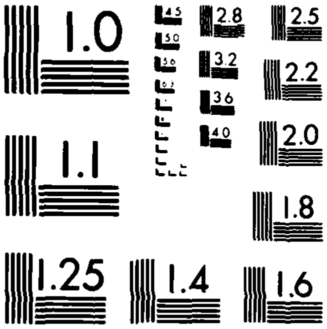
AD-A151 165 AUTOMATIC SPEAKER RECOGNITION OVER MILITARY COMMUNICATION SYSTEMS A FEASIBILITY EVALUATION(U) NAVAL RESEARCH LAB WASHINGTON DC S S EVERETT 05 MAR 85 1/1

UNCLASSIFIED NRL-MR-5519 F/G 17/2.1 NL


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NRL Memorandum Report 5519	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Naval Research Laboratory 6b OFFICE SYMBOL *(If applicable)* Code 7526	7a NAME OF MONITORING ORGANIZATION
6c ADDRESS *(City, State, and ZIP Code)* Washington, DC 20375-5000	7b ADDRESS *(City, State, and ZIP Code)*
8a NAME OF FUNDING / SPONSORING ORGANIZATION Office of Naval Research 8b OFFICE SYMBOL *(If applicable)*	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS *(City, State, and ZIP Code)* Arlington, VA 22217	10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
61153N		RR021-05-42	DN480-556

11 TITLE *(Include Security Classification)*

Automatic Speaker Recognition Over Military Communication Systems: A Feasibility Evaluation

12 PERSONAL AUTHOR(S)

Everett, S.S.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT *(Year, Month, Day)*	15 PAGE COUNT
Interim	FROM ________ TO ________	1985 March 5	16

16 SUPPLEMENTARY NOTATION

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

Voice processing — Communication security
Speaker recognition — Speech processing
Low bit rate voice processing

19 ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Automatic speaker recognition (ASR) offers potential benefit for numerous Navy situations, including identification of users of communication channels such as the telephone and channels using processed or vocoded speech. Currently the user must subjectively determine whether the person on the other end of the line is who he or she claims to be. However, past research has shown that ASR systems are capable of higher recognition accuracy than human listeners under certain circumstances. This report discusses a series of tests conducted to evaluate the feasibility of performing ASR using vocoded speech. The analog outputs of six different Department of Defense voice processors were used as input to a real-time ASR system. Data transmission rates of these processors ranged from 2400 to 64,000 bits per second. Recognition accuracy results for the processed speech were 70 to 95% using a 2500 Hz bandwidth input filter, and 75 to 95% using a 4000 Hz input filter. These results indicate that ASR using vocoded speech is definitely possible, though further research is needed to determine which speech parameters are best suited for use with each voice processor.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED	
22a NAME OF RESPONSIBLE INDIVIDUAL Stephanie S. Everett	22b TELEPHONE *(Include Area Code)* (202) 767-2116	22c OFFICE SYMBOL Code 7526

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

CONTENTS

AUTOMATIC SPEAKER RECOGNITION OVER MILITARY COMMUNICATION SYSTEMS: A FEASIBILITY EVALUATION

I. INTRODUCTION

Over the last 20 years, the research community has spent a great deal of time and money developing methods of recognizing speakers automatically based on voice input alone. Such systems could be of considerable benefit in numerous situations in both civilian and military environments. Potential Navy applications include controlling access to restricted areas, communication security, and verification of computer users through terminals accepting voice input.

One logical application of automatic speaker recognition (ASR) to communication security is the task of verifying the identity of speakers over the telephone or over military communication channels using processed or vocoded speech. Current verification methods are virtually nonexistent. The listener must subjectively determine or verify the speaker's identity based solely on his or her recollection of that person's voice. If the listener has never spoken with the person over this particular type of channel, or has never met him or her before, it can be nearly impossible to know for sure whether the speaker is who he or she claims to be. An ASR system that could automatically verify the speaker's identity or provide a confirmation and confidence rating to the listener would greatly improve communication security.

