THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Speaker Verification by Human Listeners over Several Speech Transmission Systems

By C. A. McGONEGAL, L. R. RABINER, and B. J. McDERMOTT

(Manuscript received January 6, 1978)

Although a great deal has been learned about how speakers are verified, both by humans and by machines, several factors have not yet been studied. One of these factors is the effect of the transmission system (over which the message is communicated) on the accuracy with which verification is achieved. This factor is potentially an important one for digital communications problems over telephone lines where the transmission system could vary from one which gives a high-quality coded representation of the signal (e.g., log PCM) to a low-bit-rate vocoder. The purpose of this paper is to demonstrate the effects of three speech transmission systems on verification accuracy by human listeners. It is shown that the false alarm rate (i.e., a customer is rejected) is significantly higher when the test and reference utterances are transmitted by different systems than when transmitted by the same system. The miss rate (i.e., an imposter is accepted) is not significantly different for similar comparisons except for one of the conditions. The overall conclusion of this experiment is that speaker verification by human listeners cannot be performed as accurately over mixed speech transmission systems as over the same transmission system.

I. INTRODUCTION

Speaker verification, both automatically by machine and by human listeners, is an important problem in the area of man-machine communication by voice.[1–8] The verification problem has applications in the business community for such things as voice banking by telephone, credit card transactions (including charging of telephone calls), and access of privileged or confidential information.

As shown in Fig. 1, the speaker verification problem, either by human listeners or by machine, has two aspects—the creation of a reference pattern (i.e., the training phase) and the determination of similarity between a test and a reference pattern (i.e., the testing phase). When

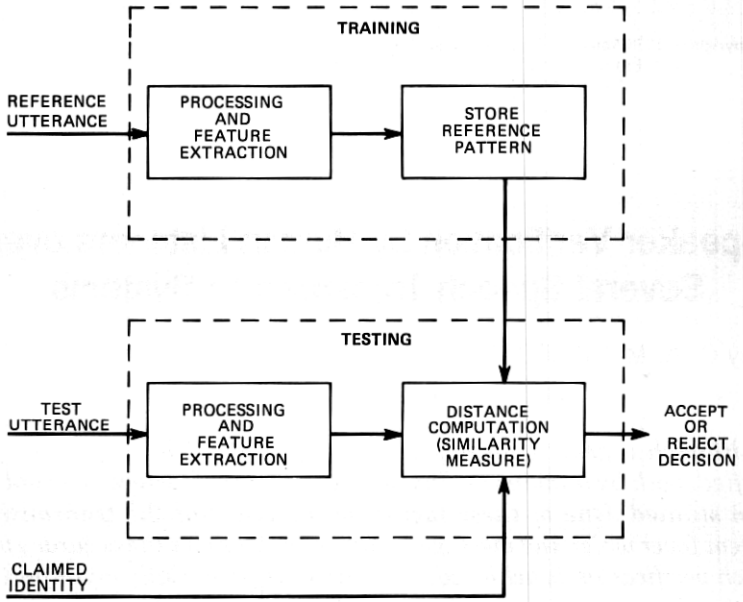

Fig. 1—Block diagram of the speaker verification problem.

verification is to be performed over telephone lines, an additional factor complicates the problem—namely, the transmission system used in the telephone plant. With the increased reliance on digital speech-processing techniques, such as waveform coders and linear prediction methods, the interesting possibility arises that the test pattern for speaker verification may have been coded or vocoded, whereas the reference pattern may not have been subjected to the same processing. Past experiments by Rosenberg[2] have studied the problem of verification both by human listeners and with an automatic system using natural speech for both the test and reference patterns. The purpose of this experiment is to evaluate how several speech transmission systems affect the process of speaker verification by human listeners. In future work, we will investigate the parallel problem—how these same factors affect automatic methods of speaker verification.

The organization of this paper is as follows. Section II describes the way in which the evaluation was carried out. This section includes a description of the speech transmission systems, as well as the experimental procedure used to measure system performance. In Section III, the experimental results are presented in terms of a signal detectability model, and in Section IV the results are discussed.

II. EXPERIMENTAL EVALUATION

For the experiment to be described below, both the reference and test utterances were preprocessed by one of the following three transmission systems:

(*i*) Bandpass filtering from 100 to 2600 Hz.

(*ii*) Adaptive differential pulse code modulation (ADPCM) coding, followed by bandpass filtering from 100 to 2600 Hz.

