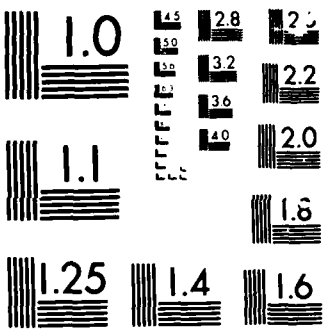
AD-A169 707 SPEAKER RECOGNITION USING PHONEME SPECIFIC SENTENCES 1/1
(U) NAVAL RESEARCH LAB WASHINGTON DC A SCHMIDT-NIELSEN
24 JUN 86 NRL-MR5801

UNCLASSIFIED F/G 17/2 NL


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

Naval Research Laboratory

Washington, DC 20375-5000 NRL Memorandum Report 5801 June 24, 1986

Speaker Recognition Using Phoneme Specific Sentences

ASTRID SCHMIDT-NIELSEN

Communication Systems Branch
Information Technology Division

AD-A169707

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 5801	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Naval Research Laboratory 6b OFFICE SYMBOL (If applicable) Code 7520	7a NAME OF MONITORING ORGANIZATION
6c ADDRESS (City, State, and ZIP Code) Washington, DC 20375	7b ADDRESS (City, State, and ZIP Code)
8a NAME OF FUNDING/SPONSORING ORGANIZATION 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 208010N PROJECT NO: X0919 TASK NO: WORK UNIT ACCESSION NO: DN260-149

11 TITLE (Include Security Classification)

Speaker Recognition Using Phoneme Specific Sentences

12 PERSONAL AUTHOR(S)

Schmidt-Nielsen, Astrid

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Final	FROM 10/84 TO 9/85	1986 June 24	19

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Speaker recognition → Linear Predictive Coding (LPC) Human listeners Voice communications

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

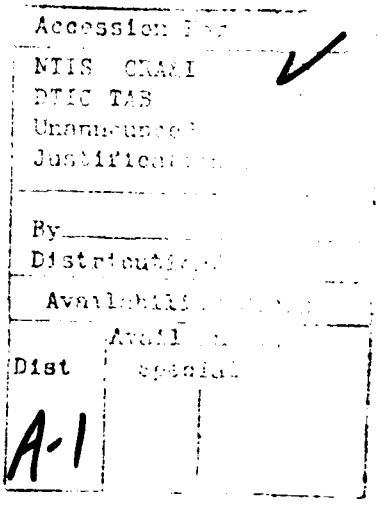
Listener tests involving the ability to distinguish between previously unfamiliar voices were conducted with phoneme specific sentences (each sentence contains only certain classes of consonant phonemes) using a familiarization-test procedure. There were three test conditions comparing unprocessed speech, LPC processed speech, and speech at 800 bits/sec. The results suggest that for unprocessed speech, speakers are better recognized when speaking sentences that contain voiced fricatives or voiced or unvoiced stops and are not as well recognized if the sentences contain only glides or only nasals. On the other hand, sentences with voiced fricatives and nasals were best for LPC speech. The results were also highly dependent on the grouping of the speakers. The recognition of more distinctive male voices and of female voices went down with LPC and 800 bit/sec processing as expected, but the recognition of less distinctive males was no worse after processing than before. The fact that the effects of voice processing vary with the composition of the speaker set is discouraging to the development of a standardized test of speaker recognition. Speaker recognition with the 800 bit/s algorithm was very poor but performance was still better than pure guessing.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION	
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED	
22a NAME OF RESPONSIBLE INDIVIDUAL Astrid Schmidt-Nielsen	22b TELEPHONE (Include Area Code) 202-767-2682	22c OFFICE SYMBOL Code 7526