Though some research has been done on ASR using telephone (or telephone-quality) speech, very little is known about ASR using processed or vocoded speech. Military voice processing systems have been carefully designed to maximize the quality and intelligibility of the synthesized speech. However, the analyses performed to allow bandwidth compression or encoding of the speech signal frequently remove or distort certain characteristics of the

Manuscript approved December 13, 1984.

original speech. It is not known what effects, if any, this processing has on the portions of the speech signal relevant to speaker identity. A recent study showed that over a 2400 bits per second (bps) linear predictive coding (LPC) voice channel pecple could identify the familiar voices of their coworkers only about 70% of the time [1]. Other studies have shown that automatic recognition systems actually perform better than human listeners under certain conditions [2, 3].

As a preliminary evaluation of the potential for performing ASR over military communication channels, a series of tests was conducted using the speech output from several DoD voice processors as input to an existing ASR system.

II. ASR SYSTEM DESCRIPTION

The ASR system used in this feasibility evaluation was developed by ITT Defense Communications Division. This system was selected to meet the following requirements:

- perform text-independent speaker identification,
- operate in real time,
- recognition procedure completely automatic (i.e., no hand-marking of speech segments, determination of silent periods, etc.),
- capable of providing recognition results using less than 3 seconds of input speech,
- capable of operating on a set of at least 20 speakers, including both males and females, and
- have a reported overall recognition accuracy of at least 90% using high-quality input speech for a set of not less than 10 speakers.

This system uses a multiple parameter algorithm based on an LPC analysis of the input speech. This approach involves calculating the mean, variance, and covariance for all the frames in the training utterance. The mean vector for the reflection coefficients is then calculated over the test utterance and compared to the mean vectors in the model using the Mahalanobis distance measure. A more complete description of this system may be found in Reference 4.

Text-independent speaker recognition was required for this series of tests because it was felt that this would give a better indication of how well the voice processors reproduce all sounds for a variety of voices, rather than only those sounds in each individual's code word or phrase as with text-dependent recognition. The task chosen was speaker identification, where the ASR system must choose the speaker's identity from a set of known voices. This is somewhat more difficult than speaker verification, where the ASR system is given an identity claim and need only accept or reject that claim based on a preset distance measure threshold.