(*iii*) Linear predictive vocoding (LPC), followed by bandpass filtering from 100 to 2600 Hz.

The bandwidth of all three systems was set to 2500 Hz, in accordance with the requirements of the ADPCM coder, to ensure that the speech bandwidth was not a factor in determining relative verification accuracy.

The ADPCM coder used in this experiment was a simulation of the coder built by Bates,[9] based on the work of Cummiskey et al.[10] Figure 2 is a block diagram of the ADPCM system. The input signal is band-pass-filtered from 100 to 2600 Hz and sampled at a 6000-Hz rate. A 4-bit adaptive quantizer was used to code the difference signal, giving an overall bit rate of 24 kb/s for the coder. The step-size multiplier of the quantizer ranged over a 41-dB range (i.e., the ratio between the largest and smallest step size was 114 to 1). A first-order predictor was used with a multiplier of $\alpha = 0.9375$. Signal levels were chosen so that the coder was operating at approximately the optimum point.

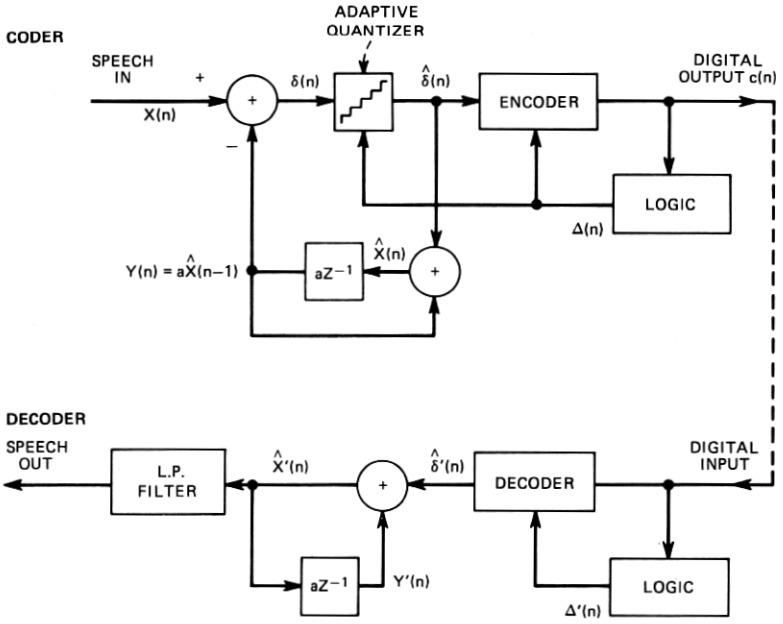

Fig. 2—Block diagram of an ADPCM coder.

A block diagram of the LPC vocoder is given in Fig. 3. The implementation was based on the autocorrelation method of linear prediction.[11–13] Pitch detection and voiced-unvoiced decision were performed using the modified autocorrelation pitch detector of Dubnowski et al.[14] The input signal was sampled at a 10-kHz rate, and a 12-pole LPC analysis was done with a pitch adaptive, variable frame size, at a rate of 100 frames per second.[15] No quantization of the LPC parameters was used in this experiment.

2.1 Data base for the evaluation

To evaluate the three transmission systems, a data base was designed which included

(*i*) 16 speakers designated "customers":
8 male.
8 female.
(*ii*) 62 speakers designated "imposters":
31 male.
31 female.
(*iii*) 2 sentences:
"We were away a year ago"—male utterance.
"I know when my lawyer is due."—female utterance.
(*iv*) 3 versions of each utterance:
bandpass filtered speech—SP.
ADPCM coded and filtered speech—ADPCM.
LPC vocoded and filtered speech—LPC.

The set of male utterances used in this study were those used by Rosenberg in his earlier work.[4] New recordings were made for the set of female utterances. Both male and female speakers recorded 10 utterances over a period of several weeks. The imposters provided just one recording each.

2.2 Experimental procedures

To test the effects on verification of combinations of different speech systems for the reference and test utterances, a paired-comparison test was used. A block diagram of the experimental arrangement used is shown in Fig. 4. Each test pair consisted of a comparison utterance and a challenge utterance. The comparison utterance was always a customer utterance processed by one of the three transmission systems. The challenge utterance was either an imposter utterance (customer-imposter pair) or one of the remaining nine utterances of the same customer (customer-customer pair) processed by one of the three systems.