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

CONTENTS



SPEAKER RECOGNITION USING PHONEME SPECIFIC SENTENCES

INTRODUCTION

The ability to recognize the speaker is impaired when talking over a narrowband voice communication system such as the DoD standard LPC (linear predictive coder) at 2400 bits/s (Federal Standard 1015 or MIL-STD-188-113). In previous research (Schmidt-Nielsen and Stern, 1985a), we found that the recognition of familiar speakers over LPC was nearly 80% as good as recognition of the same speakers with unprocessed speech (69% correct and 88% correct, respectively). Recognition tests using familiar speakers are representative of many ordinary situations for recognizing speakers. Tests with previously unfamiliar speakers are representative of some types of conferencing situations and have the advantage that they allow the experimenter to control the degree of exposure to each of the voices. Speaker recognition tests with unfamiliar speakers are limited to a relatively small set of speakers tested at the same time because of the listeners' memory limitations and practical considerations of testing time. With small sets of unfamiliar speakers there is a large effect of speaker selection in that the context of the other voices in the set and how difficult they are to tell apart has an effect on the degree of speaker recognition. We found that speaker selection and the way speakers were grouped also had different effects on the magnitude of the recognition loss due to LPC processing (Schmidt-Nielsen and Stern, 1985b). When more distinctive speakers, who were relatively easy to tell apart in the unprocessed condition, were grouped

Manuscript approved April 18, 1986.

together, recognition over LPC dropped considerably and was approximately 70% of what it was with unprocessed speech; when less distinctive speakers, who were more difficult to tell apart in the unprocessed condition, were grouped together, there was no additional loss in speaker recognition due to LPC processing (although recognition was poor in both conditions).

It has been shown (Bricker and Pruzansky, 1966) that speakers are better recognized when the speech has a varied phonemic content than when utterances of the same duration with little phonemic variation (e.g. sustained vowels) are used. This is not surprising since the more varied utterances contain more information about the source (i.e. the speaker's vocal apparatus). A set of phoneme specific sentences designed for subjective evaluations of LPC based speech coders has been developed (Huggins and Nickerson, 1985) such that consonants with similar acoustic parameters are used in the same sentence. For example, the sentence Nanny may know my meaning has only vowels and nasal consonants. They showed that different sentence types were sensitive to different aspects of degradation due to LPC processing.

The research reported here was conducted to investigate the effect of the phoneme content of the speech on speaker recognition over the LPC system, to replicate the results of the previous research with regard to the effect of speaker selection on recognition over LPC, and to extend the results to an 800 bit/s algorithm using a pattern matching strategy based on the standard LPC algorithm. Six phoneme categories (glides, nasals, voiced stops, unvoiced stops, voiced fricatives, and unvoiced fricatives) were selected to determine how the phonemic content of the speech affects the ability to recognize speakers. The various distortions that may occur due to LPC processing for particular types of phonemes could have an effect on listeners' ability to recognize speakers. The sentences with only glides do not contain abrupt

offsets and onsets or rapid transitions and are likely to be well modeled by the LPC system. The sentences with all nasal sounds may not be adequately modeled by the LPC all-pole model since their spectra contain zeros, and this could produce some distortion. On the other hand, nasal resonances have been shown to be useful for automatic speaker recognition (Glenn and Kleiner, 1967), and sentences with mostly nasals might also be helpful to recognition by listeners. The sentences with voiced and unvoiced stops contain abrupt onsets and offsets as well as rapid transitions, both of which are likely to be degraded because of the averaging that occurs over the relatively long frame rate used by the LPC algorithm. Fricatives have less amplitude change than the stops, but the noise excitation is very different from voiced sounds, and voiced fricatives are particularly sensitive to distortions from incorrect voicing decisions. Some distortions introduced by LPC processing could conceivably increase certain speaker differences, but it seems more likely that most forms of distortion would destroy useful vocal tract information and thereby reduce speaker recognition. In addition to the two conditions used in the preceding experiments -- unprocessed and LPC -- an 800 bit/s algorithm using a pattern matching strategy based on the standard LPC algorithm was also included in the tests. Even though several of the speakers in this experiment were also included in the data base for the pattern matching algorithm, there was almost certainly even less speaker specific vocal tract information preserved using this system than with the LPC system.

METHOD

Speakers and speech materials. Three sets of 5 speakers were selected from a group of 24 speakers used in a previous experiment (Schmidt-Nielsen and Stern, 1985a). There were two sets of male speakers; the first group consisted of speakers who had been rated as having more distinctive or characteristic voices and the second group was rated as having less distinctive voices. For

the voices in this experiment there were two independent sets of distinctiveness ratings -- one by 24 people who knew the speakers and one by 54 listeners unfamiliar with the speakers, none of whom were listeners in the present experiment. Both groups used a 7-point scale to answer the question "How distinctive or characteristic is this person's voice?" The raters who were familiar with the speakers made their ratings from memory, and the unfamiliar raters heard tape recorded voice samples. The average of the two sets of distinctiveness ratings was used to assign the male voices to two groups. As there were not enough females for two groups, the third group, consisting of five female voices varying in distinctiveness, was selected from the original nine females.

