over all six speakers.

- The individual-speaker and all-speaker correlations of the best objective measure for the real-time coders were, respectively, 0.93 and 0.97 for waveform coders, 0.72 and 0.92 for vocoders, and 0.9 and 0.96 for all coders. The all-speaker correlation exceeded our goal of 0.9 in each case.

- We evaluated the objective measures over a test set of coder conditions, which was different from the set used for training or optimizing the measures. The changes in correlations over the training and test sets were small when both sets were sufficiently large, which was true for the all-coder class. For each of the other two classes, we had only a limited database of coder conditions, and the resulting modest-to-large correlation changes are therefore not reliable. We must point out that if the test set is quite different from the training set, the test-set correlations could be drastically lower than the training-set correlations, even with large databases of coder conditions. We consider a test coder to be "quite different" from the coders in the training set, if it produces dstortions in the output speech that are not covered by the training-set coders. As an example, we point out the case where the training set has coders all operating over noise-free channels and the test set has coders all operating over fairly noisy channels.

Finally, we suggest four problems for future work. First, how can the objective measures for vocoders be improved? Three specific methods seem to hold promise: average frame-by-frame distances over only the top n-th percentile values (e.g., n=20, see Section 6.2.2.2) [15]; use a variable frame rate model for computing the parametric measures (see Section 3.1) [15]; incorporate pitch errors into objective measures.

Second, how can the signal-to-noise ratio measure STB-SNR for real-time coders be improved? We believe that with appropriate phase equalization (using an all-pass network) of the coder output relative to the coder input, the STB-SNR measure should produce excellent correlation with subjective judgments.

The third problem is based on our experience that the objective measures presented in this report do not predict well the speech quality of tandem links, particularly when each coder in the tandem produces different types of distortions in the output speech. In the approach we pursued in this work, we compared the overall tandem output with the overall tandem input; in so doing, we lumped the distortions produced by both coders. An alternate approach is to evaluate the first coder by comparing its output with its input; evaluate the second coder by comparing its output with its input, which is the first coder's output; and then somehow (linearly or nonlinearly) combine the two objective scores to produce one that predicts the quality of the tandem link.

The fourth problem is somewhat related to the third one presented above: how to predict the speech quality of a coder operating in acoustic background noise (e.g., ABCP noise, helicopter noise, ship noise, etc.)? The coder input speech in this case is the noise-corrupted speech. The approach of comparing the coder output with its input will not work well in this case. For example, we found that the APC-NS coder produced a higher SNR with ABCP noise-corrupted speech as input than with clean speech as input [1], although the coder produced inferior speech in ABCP noise than in the noise-free case. We suggest an alternate approach in which one evaluates the objective measure of the coder with clean speech as input and combines this somehow (linearly or nonlinearly) with the DAM test score of the ABCP noise-corrupted speech, to obtain the desired score. (The government has available the DAM scores for speech in a number of different noise environments.)

## 11. REFERENCES

1. V. Viswanathan, W. Russell, and A. Higgins, "Design and Real-Time Implementation of a Robust APC Coder for Speech Transmission over 16 kb/s Noisy Channels," Final Report, Vol. I: Algorithm Design and Simulation, BBN Report No. 4565, Bolt Beranek and Newman Inc., December 1980, AD No. A096092, Contract No. DCA100-79-C-0037.

2. J. Wolf, K. Field, and W. Russell, "Design and Real-Time Implementation of a Robust APC Coder for Speech Transmission Over 16 kb/s Noisy Channels," Final Report, Vol. II: Real Time Implementation, BBN Report No. 4565, Bolt Beranek and Newman Inc., December 1980, AD No. A096092, Contract No. DCA100-79-C-0037.

3. V. Viswanathan, W. Russell, A. Higgins, M. Berouti, and J.Makhoul, "Speech-Quality Optimization of 16 kb/s Adaptive Predictive Coders," IEEE International Conference on Acoustics, Speech and Signal Processing, Denver, CO, April 1980, pp. 520-525, Vol. 2.