Ten analog tapes were prepared. Each tape consisted of only male or female utterances with 48 customer-customer and 48 customer-imposter pairs randomly presented. The eight customers were presented in each

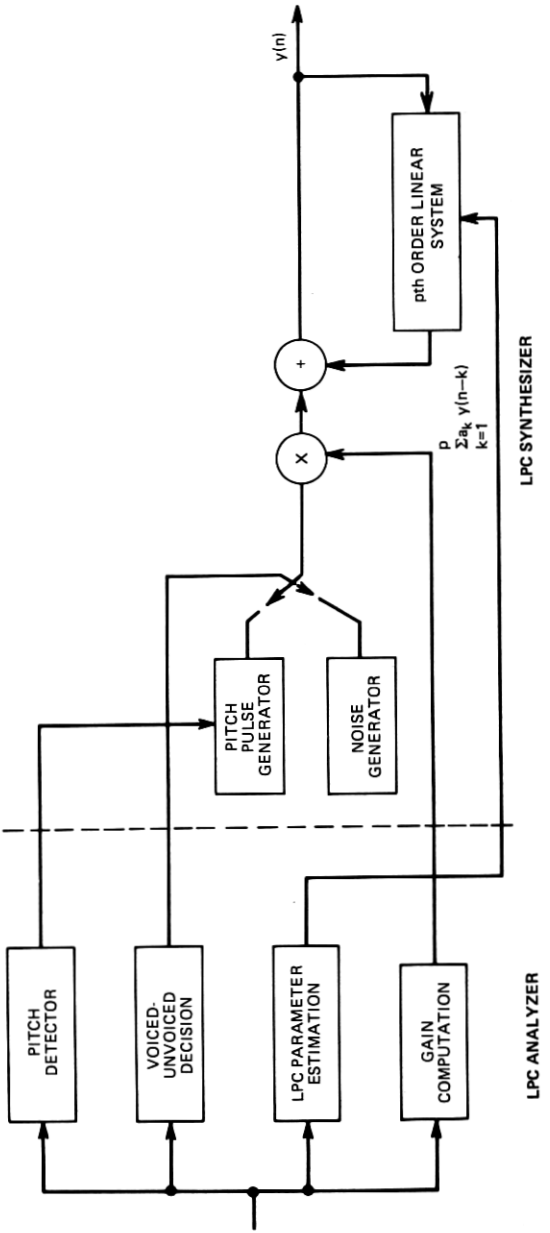

Fig. 3—Block diagram of an LPC vocoder.

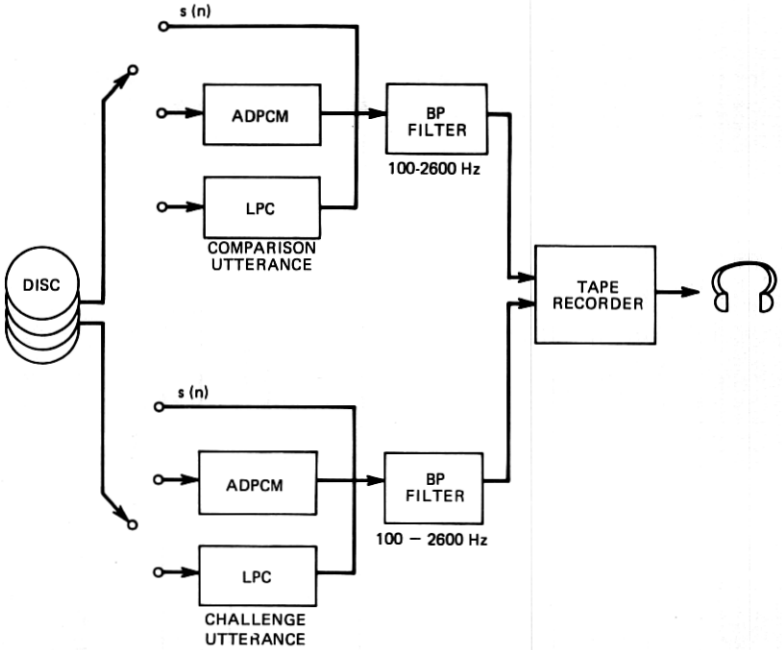

Fig. 4—Block diagram of the experimental arrangement.

of the transmission system combinations in both types of paired comparisons. When the transmission system combinations were heterogeneous, four of the eight customers were presented in each of the two orders.