All 15 speakers read the same familiarization paragraph lasting about 30 s, and each speaker also recorded 14 sentences, 2 for each of the 6 phoneme classes (used in the test portion of the experiment) and two sentences containing all of the consonant phonemes in English (used for the practice with feedback portion of the tests). The sentences are listed below:

GLIDES

Why were you weary?
Why were you away a year, Roy?

NASALS

Nanny may know my meaning.
Many young men owe money.

VOICED STOPS

Bobby did a good deed.
Grab a doggie bag.

UNVOICED STOPS

Take a copy to Pete.
Patty cut up a potato cake.

VOICED FRICATIVES

There's usually a valve.
View these azure vases.

UNVOICED FRICATIVES

A thief saw a fish.
Three chefs face a thief.

ALL CONSONANT PHONEMES IN ENGLISH

Nothing could be further from reality than his illusion of taking your gorgeous sheep away.
The voyagers have ground the crankshaft with unimpeachable precision.

The 12 phoneme sentences for each of the 3 sets of 5 speakers were spliced apart and assembled into two counterbalanced sublists of 30 items, each preceded by its own set of 5 practice sentences. The control tapes were unprocessed, and the experimental tapes were prepared by processing the familiarization and test tapes for each of the 3 sets of speakers through the analysis and synthesis of a low data rate voice terminal once using the standard LPC algorithm and once using the 800 bit/s pattern matching algorithm.

Procedure. The experiment consisted of a familiarization phase in which the speakers' voices were introduced followed by a test phase during which the listeners tried to identify the phoneme sentences spoken by the different speakers. In the familiarization phase, listeners heard the tape recording on which each speaker introduced himself or herself giving a fictitious name starting with one of the letters from A to E, by saying, "Hello, my name is _____," and then reading the familiarization paragraph, which was about quicksand. To minimize confusion for the listeners, the familiarization paragraphs were always presented by speakers in order from A to E. The listeners were given typed copies of the text so that they could concentrate on the voice rather than the content. The listeners were asked to rate each of the voices on two 7-point rating scales (pleasant/unpleasant and ordinary/unusual) in order to insure that they attended to the characteristics of the voices. The five paragraphs were followed by a test, which was given in two halves. Each half started with a practice sequence in which five exemplars, one for each speaker, of one of the sentences containing all English consonant phonemes were presented in random order. Feedback about the identity of the speakers was given immediately afterward. This was followed by the 30 counterbalanced test sentences, which were given without feedback. In the second half, the second all-phoneme sentence was used for practice, and the remaining 30 test sentences were presented. The subjects wrote the letter

corrresponding to the speaker's name on a numbered answer sheet and checked a confidence rating of "very sure," "fairly sure," or "guessing." The subjects were instructed not to leave any blanks and to quess if they had to.

The listeners were 72 psychology students from the University of Maryland who volunteered to participate for extra course credit. The subjects were tested in groups of from 1 to 5 in a quiet room. Subjects were randomly assigned to one of the three processing conditions -- unprocessed, LPC, or 800 bits/s -- and were tested on all three speaker sets with the order of the speaker sets balanced across subjects. There were 24 listeners for each processing condition.

RESULTS

The overall recognition rates were relatively poor -- 41.2% for the unprocessed condition, 31.3% for LPC, and 26.6% for 800 bit/s. As pointed out previously (Schmidt-Nielsen and Stern, 1985b), this is a very difficult task. The listener hears five new voices for about 30s each and must remember the characteristics of each voice while trying to identify the correct voice for each of a series of sentences. There was no improvement in performance from the first 30 test items to the second 30 (in spite of the interpolated 5 practice sentences with feedback), so the results for the two test halves were combined. The effects of processing condition, speaker selection, and phoneme type were analyzed using analysis of variance; the results are shown in Table 1.

Processor and speaker set effects. The speaker set and processor effects (Figure 1) were significant and were similar to those of previous experiments (Schmidt-Nielsen and Stern, 1985b), confirming the strong context effect and the disproportional losses in recognition for the different speaker sets. For

Table 1. Analysis of variance for speaker recognition using phoneme specific sentences.