4. V. Viswanathan, W. Russell, and A. Higgins, "Noisy Channel Performance of 16 kb/s APC Coders," IEEE International Conference on Acoustics, Speech and Signal Processing, Atlanta, GA, April 1981, pp. 615-618, Vol. 2.

5. W.D. Voiers, A.D. Sharpley, and C.J. Hehmsoth, "Research on Diagnostic Evaluation of Speech Intelligibility," Tech. Report AFCR-72-0694, TRACOR Inc., January 1973.

6. W.D. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems," IEEE International Conference on Acoustics, Speech and Signal Processing, Hartford, CT, May 1977, pp. 204-207.

7. T.E. Tremain, J.W. Fussell, R.A. Dean, B.M. Abzug, M.D. Cowing, and P.W. Boudra, Jr., "Implementation of Two Real-Time Narrowband Speech Algorithms," Proc. EASCON '78, Washington, DC, September 1978, pp. 698-708.

8. J. Wolf, K. Field, and W. Russell, "Real-Time Implementation of the APC-SQ and LPC-10 Speech coding Algorithms," Final Report, BBN Report No. 4855, Bolt Beranek and Newman Inc., June 1982, Contract No. DCA100-80-C-0034, Ad No. A116902.

9. V. Viswanathan, J. Wolf, L. Cosell, K. Field, A. Higgins, and W. Russell, "Design and Real-Time Implementation of a Baseband LPC Coder for Speech Transmission Over 9600 Bps Noisy Channels," Final Report BBN Report No. 4327, Bolt Beranek and Newman Inc., February 1980, Vol. I & II, AD No. A083079 and A083238.

10. V. Viswanathan, A. Higgins, and W. Russell, "Design of a Robust Baseband LPC Coder for Speech Transmission Over 9.6 Kbit/s Noisy Channels," IEEE Trans. Communications, Vol. COM-30, No. 4, April 1982, pp. 663-673.

11. T.P. Barnwell III and W.D. Voiers, "An Analysis of Objective Measures for User Acceptance of Voice Communication," Final Report, Contract DCA100-78-C-0003, Georgia Institute of Technology, September 1979, AD-A089210.

12. T.P. Barnwell, III, "Correlation Analysis of Subjective and Objective Measures for Speech Quality," IEEE International Conference on Acoustics, Speech and Signal Processing, Denver, CO, April 1980, pp. 706-709, Vol. 3.

13. J. Makhoul, V. Viswanathan, and W. Russell, "A Framework for the Objective Evaluation of Vocoder Speech Quality," IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, PA, April 1976, pp. 103-106.

14. V. Viswanathan, J. Makhoul, and W. Russell, "Towards Perceptually Consistent Measures of Spectral Distance," IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, PA, April 1976, pp. 485-488.

15. V. Viswanathan, W. Russell, and J. Makhoul, "Objective Speech Quality Evaluation of Narrowband LPC Vocoders," IEEE International Conference on Acoustics, Speech and Signal Processing, Tulsa, OK, April 1978, pp. 591-594.

16. B.J. McDermott, C. Scagliola, and D. Goodman, "Perceptual and Objective Evaluation of Speech Processed by Adaptive Differential PCM," Bell System Technical Journal, Vol. 57, May-June 1978, pp. 1597-1618.

17. J.M. Tribolet, P. Noll, B.J. McDermott, and R.E. Crochiere, "A Study of Complexity and Quality of Speech Waveform Coders," IEEE International Conference on Acoustics, Speech and Signal Processing, Tulsa, OK, April 1978, pp. 586-590.

18. P. Mermelstein, "Evaluation of a Segmental SNR Measure as an Indicator of the Quality of ADPCM Coded Speech," J. Acoust. Soc. Am., Vol. 66, No. 6, December 1979, pp. 1664-1667.

19. V. Viswanathan, J. Makhoul, R. Schwartz, A.W.F. Huggins, "Variable-Frame-Rate Transmission: A Review of the Methodology and Application to Narrowband LPC Speech Coding," IEEE Trans. Communications, Vol. COM-30, No. 4, April 1982, pp. 674-686.

20. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 1, February 1975, pp. 67-72.

21. D.P. Kemp and J.W. Fussell, "Evaluation of Two Narrowband Speech Algorithms," IEEE International Conference on Acoustics, Speech and Signal Processing, Washington, DC, April 1979, pp. 990-993.

22. L. Kanal, "Patterns in Pattern Recognition: 1968-1974," IEEE Trans. Information Theory, Vol. IT-20, No. 6, November 1974, pp. 697-722.

23. A.W.F. Huggins, R. Viswanathan, and J. Makhoul, "Speech-Quality Testing of Some Variable-Frame-Rate (VFR) Linear-Predictive (LPC) Vocoders," J. Acoust. Soc. Am., Vol. 62, No. 2, August 1977, pp. 430-434.

24. Y. Takane, F.W. Young, and J. de Leeuw, "Non-parametric Individual Differences Multi-Dimensional Scaing: An Alternating Least-Squares Method with Optimum Scaling Features," Psychometrika, Vol. 42, 1977, pp. 7-67.

25. Special Issue on, "Time Delay Estimation," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-29, No. 3, Part II, June 1981, pp. .

26. W. Russell, K. Field, and V. Viswanathan, "User's Guide for the BBN Objective Speech Quality Measurement System," Bolt Beranek and Newman Inc., Tech. Report 5371R , January 1984, Prepared for the Defense Communications Agency.

# APPENDIX A
# DATA MEASUREMENT AND DIGITIZATION FOR QUALITY EVALUATION OF REAL-TIME SPEECH CODERS

In this appendix we describe the method for generating the digital input data required by the software for measurement of objective speech quality of real-time coders. The two audio tapes containing the analog input signal to be processed on a real-time coder have been delivered to the sponsor. The coder input and output signals are recorded on a two-channel tape recorder and then digitized using a two-channel A/D converter. The resulting files of digitized data provide the required input data for the software mentioned above. Below, we describe the contents of the analog tapes and discuss the steps required for two-channel tape-recording and A/D. We note that the material given below has been taken from the User's Guide [26].

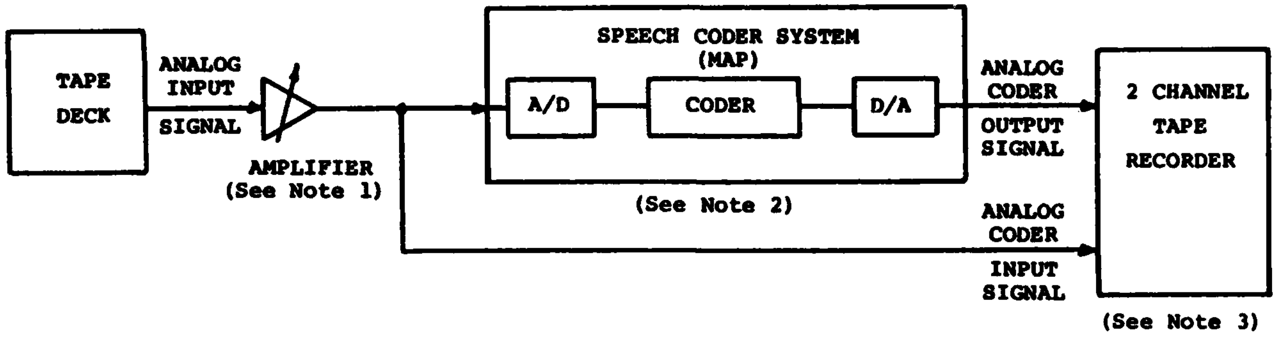
## A.1 Audio Tapes of Input Signal

Of the two audio tapes provided, one is used as input for waveform coders and the other as input to vocoders and all coders (see Section 7.3). Each tape contains an initialization signal (a five-second steady random noise signal with a mean-square value of 60 dB), followed immediately by the time-delay estimation (TDE) signal (a five-second random noise signal containing alternating regions of low and high energy). Following the TDE signal are 4 consecutive sets of DAM test sentences. Each set contains 6 utterances recorded in 3 blocks with each block containing two sentences from a single speaker.