Two tapes, one containing the male utterances and one containing the female utterances, were presented over headphones to five different groups of six naive subjects who were seated in a soundproof booth. The subjects were asked to indicate whether the comparison and challenge utterances were spoken by the same or by different speakers. They received no training for the experiment and were given no instructions as to the costs of any type of error.

III. EXPERIMENTAL RESULTS

The subject responses can be interpreted according to signal detection theory[16]—i.e., as a hypothesis test. For any trial in the test, there were two input hypotheses (same speakers or different speakers) and two possible subject responses (SAME and DIFFERENT). Therefore, each trial can be represented by the intersection of one of the input alternatives and one of the response alternatives as indicated in Fig. 5. There are two types of errors (*false alarm* and *miss*) and two types of correct responses (*hit* and *correct rejection*) associated with each trial. A *false alarm* (the rejection of a customer) is defined as a subject response of DIFFERENT

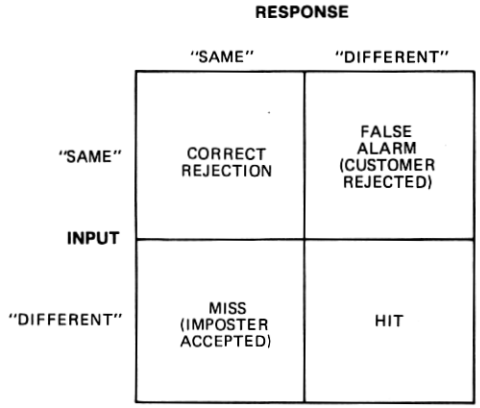

Fig. 5—The various response classifications for detecting an imposter.

when both utterances are spoken by the same speaker. A *miss* (the acceptance of an imposter) is defined as a subject response of SAME when the challenge utterance was spoken by a different speaker. A *hit* (acceptance of a customer) is defined as a subject response of DIFFERENT when the challenge utterance was spoken by a different speaker. A *correct rejection* (rejection of an imposter) is defined as SAME when both utterances are spoken by the same speaker.

The false alarm rates for male and female customers are shown in Fig. 6. The customer false alarm rates are represented by vertical bars—one bar per customer for each pair of transmission systems. The percentage of time a customer was rejected varied for each customer in a group and also between groups. Although customers were asked to record their sentence the same way at each session, several had dramatic pitch changes. Since the subjects were not familiar with the customer voices, they tended to reject those customers. In general, the false alarm rates were fairly low and in many cases less than 10 percent.

The miss rates for the male and female customers are shown in Fig. 7. The percentage of time an imposter was accepted also varied greatly among customers. As seen in this figure, the miss rates were generally higher than 15 percent for all transmission pairs except for the LPC-ADPCM pair.

An alternative way of displaying the information in the subject data is in terms of the likelihood ratio. The likelihood ratio, l, is a good measure of signal detectability and is defined as

$$l = \frac{P\,(\text{hit})}{P\,(\text{false alarm})},$$

where

$$P\,(\text{hit}) = 1 - P\,(\text{miss}).$$

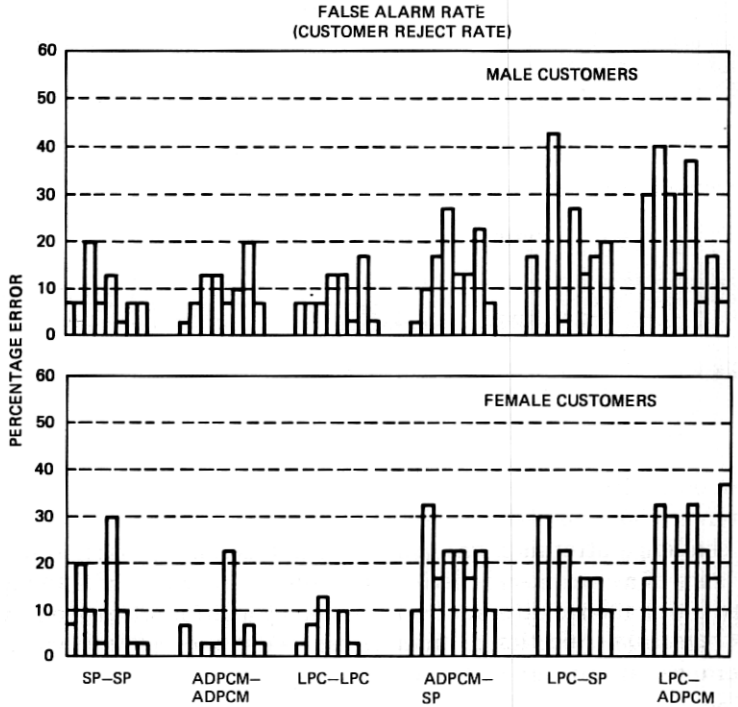

Fig. 6—False alarm rates for male and female customers for all transmission system pairs.