III. TEST DESCRIPTION

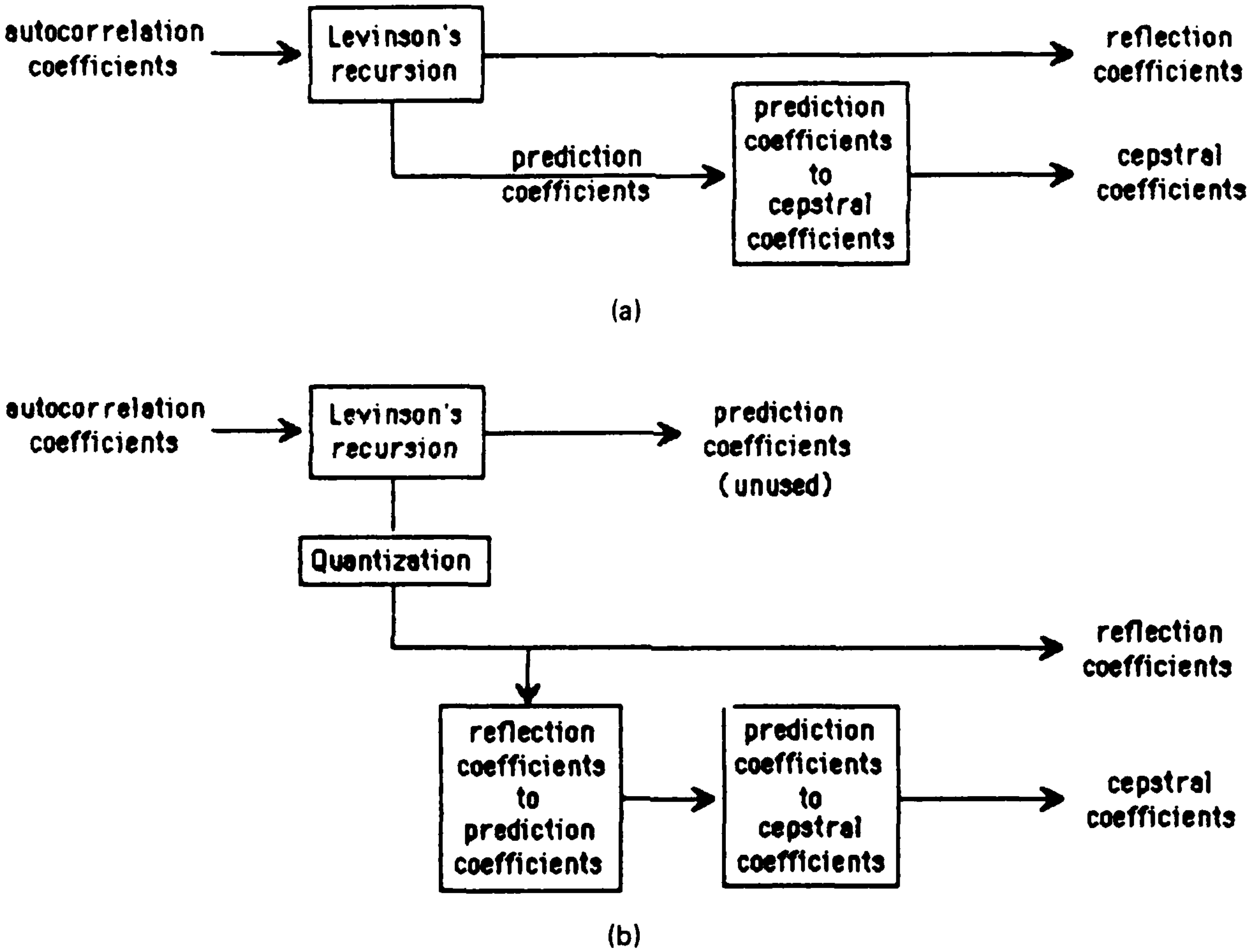
To test the performance of the ASR system an audio source tape was generated containing five phonetically balanced sentences from each of 20 speakers (10 males, 10 females) for a total of 100 different utterances. For each test condition the first two sentences from a given speaker were used for model generation and the remaining three sentences were used as test utterances.

The tape was played through each of the six different Department of Defense (DoD) voice processors listed in Table 1. The processed output speech was recorded for use as input to the ASR system. The ASR system was not specifically tailored to any of the voice processors used in this investigation. For complete descriptions of the voice processing systems used in this study, consult References 5-7.

Table 1. Coding methods and transmission rates of voice processing systems used for evaluating the feasibility of performing ASR over military communication channels.

Coding Method	Data rate (bps)
Pulse Code Modulation (PCM)	64000
Residual-Excited Linear Prediction (RELP)	16,000 and 9,600
Continuously Variable Slope Delta (CVSD)	16000
Adaptive Predictive Coding (APC)	9600
Linear Predictive Coding (LPC)	2400

One other test condition was designed to simulate the performance of speaker recognition prior to resynthesis at the receiver of a 2400 bps LPC voice processing system (assuming an ideal transmission channel). This test used the clear text source tape and required a slight modification of the ASR algorithm to include quantization of the LPC parameters. The ASR algorithm uses 20 acoustic features derived from LPC analysis: 10 reflection coefficients and 10 cepstral coefficients. In the standard algorithm implementation the LPC prediction and reflection coefficients are derived from the autocorrelation coefficients using Levinson's recursion [8]. The prediction coefficients are then converted to cepstral coefficients as shown in Figure 1(a). For the test using quantized LPC parameters the ASR system was modified as shown in Figure 1(b). The reflection coefficients were quantized, then converted first to prediction coefficients and finally to cepstral coefficients. The reflection coefficients were quantized using the standard DoD quantization. The modified ASR algorithm was used in both training and recognition for this one test condition.



autocorrelation coefficients
Levinson's recursion
reflection coefficients
prediction coefficients
prediction coefficients to cepstral coefficients
cepstral coefficients
autocorrelation coefficients
Levinson's recursion
prediction coefficients (unused)
Quantization
reflection coefficients
reflection coefficients to prediction coefficients
prediction coefficients to cepstral coefficients
cepstral coefficients

(a)

(b)

Figure 1. ASR signal processing procedure. Figure 1(a) shows the usual ASR algorithm; 1(b) shows the modifications used for the quantized LPC parameters test condition [9].