Source	df	SS	MS	F	p
Between Ss	71	1345.08			
Processors	2	482.50	241.25	19.30	<0.0001
S / processors	69	862.58	12.50		
Within Ss	1224	3419.56			
Speaker sets	2	336.50	168.25	21.87	<0.0001
Processors X spkr sets	4	242.90	60.72	7.89	<0.0001
S X spkr sets / processors	138	1061.82	7.69		
Phoneme type	5	32.92	6.58	3.81	<0.01
Processors X phon type	10	33.41	3.34	1.93	<0.05
S X phon type / processors	345	596.56	1.73		
Spkr sets X phon type	10	12.47	1.25	<1	NS
Processors X spkr sets X phon type	20	39.69	1.98	1.29	NS
S X spkr sets X phon type / processors	690	1063.29	1.54		
Total	1295	4764.64			

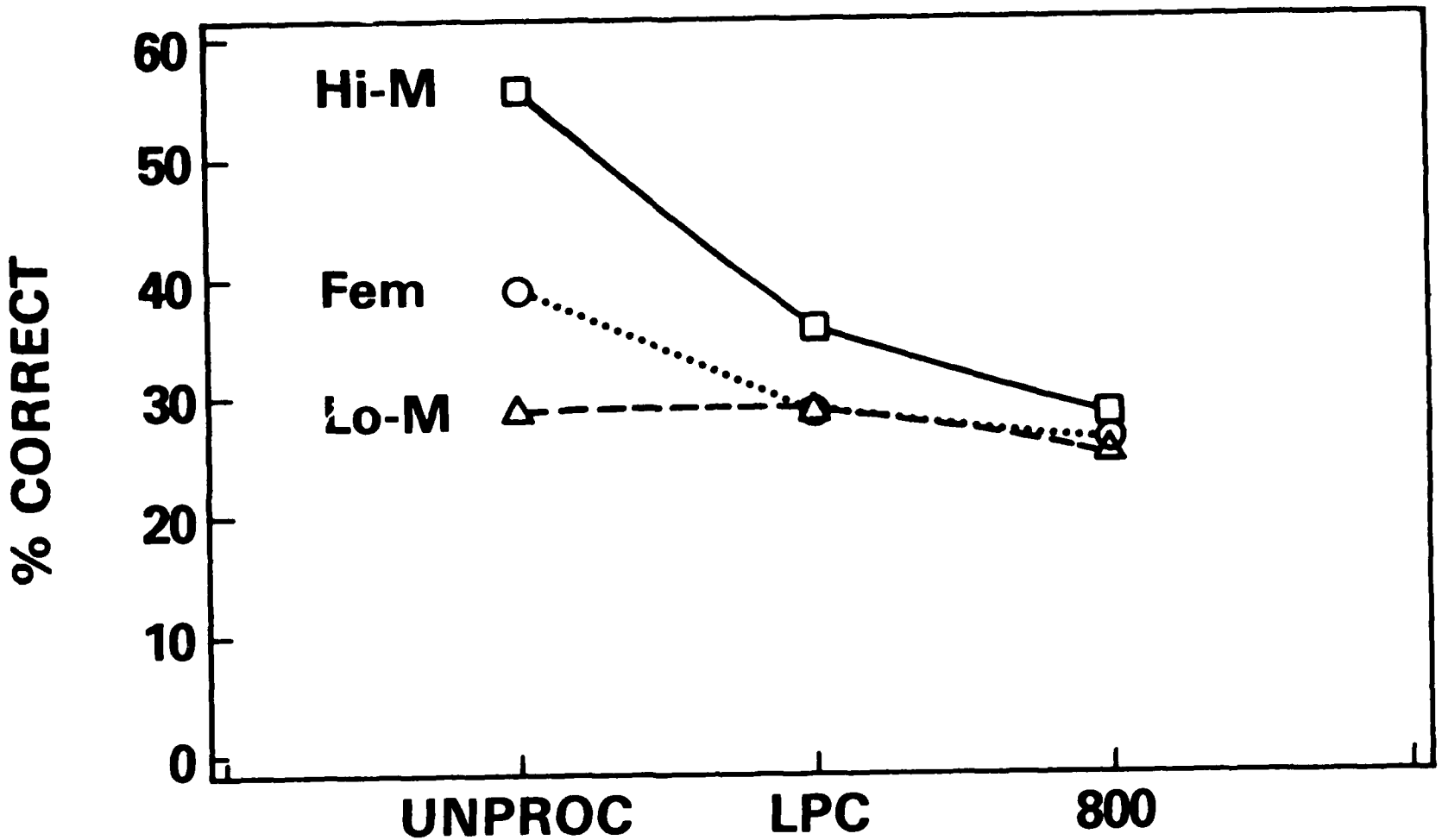


Figure 1. Speaker recognition scores for the voice processing conditions -- unprocessed, LPC processed, and 800 bits/s. Score are shown by speaker sets. Chance performance is 20% correct.