Figure 8 is a plot of the likelihood ratios for the male and female customers for transmission system pairs. The height of a vertical bar is the ratio of the number of times a subject correctly detected an imposter to the number of times a subject rejected a customer. Therefore, the larger the ratio (the higher the vertical bar), the better the subject performance for that transmission pair. As seen in this figure, the ratios are highest among the homogeneous systems for both male and female customers and lowest among the mixed systems. For several of the female customers, there were no false alarms so the likelihood ratios are infinite. Again, the large amount of variation among customers can be seen.

Because of the high variation among the customers (as seen in the preceding figures), the data for the false alarm and miss rates were pooled on the basis of median error scores rather than mean error scores. The median errors of the eight customers for each pair of transmission systems are shown in Fig. 9. Both a chi-square and a Fisher test[17] were applied to the median data to determine when significant differences existed between (*i*) the male and female customer medians for each pair of transmission systems (no significant differences were indicated) and (*ii*) the combined male and female customer medians of each pair of

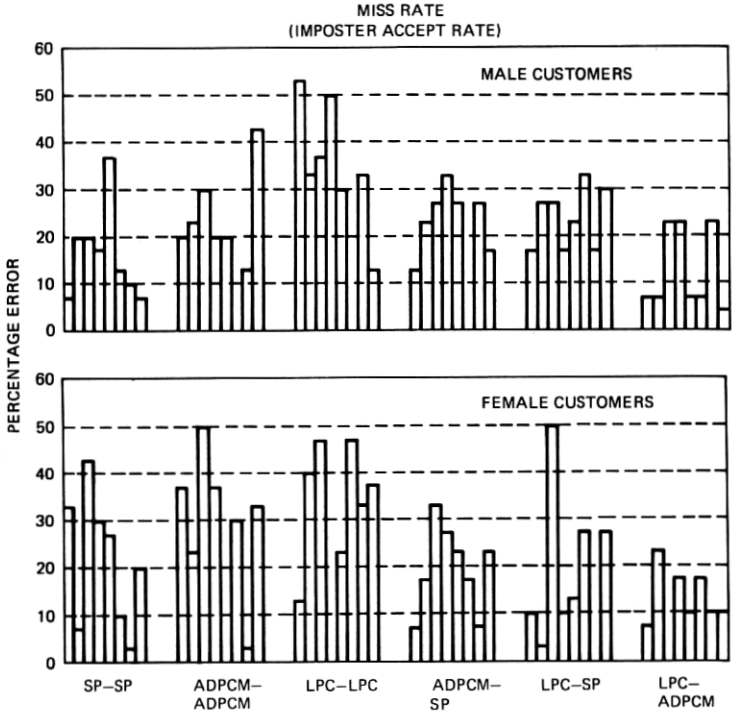

Fig. 7—Miss rates for male and female customers for all transmission pairs.

transmission systems. The following significant differences were found.

(*i*) The false alarm rate for mixed systems, that is, when the test and reference utterances were processed by different transmission systems, is significantly different from that of homogeneous systems.

(*ii*) The false alarm rate for the SP-SP pair was significantly different from all other transmission pairs.

(*iii*) The miss rates for mixed and homogeneous systems were not significantly different except for two transmission pairs. The LPC-ADPCM system had a significantly lower miss rate than the other system pairs, and the LPC-LPC pair had a significantly higher miss rate than any other transmission system pair.

Finally, Fig. 10 shows the overall error rates, that is, the average of the false alarm rates and miss rates, for each speech transmission pair. The overall error rate is between 10 and 20 percent for all speech transmission pairs. The lowest overall error rate is observed for SP-SP and ADPCM-ADPCM transmission pairs. There is no significant difference in the overall error rate for any of the system combinations.