A block diagram of the testing configuration is shown in Fig. 2. The audio signal from the tape recorder was passed through an input filter and into a 12 bit analog-to-digital (A/D) converter. The signal was sampled at 10,000 samples per second. For each test condition the two or three sentences of each speaker's training or test set were grouped together and digitized as a single file, giving 40 such files for each condition. Excess silence at the beginning and end of the files was removed to reduce the amount of memory required to store the data. This was done using a simple energy-thresholding algorithm, leaving approximately one half second of silence as padding outside the detected endpoints to ensure that no speech sounds were removed. In order to use the entire 12 bit range of the digital representation the samples in each file were multiplied by a factor calculated to clip 1% of the samples.

Two series of tests were conducted, using input filters of 2500 Hz and 4000 Hz bandwidth, respectively. The results for both sets of tests are discussed in the next section.

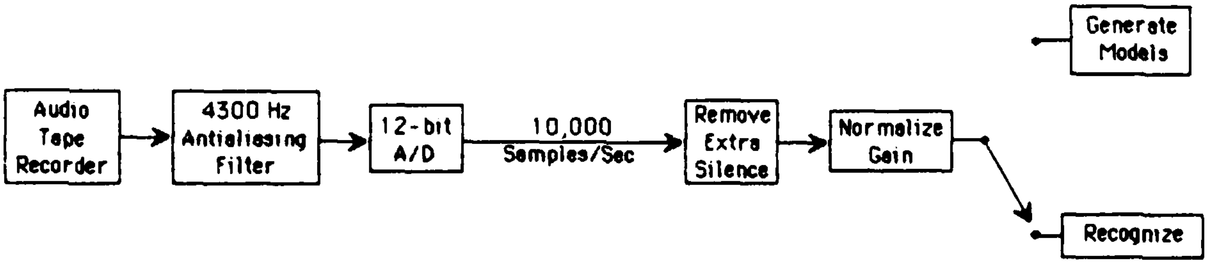


Figure 2. Block diagram of ASR system performance testing configuration.

IV. TEST RESULTS

Tables 2 and 3 summarize the results of the ASR system performance tests described in the preceding section. The results are based upon all of the available speech data. Tests were performed only within conditions, i.e., recognition tests were always conducted using the same input speech condition as had been used in the generation of the speaker models.

With the 2500 Hz bandwidth input filter recognition accuracy for the clear speech was 90%. Accuracy for the processed speech ranged from 70% to 95%.

LPC parameter quantization alone lowered the accuracy to 80%. It is interesting that the synthesized speech from the two residual-excited linear predictive (RELP) systems gave better results than the original clear speech, however the difference is not significant. It is conceivable that the analysis and synthesis procedures used in these processors involve a quantization or smoothing that actually improves speaker discrimination for this ASR system, but further investigation would be needed to support this hypothesis.

With the 4000 Hz bandwidth filter the range of scores was roughly the same, but recognition accuracy for individual conditions changed by as much as 10%. Accuracy improved or remained the same for all vocoders except the RELP processors which both degraded somewhat. Accuracy for clear speech and the pulse code modulated (PCM) processor improved to 95%, and accuracy using the quantized LPC parameters improved to 90%. It will be noted that Speaker 4 was recognized correctly under only two conditions. There is no readily apparent explanation for this, however it was verified that there were no procedural errors and that there was no unusual imbalance in the phonetic content of the training and testing sentences for this person.