the unprocessed condition, the more distinctive males were better recognized than the females, who in turn were better recognized than the less distinctive males. The recognition of more distinctive male voices and of female voices went down with LPC and 800 bit/sec processing as expected, but the recognition of less distinctive males was no worse after processing than before. The overall difficulty of a speaker recognition task can be manipulated by various means such as changing the size or the composition of the speaker set. Varying the amount of exposure to the different voices should also have an effect, but this is not necessarily the case and may require the right kind of repeated exposures since there was no improvement from the first to the second half of this experiment, and Legge et al. (1984) found slightly poorer recognition with longer (60 s) than with shorter (30 s) initial exposures to each voice. The initial level of recognition should not be an important factor so long as one can expect the effects of the experimental manipulations to be similar for the different levels of initial performance. This expectation is met for speech intelligibility, where the results of tests of varying difficulty are all very highly correlated. When the difficulty of the speaker recognition task is varied by manipulating the composition of the speaker set, this expectation is not met. Averaging the results over all three speaker sets did give results that were very similar to those obtained in the previously reported experiment with familiar speakers (Schmidt-Nielsen and Stern, 1985a) even though the overall recognition of the familiar speakers was considerably better than for unfamiliar speakers in spite of a much larger speaker set to choose from. The inconsistency of the results of tests using unfamiliar speakers, depending on how the speakers are selected, is discouraging for the prospects of developing a reliable procedure for testing speaker recognition over various voice communication systems, like the tests that presently exist for testing speech intelligibility (e.g. the Diagnostic Rhyme Test; Voiers, 1977). A speaker recognition test proposed by

Papamichalis and Doddington (1984) used speakers selected to vary in the degree to which they were confused with one another, and this may be a promising approach.

The 800 bit/s speech had poorer speaker recogniton than the LPC, but even so performance was better than chance. This suggests that even when there is little specific vocal tract information available, listeners can still distinguish somewhat among speakers, probably on the basis of speech habits, suprasegmentals, and other global speech characteristics. Some of the speakers in this experiment happened to be among those whose voices were used to generate the table of patterns for this processor. Table 2 shows that when the 2400 bit/s LPC system was compared with the 800 bit/s system, the loss in speaker recognition was about the same whether or not the speaker's voice was included in the table of patterns. Ratings of voice unusualness and voice pleasantness made while listening to the familiarization paragraphs also showed no differences between the two groups of speakers.

Many of the low recognition scores were close to the chance level of 20%. This suggested the possibility of a floor effect, which could be a reason that the scores showed little or no loss for the low males in the LPC condition or at 800 bits/s. To test this possibility binomial tests were conducted and these showed that all of the scores shown in Figure 1 were significantly above chance, so there was still some opportunity for additional recognition loss.

Phoneme class effects. There was a significant effect due to phoneme type as well as a significant interaction of phoneme type by processing condition. Figure 2 shows the percentage of correct recognitions for each phoneme class for each of the three processing conditions. The results are averaged over speaker sets since there were no significant interactions of speaker set with