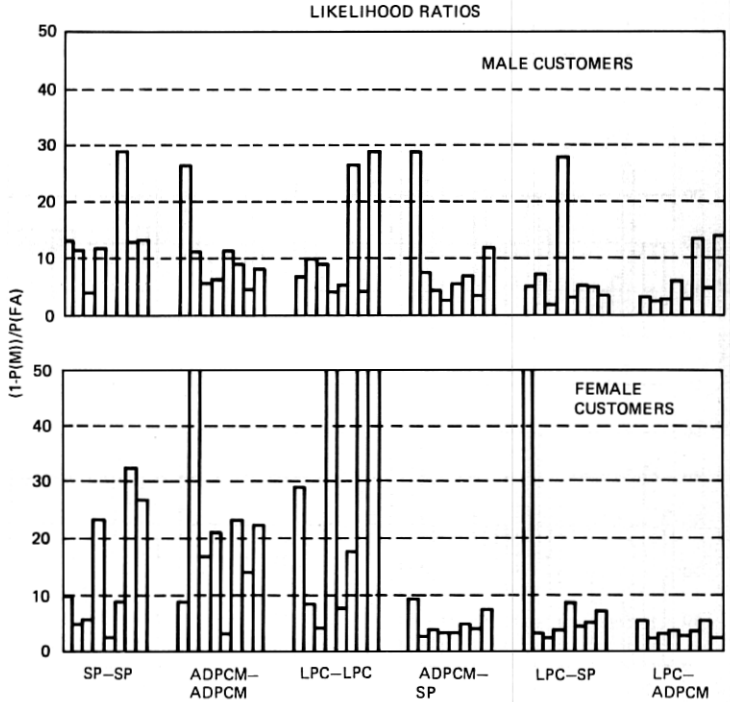

Fig. 8—Likelihood ratios for male and female customers for all transmission pairs.

IV. DISCUSSION OF THE RESULTS

The results (i.e., the false alarm and miss rates for the speech transmission pairs) can be interpreted in terms of the dimensions of difference between transmission systems. There are three main differences between system pairs, as shown in Fig. 11. These are:

(*i*) No difference—the same transmission system is used for both the test and reference utterances. The transmission system pairs in this category are SP-SP, ADPCM-ADPCM, and LPC-LPC.

(*ii*) One dimension of difference—the transmission system pair consists of one utterance transmitted over a clear channel and the other utterance processed over an ADPCM or LPC system. The transmission system pairs in this category are ADPCM-SP and LPC-SP.

(*iii*) Two dimensions of difference—two very different transmission systems are used for the test and reference utterances. Only the LPC-ADPCM pair is in this category.

In category (*i*), the median customer rejection (false alarm) rates were very low, and the median imposter acceptance (miss) rates were very high. This result reflects a subject bias toward responding SAME when the test and reference utterances are processed over the same transmission system. The low customer rejection rates in this category also indicate that subjects can easily verify customer-customer pairs.

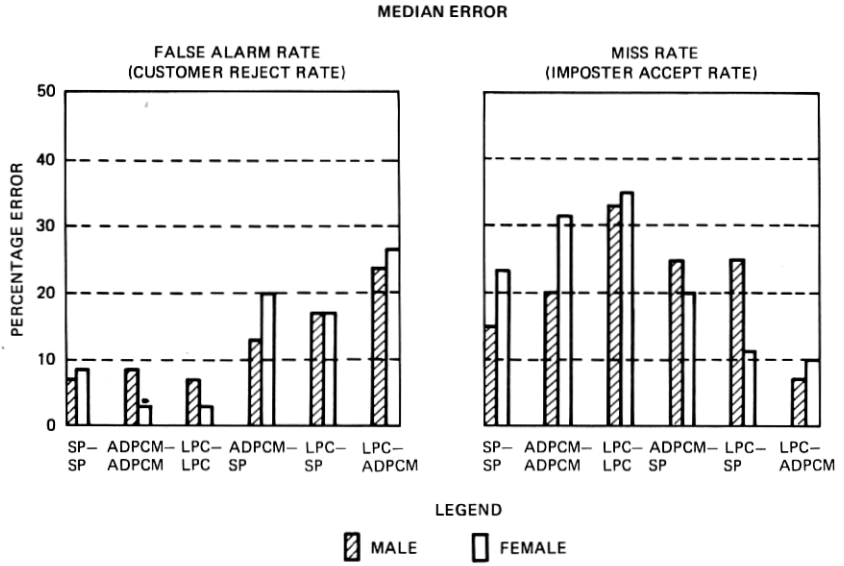

Fig. 9—Median false alarm rates and median miss rates for all customers and for all transmission pairs.