The first set contains utterances spoken by 3 male speakers. The second set is a repeat of the first. The third set contains utterances spoken by 3 female speakers and the last set is a repeat of the third. Duplication of the DAM test sentences, as described above, has been provided so that a coder can adapt, if necessary, to speaker characteristics during the input of the first and third sets of utterances (utterances 1 to 6 and 13 to 18 on the tape). A two-channel recording should be made of all utterances on the tape. However, only the second and fourth sets of utterances (utterances 7 to 12 and 19 to 24 on the tape) and the TDE signal should be digitized and processed by the objective measures programs. A gap has been provided between blocks so that each block can be easily cued and digitized as a single speaker file. The initialization and TDE signals were recorded at roughly the same level as the DAM sentences so that all the analog signals (initialization signal, TDE signal, and the DAM sentences) can be processed by the real-time coder under evaluation, using the same amplifier settings.

### A.2 Two-Channel Tape-Recording

A block diagram of the setup for two-channel tape-recording is given in Fig. 7. Before an actual recording is made the signal levels must be adjusted at different places in the system. The correct settings of signal levels are obtained as follows. First, adjust the coder input signal level, using the input amplifier (see Fig. 7), so that it is in the normal operating range of the coder A/D. It is advisable that sufficient amplification of the input signal is provided so that the DAM test sentence from the tape with the strongest signal level fills



NOTES:

1. Adjust amplifier for an input signal level in the operating range of the speech coder A/D. For the MAP (12 bits A/D) the peak signal must be below 2047 - normal range is 1700 to 1900.

2. Speech coding system must have an overall gain of one. If losses occur in the speech coding system, the "Analog Coder Output Signal" must be amplified to compensate for the loss before 2-channel tape recording. This amplifier is not shown in the figure.

3. If the tape recorder has two separate amplifiers, one for each channel (or if one uses two external amplifiers), they must both be set to the <u>same</u> level. The reason for using amplification before taping is to get a reasonable recording level (-3 to -8 dB on the VU meter.)

4. The settings of the various amplifiers should be left unchanged during the entire recording session.

Figure 7. Block diagram showing the setup for two-channel tape-recording of coder input and output.

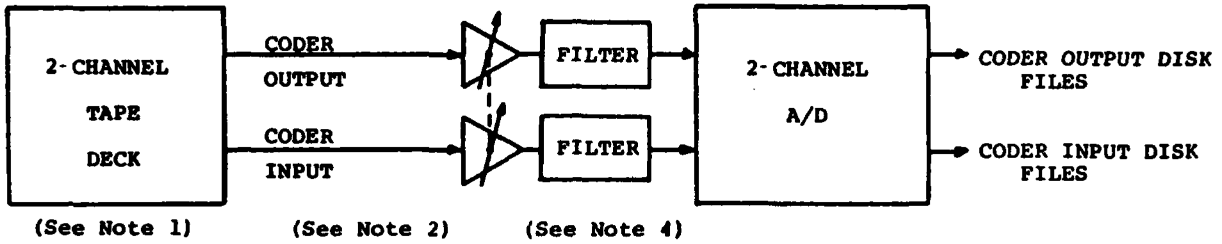
the full operating range of the coder A/D without causing overload. (For the coders BBN has implemented on the MAP, the peak signal level can be monitored, and the normal range is 1750 to 1950. The DAM test sentence with the strcngest signal level should give a peak signal level reading between 1900 and 1950.) Second, adjust any additional amplifier at the coder output (used to maintain an overall coder system gain of unity) so that the coder output signal level is equal to the coder input signal level. The coder input and output signal voltages should be monitored on a multimeter or oscilloscope to insure the correct coder output amplifier setting. Do not change the input amplifier level during this adjustment. Third, adjust the settings of the tape recorder's two input amplifiers. To make this adjustment, the coder input signal is applied to channel one of the tape recorder, and its amplifier is adjusted so that the peak signal level, as indicated on the VU meter is between -5 and -8 dB. Next, a 2-kHz sine wave from a signal generator is applied to both channels of the tape recorder, and the recorder amplifier of channel two is adjusted to provide equal output voltages from both channels. Finally, with these settings, play the entire tape through the system. Fine tune the appropriate amplifiers should the coder A/D or the VU meters of the tape recorder indicate that "overload" occurred at any time. Repeat this last step of fine tuning until no "overload" occurs.