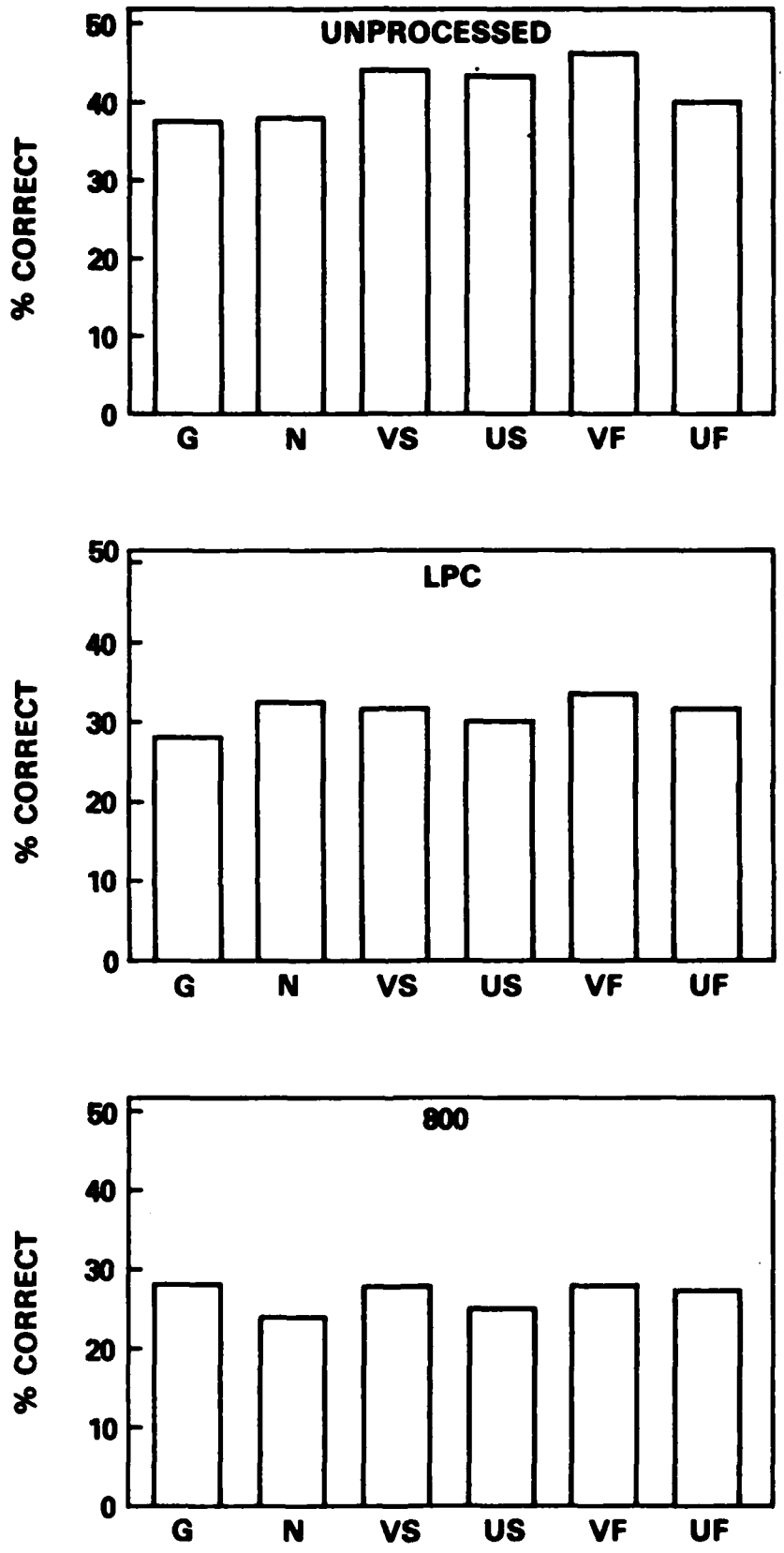


Figure 2. Speaker recognition scores by phoneme specfic sentence type for unprocessed, LPC, and 800 bits/s processed speech. G = glides, N = nasals, VS = voiced stops, US = unvoiced stops, VF = voiced fricatives, UF = unvoiced fricatives.

Table 2. A comparison of the speakers whose voices were incuded in the pattern table for the 800 bit/s system with the speakers whose voices were not included. There were no significant differeences between the two groups.

	10 Voices on 800 System			5 Voices not on 800 System		
	LPC	800	Difference	LPC	800	Difference
% Correct Speaker Recognition	31.8	26.8	5.0	29.9	26.3	3.6
Ordinary/ Unusual Voice Rating	3.27	1.87	1.40	3.34	2.04	1.30
Pleasant/ Unnpleasant Voice Rating	3.43	2.19	1.24	3.38	2.36	1.02

Table 3. Speaker recognition for phoneme specific sentences as a function of narrowband processing with the results expressed as a proportion of recognition performance for unprocessed speech.