It is curious that recognition accuracy for the RELP processors should degrade slightly with the wider bandwidth filter, while accuracy for all the other vocoders improved or remained the same. One possible explanation for this is the type of signal processing performed in these systems. These processors synthesize an accurate representation of the low-frequency portion, or baseband, of the speech signal and approximate the high-frequency portion using information in the baseband residual. For this reason the high frequency spectrum of the synthesized speech often differs significantly from that of the original speech, particularly for voiced sounds. This inaccuracy is not noticeable to human listeners, but could have a marked effect on an ASR algorithm that weights all portions of the speech spectrum equally. The speaker information is therefore contained primarily in the baseband of these coders, and the amount of additional information in the high frequencies is outweighed by the inaccuracy of the spectral representation. With this in mind, it is not surprising that these processors do not benefit from the wider bandwidth. The 16,000 bps RELP degraded less than the 9600 bps because it has a wider baseband.

Table 2. Summary of test results for 2500 Hz bandwidth tests. Speakers 1-10 are males, 11-20 are females. A dot indicates that the speaker was identified correctly under the given condition; a number indicates an identification error. The number entered is the identified speaker, while the row number is the actual speaker. The total number of errors is listed at the bottom of the table for each condition, along with the recognition accuracy measured with a granularity of 5%.

Speaker Number	Clear	PCM 64000	RELP 16000	CVSD 16000	RELP 9600	APC 9600	LPC 2400	[illegible] LPC	Phone
1	•	•	•	•	•	•	•	10	•
2	•	•	•	1	•	1	1		•
3	•	•	•	•	•	•	•		13
4	5	5	5	6	10	6	5	10	18
5	•	•	•	•	•	13	10	10	•
6	•	•	•	•	•	•	•	•	9
7	•	•	•	•	•	•	•	•	18
8	•	•	•	14	•	14	•	•	10
9	•	•	•	•	•	•	•	•	•
10	•	•	•	•	•	•	•	•	13
11	•	•	•	•	•	19	•	20	•
12	14	•	•	14	•	14	14	•	7
13	•	•	•	16	•	•	•	•	16
14	•	•	•	•	•	•	•	•	19
15	•	•	•	•	•	•	•	•	•
16	•	•	•	•	•	•	•	•	5
17	•	•	•	19	•	•	•	•	•
18	•	•	•	•	•	•	•	•	•
19	•	•	•	•	•	•	•	•	7
20	•	16	•	•	•	•	•	•	19
Total Errors	2	2	1	6	1	6	4	4	12
Percent Recognition	90	90	95	70	95	70	80	80	40

Table 3. Summary of test results for 4000 Hz bandwidth tests. As in Table 4 a dot indicates that the speaker was identified correctly for the given condition, and a number indicates an identification error.

Speaker Number	Clear	PCM 64000	RELP 16000	CVSD 16000	RELP 9600	APC 9600	LPC 2400	Q-LPC 2400	Phone
1	•	•	•	•	•	•	•	•	•
2	•	•	•	•	•	•	1	•	•
3	•	•	•	•	•	•	•	•	13
4	•	5	5	6	5	5	5	•	18
5	10	•	•	•	10	•	10	6	6
6	•	•	•	•	•	•	•	•	3
7	•	•	•	•	•	•	•	•	18
8	•	•	•	14	•	14	•	•	19
9	•	•	•	•	•	•	•	•	•
10	•	•	•	•	•	•	•	•	18
11	•	•	•	•	•	14	•	•	•
12	•	•	•	14	•	14	14	•	7
13	•	•	20	19	19	•	•	20	19
14	•	•	•	•	•	•	•	•	19
15	•	•	•	•	•	•	•	•	•
16	•	•	•	•	•	•	•	•	20
17	•	•	•	•	•	•	•	•	•
18	•	•	•	•	•	•	•	•	7
19	•	•	•	•	•	•	•	•	7
20	•	•	•	•	•	19	•	•	7
Total Errors	1	1	2	4	3	5	4	2	14
Percent Recognition	95	95	90	80	85	75	80	90	30

The relatively poor performance of the 16000 bps CVSD and the 9600 bps APC vocoders can be attributed primarily to the wideband quantization noise generated by these algorithms. Though not overly distracting to human listeners, this noise corrupts the speech signal enough that the subsequent LPC analysis is unable to extract the necessary speaker identification characteristics. This problem is also seen in the similar pattern of speech intelligibility scores shown in Table 4. These scores were generated by combining the processors used in this study with the 2400 bps LPC vocoder in a tandem configuration [10]. This does not necessarily indicate that these vocoders are unsuitable for ASR. It does suggest, however, that a non-LPC based approach might be better suited to systems of this sort.