After the correct signal levels have been set, cue the input tape for recording. Most real-time coders introduce a delay that will cause the coder output signal to lag the input signal. Because of the delay, the user must aurally monitor the coder input and output signals to cue properly for recording. Recording should begin just prior to the initialization signal; aurally monitor the coder input signal to locate this position on the tape. The coder output signal is monitored to determine when the last sentence of the last speaker has ended and the recording can be stopped. Record all signals (initialization, TDE, and the 24 DAM sentences) in a single session without changing any of the amplifier settings.

### A.3 Two-Channel A/D

A block diagram of the setup for two-channel A/D is shown in Fig. 8. Although not indicated in this figure, lowpass filters with cutoff at 3.2 kHz must be placed in the two channels (labelled as coder output and input in the figure) at the input of the A/D converter. Using the two-channel tape recorded from a real-time coder, adjust the signal levels as specified in the notes in Fig. 8. The two-channel A/D shown in the figure provides as output two sets of digitized signal files, one set for each channel. A typical A/D program (e.g., the one we have used) may produce one file containing samples from both channels in an interleaved fashion. A separate program is then required to split the interleaved data file into two individual files, one for each channel. All data must be sampled at 6.67 kHz (sampling period = 150 microseconds).

From a two-channel tape corresponding to a given coder condition, fourteen digital files are created: a pair of TDE files (coder input and coder output) and six pairs of speaker files, 3 pairs of files (3 male speakers) from the second set of utterances on the tape and 3 pairs of files (3 female speakers) from the fourth set of utterances on the tape. Recall that the first and third set of utterances (utterances 1 to 6 and 13 to 18



NOTES:

1. If the 2-channel tape deck has separate amplifiers for the two output channels, they must both be set to the same level.

2. Adjust amplifiers (if they are necessary) for a signal level in the operating range of the two channel A/D. Both amplifiers must be set to the same level. If MAP A/D (12-bit) is used, the peak signal must be below 2047 - normal range is 1700 to 1900.

3. Our plan is to digitize separately the synchronization (or time delay estimation) signal and each of the six speaker files. The synch signal is about 7 sec long (per channel), and each speaker file is about 10 sec long (per channel).

4. The two filters are identical 3.2 kHz lowpass filters. The filters we use have 3 dB cutoff at 3.2 kHz and -40 dB attenuation at 3.4 kHz (TT Electronics, Los Angeles; Model No. J576C5707).

5. We use a sampling rate of 6.67 kHz (T = 150 microseconds) on our VAX. If one uses the MAP A/D, we recommend a sampling rate of 6.621 kHz ( = 384/58).

Figure 8. Block diagram showing the setup for two-channel A/D.

on the tape) should not be digitized. Before digitizing, aurally monitor the input and the output signals to determine the tape counter positions where digitization should start and stop. Monitor the input signal to start digitizing and the output signal to stop digitizing. For correct operation of the TDE algorithm, digitization of the TDE signal must begin not more than one second before the beginning of the TDE signal. For each of the six pairs of speaker files, aurally monitor the coder input signal to locate where the first sentence in each speaker block begins. Start digitizing just prior to the beginning of this sentence. Stop digitization just after the end of the second sentence of each speaker block, which is located by aurally monitoring the coder output signal. Digitize all seven pairs of files for a coder condition, without changing any of the amplifier settings.

END

FILMED

6-84

DTIC