	Sentence Type					
Processor	Glides	Nasals	Voiced Stops	Unv'd Stops	Voiced Frics	Unv'd Frics
LPC	0.751	0.853	0.723	0.697	0.730	0.800
800 Bits/s	0.751	0.634	0.637	0.577	0.609	0.684

phoneme type. This lack of interaction suggests that differences in recognition due to the content of the speech are similar across different speaker sets. Another way of viewing the pattern of phoneme results is to look at the relative losses within phoneme classes when the two types of processed speech are compared with the unprocessed speech. The proportion of correct recognitions relative to performance for unprocessed speech is shown in Table 3. For the unprocessed speech the glides and nasals seemed to be slightly worse for recognizing speakers than the other types of consonant sentences with more rapid formant and amplitude changes. In the LPC condition nasals and unvoiced fricatives seemed to show less loss in speaker recognition than the other types of sentences, and in the 800 bit/s condition the glides showed the least loss. Of necessity, the sentences contained a variety of vowels in addition to their consonant content, which means that the phoneme content of all of the sentences was reasonably varied even though only specific consonants classes were included. For this reason the magnitude of the differences between the sentence types was fairly small, and only some of the larger pairwise differences reached statistical significance. The results obtained here are suggestive and should be regarded as preliminary until a larger sample of speakers, sentences, and listeners can be tested. For most applications where speaker recognition is done by listeners, a varied phonemic content of the speech is to be expected, and the small differences obtained here suggest that using specific types of sentences is not likely to be particularly helpful. However, the fact that different phoneme types led to better recognition for processed than for unprocessed speech does suggest that different speech parameters may be the most useful for recognizing speakers under LPC and very low data rate voice processing conditions than for unprocessed speech, and this could have some application in developing automatic speaker recognition techniques for these voice transmission methods.

CONCLUSIONS

The phoneme content of the speech had a small but significant effect on speaker recognition performance and this effect varied with processing condition but was similar across speaker sets. The effects of speaker selection and the differences in the extent of recognition loss after LPC processing replicated the earlier results. This suggests that it would be very difficult to develop reliable methods for predicting speaker recognition over various voice communication systems using tests with small sets of unfamiliar speakers. In addition to the effects of LPC processing the results were extended to the 800 bits/s pattern matching algorithm, which showed very little additional loss in speaker recognition.

REFERENCES

Bricker, P. D., and Pruzansky, S. (1966). "Effects of stimulus content and duration on talker identification," J. Acoust. Soc. Am. 40, 1441-1449.

Glenn, J. W., & Kleiner, N. (1968). Speaker identification based on nasal phonation. J. Acoust. Soc. Am. 43, 368-372.

Huggins, A. W. F., & Nickerson, R. S. (1985). Speech quality evaluation using "phoneme-specific" sentences, J. Acoust. Soc. Am. 77, 1896-1906.

Legge, G. E., Grosmann, C., and Peiper, C. M. (1984). "Learning unfamiliar voices," J. Exp. Psychol.: Learn. Mem. Cognit. 10, 298-303.

Military Standard 188-113: Common Long Haul/Tactical Standards for Analog/ Digital Conversion Techniques. Naval Publications Forms Center, Philadelphia, PA.

National Communications System Office of Technology & Standards (1984). Federal Standard 1015: Analog to digital conversion of voice by 2400 bit/second linear predictive coding. Washington, DC: General Services Administration Office of Information Resources Management.

Papamichalis, P. E., and Doddington, G. R. (1984). "A speaker recognizability test," ICASSP 84. 1984 International Conference on Acoustics, Speech, and Signal Processing, San Diego, CA, (IEEE, New York), pp. 18B6.1-18B6.4.

Schmidt-Nielsen A., and Stern, Karen R. (1985a). Identification of known voices as a function of familiarity and narrow-band coding, J. Acoust. Soc. Am., 77 (2), 658-663.

Schmidt-Nielsen, A., and Stern, Karen R. (1985b). The Effect of LPC Processing on the Recognition of Unfamiliar Speakers. (Report No. 8926). Washington, DC: Naval Research Laboratory.

Voiers, W. D., (1977). "Diagnostic evaluation of speech intelligibility," Speech Intelligibility and Recognition, edited by Hawley, M. E. Strondsburg, Pa: Dowden, Hutchinson, and Ross.

END

DTIC

8-86