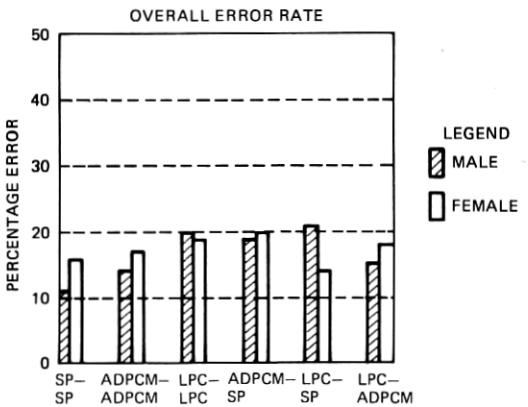

Fig. 10—Overall error rates for all customers and for all transmission pairs.

In category (*ii*), the median error rates were approximately the same for both customer-customer and customer-imposter pairs. For these cases, about 15 to 20 percent of the time a customer would be rejected and an imposter would be accepted. Whether a customer or an imposter was processed over either one of the transmission systems seemed to make very little difference. This result indicates that the subjects were confused by the pairing of an ADPCM or LPC system with a natural speech utterance.

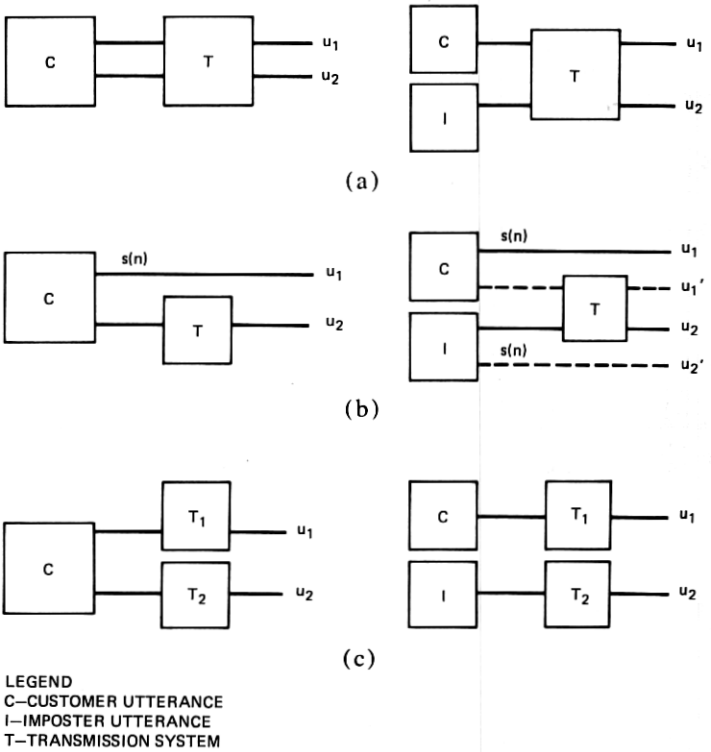

Fig. 11—Dimensions of difference between transmission system pairs. (a) No difference—SP-SP, ADPCM-ADPCM, LPC-LPC. (b) One dimension of difference—ADPCM-SP, LPC-SP. (c) Two dimensions of difference—LPC-ADPCM.

In category (*iii*) (i.e., LPC-ADPCM pair), the median customer rejection rate was very high and the median imposter acceptance rate was very low. The speech quality produced by the ADPCM and LPC systems is extremely different, and the results seem to reflect a subject bias toward responding DIFFERENT in this situation.

The overall conclusion is that the speaker verification task by human listeners is easiest when homogeneous systems are used and is significantly more difficult when mixed systems are used for the test and reference patterns.

V. COMPARISONS WITH PREVIOUS WORK

Since the male utterances used in Rosenberg's experiment[4] were also used in this experiment, the male SP-SP error rates can be compared. The median false alarm and miss rates for the two experiments are shown in Table I. The median false alarm rate observed by Rosenberg was 3.3 percent, which is about two times smaller than the error rate seen in this study. The median miss rate observed by Rosenberg was only 2.8 percent,

Table I — Comparison of male SP-SP median error rates

Experiments	False Alarm Rate	Miss Rate
Current experiment	7%	15%
Rosenberg[4]	3.3%	2.8%*

* The average miss rate after completion of 25 percent of the listening sessions was six percent.

which is considerably less than that observed here. Ideally, the error rates for the two experiments should be the same. However, several differences between the two experiments may have influenced the results. These are:

(*i*) The bandwidth used in Rosenberg's experiment was 4 kHz, whereas the bandwidth used here was 2.5 kHz. The 2.5 kHz bandwidth was required by the ADPCM system, which had a sampling frequency of 6 kHz.