High bit rate channels such as the 64,000 bps PCM system tested here probably generate speech of sufficient quality to be used with any type of ASR algorithm. Further research is required to determine exactly which parameters, analysis methods and distance measures produce optimum ASR results for a given voice processing system or type of system.

Table 4. Diagnostic rhyme test (DRT) speech intelligibility scores from tandem configurations of the given processor into the DoD LPC-10.

	Clear	PCM 64000	RELP 16000	CVSD 16000	RELP 9600	APC 9600	LPC 2400
Recognition Accuracy 2500 Hz bw	90	90	95	70	95	70	80
Recognition Accuracy 4000 Hz bw	95	95	90	80	85	75	80
Tandem Speech quality (DRT)	87	--	--	75	79	77	80

V. SUMMARY

Automatic speaker recognition offers great potential benefit to the Navy for a variety of applications, including communication security. The ability to perform ASR using processed or vocoded speech would allow verification of communication channel users, authentication of reports from remote sites, monitoring of channel activity, and access control for the channel itself.

A series of tests was conducted to evaluate the potential for performing ASR over military communication channels. In these tests the vocoded output of six DoD voice processors was used as input to an existing real-time ASR system. The vocoders used a variety of processing algorithms and had data transmission rates ranging from 2400 to 64,000 bps.

The results of the tests indicate that ASR using processed or vocoded speech is definitely feasible. However, further research is needed to raise the recognition accuracies to 99% and above as required by the military. In addition, research is needed to determine which acoustic parameters produce the highest recognition accuracy for a given voice processing system. Despite the many years of research in speech acoustics, relatively little is known about the speaker-specific characteristics of voices, or about which cues human listeners use in identifying speakers. Further investigation in these areas would also lead to increased recognition accuracy for ASR systems.

VI. ACKNOWLEDGMENTS

The author wishes to acknowledge the contributions to this study of Alan Higgins and Joe Naylor of ITT Defense Communications Division in San Diego who performed the testing and provided numerous insights into this problem.

Many thanks are also due to George Kang of NRL for his valuable advice and assistance with this investigation.

VII. REFERENCES

1. A. Schmidt-Nielsen and K. R. Stern, "Identification as a Function of Familiarity of Known Voices Talking Over an Unprocessed Channel and an LPC Voice Processor," NRL Memorandum Report 5382, July 1984. AD A145545

2. A. E. Rosenberg, "Listener Performance in Speaker Verification Tasks," IEEE Trans. Audio Electroacoustics, vol. AU-21, pp. 221-225, 1973.

3. C. A. McGonegal, A. E. Rosenberg and L. R. Rabiner, "The Effects of Several Transmission Systems on an Automatic Speaker Verification System," Bell Sys. Tech. Jour., vol. 58, pp. 2071-2087, 1979.

4. E. H. Wrench, Jr., "Automatic Speaker Recognition System, Appendix A," ITT Defense Comm. Div., San Diego, CA, Proposal 36031 (to Naval Res. Lab., Washington, DC), Feb. 1984.

5. G. S. Kang and L. J. Fransen, "Second Report of the Multirate Processor (MRP) for Digital Voice Communications," NRL Report 8614, Sept. 1982. AD A120591

6. MIL-STD-188-113, "Common Long Haul/Tactical Standards for Analog-to-Digital Conversion Techniques."

7. T. E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," Speech Technology, vol. 1(2), pp. 40-49, April 1982.

8. J.D. Markel and A.H. Gray, Jr., Linear Prediction of Speech, Spring-Verlag, NY, 1976

9. A. Higgins and J. Naylor, Final Report on Contract N00014-84-C-2130, ITT Defense Comm. Div., San Diego, CA, July 1984.

10. G.F. Sandy and J.E. Parker, "Digital Voice Processor Consortium Final Report," Mitre Corp. Report #MTR-84W00053-01, March 1984.