(*ii*) The written instructions given to the subjects differed in both experiments. In Rosenberg's experiment, the subjects were divided into two groups. One group was provided with instructions intended to lower the false alarm rate, while the other group was provided with instructions intended to lower the miss rate. In this experiment, all subjects received the same instructions with no intent to lower either type of error rate.

(*iii*) In this experiment, no repeat judgments were obtained from any one subject, but in Rosenberg's experiment, 32 repeated judgments were obtained from each subject for customer-customer pairs and 4 judgments were obtained for each customer-imposter pair. Even though there was no prior training in either experiment, Rosenberg noted a training effect imbedded in his data. He found an average miss rate of 6 percent after the completion of 25 percent of the listening sessions. This rate is approximately two times less than the miss rate we observed. No drop was noticed with regard to the false alarm rate.

(*iv*) The last factor that may have influenced the difference in the results is the fact that Rosenberg's experiment consisted entirely of SP-SP test presentations. In this experiment, the SP-SP presentations were randomly combined with all other transmission pair presentations.

VI. SUMMARY

The purpose of this experiment was to show the effect of different transmission systems on speaker verification accuracy by human listeners. It was shown that when the reference and test utterances were recorded from different transmission systems, the false alarm rate was significantly larger than when they were recorded from the same transmission system. However, with one exception, the miss rates were essentially equivalent, independent of the transmission system. As such, it is concluded that speaker verification by humans cannot be performed as accurately when different transmission systems are used.

VII. ACKNOWLEDGMENT

The authors wish to acknowledge the assistance of A. E. Rosenberg in providing the male data base from his previous experiments in speaker verification tasks.

REFERENCES

1. J. L. Flanagan, "Computers that Talk and Listen: Man-Machine Communication by Voice," Proc. IEEE, *64,* No. 4 (April 1976), pp. 405–433.
2. A. E. Rosenberg, "Automatic Speaker Verification: A Review," Proc. IEEE, *64,* No. 4 (April 1976), pp. 475–487.
3. R. C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans. Audio Electroacoust., *AU-21* (1973), pp. 80–89.
4. A. E. Rosenberg, "Listener Performance in Speaker Verification Tasks," IEEE Trans. Audio Electroacoust., *AU-21* (1973), pp. 221–225.
5. A. E. Rosenberg, "Evaluation of an Automatic Speaker Verification System over Telephone Lines," B.S.T.J., *55,* No. 6 (JULY–August 1976), pp. 723–744.
6. S. K. Das and W. S. Mohn, "A Scheme for Speech Processing in Automatic Speaker Verification," IEEE Trans. Audio Electroacoust., *AU-19* (1971), pp. 32–43.
7. G. R. Doddington, "A Method of Speaker Verification," Ph.D. dissertation, Univ. Wisconsin, Madison, 1970.
8. E. Bunge, "Automatic Speaker Recognition by Computers," in Proc. Carnahan Conf. Crime Countermeasures, 1975.
9. S. L. Bates, "A Hardware Realization of a PCM-ADPCM Code Converter," M.I.T. M.S. Thesis, Dept. of Elec. Eng. and Comp. Sc., January 1976.
10. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52,* No. 7 (September 1973), pp. 1105–1118.
11. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech,* Springer-Verlag, 1976.
12. J. D. Markel and A. H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based Upon an Autocorrelation Method," IEEE Trans. on Acoustics, Speech, and Signal Proc., *ASSP-22,* No. 2 (April 1974), pp. 124–134.
13. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Amer., *50,* (1971), pp. 637–655.
14. J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-Time Digital Hardware Pitch Detector," IEEE Trans. Acoust., Speech, Signal Proc., *ASSP-24* (February 1976), pp. 2–8.
15. L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," IEEE Trans. Acoust., Speech, and Signal Proc., *ASSP-25,* No. 1 (February 1977), pp. 24–33.
16. D. M. Green and J. A. Swets, *Signal Detection and Psychophysics,* New York: Wiley, 1966.
17. S. Siegel, *Nonparametric Statistics for Behavioral Sciences,* New York: McGraw-Hill, 1956.