Listener Performance in Speaker Verification Tasks

AARON E. ROSENBERG

Abstract—The ability of listeners to perform some speaker verification tasks has been measured experimentally and compared with the performance of an automatic system for speaker verification. A test presentation in the subjective experiments consists of a pair of utterances. One of these is drawn from the recordings of a group of speakers designated customers while the second utterance is either a distinct recording from the same customer or the recording of an impostor. Listeners must respond whether the utterances are from the same or different speakers. The impostor classes that have been considered are casual impostors making no attempt to mimic customers, trained professional mimics, and an identical twin of a customer. Listener performance is specified by the two types of error that can be committed.

Two experiments have been carried out to evaluate the performance of listeners in speaker verification tasks. These experiments were intended to provide a basis for comparison with the performance of an automatic system for speaker verification developed at Bell Laboratories [1], [2].

Accordingly the same speech material used in the evaluation of the automatic system was used in the subjective tests. This material consists of recordings by male speakers of the all-voiced test sentence: "We were away a year ago."

In the first experiment these recordings were obtained from eight speakers designated as "customers" plus 32 other speakers designated "impostors." Several recordings were obtained from each of the customers over a period of several weeks. The impostors provided just one recording each. The designation impostor is arbitrary since these speakers simply uttered the same test sentence with no attempt to mimic the customers. We refer to them as casual impostors. A 33rd special impostor included in this experiment is the identical twin of one of the customers.

In the second experiment the same customer recordings were used but the casual impostors were replaced by professional mimics. After intensive training sessions in which they had free access to the original customer recordings, the mimics provided recorded test utterances intended to mimic each of the customers. In addition they provided recordings in their own natural voices.

The subjective tests were of the paired-comparison type in which each test presentation consists of a challenge utterance and a comparison utterance. Listeners are required to respond whether the challenge and comparison voices are from the same or different speakers. The comparison utterances were always customer utterances while the challenge utterances were drawn either from a set of impostor utterances or from a set of customer utterances distinct from those used as comparison utterances.

In every detection task there are two possible types of error associated with each response. In these experiments the response may be "different" when in fact the speakers are the same, which is rejection of a customer or a false alarm, or, the response may be "same" when in fact the speakers are different which is acceptance of an impostor or a miss.

Test stimuli were presented by means of headphones to each listener seated in a sound booth. Both the presentation of stimuli and the recording of responses were under the programmed control of a Honeywell DDP-516 laboratory computer.

Experiment I

With 8 customers and 33 impostors there are 264 possible different customer–impostor pairings. Four judgments were obtained from each listener for each pair. In addition there were 32 stimulus pairs consisting of a customer comparison utterance paired with a distinct challenge utterance from the same customer. Thirty-two judgments were obtained for each listener for each of these stimulus pairs so that, overall, the number of customer–impostor presentations was approximately equal to the number of customer–customer presentations. The stimulus-pairs were grouped into a series of 32 approximately 10-min long listening sessions and presented in a semi-randomized order.

The initial group of listeners consisted of ten young female clerks divided into two groups of five each. Group I was provided with written instructions intended to lower the accept impostor error rate (miss rate) while Group II was provided with written instructions intended to lower the reject customer error rate (false-alarm rate).[1] The listeners received no training except for a short familiarization session.

Results

The results are shown in the top half of Table I (labeled Experiment I). Mean and median miss rates, false-alarm rates, and overall error rates over nine listeners are shown as well as error rates for the indi-

Manuscript received April 30, 1972.
The author is with Bell Laboratories, Murray Hill, N.J. 07974.

[1] The results of one listener in Group II are omitted because an extraordinarily high miss rate was obtained.

TABLE I
Mean and Median Error Rates Over Listener Panels and Best Individual Error Rates

		MISS RATE (%)		F-A RATE (%)	OVERALL (%)
		MIMICS	NATURALS		
EXP. I	MEDIAN	—	2.8	3.3	3.6
	MEAN	—	3.6	4.0	3.9
	LOWEST MISS	—	1.0	8.9	4.9
	LOWEST F-A	—	3.4	0.8	2.3
EXP. II	MEDIAN	18.0	2.7	1.1	7.3
	MEAN	22.2	4.2	2.6	9.7
	LOWEST MM	4.1	2.0	1.1	2.4
	LOWEST MN	18.0	0.6	2.9	7.2
	LOWEST F-A	14.6	2.7	0.1	5.8

Note: F-A indicates false alarm. For Experiment II, the miss rate is specified separately for intentional mimic utterances and for natural voice utterances. The individuals labeled lowest MM and lowest MN in Experiment II are those with the lowest miss rates with respect to intentional mimic utterances and with respect to natural voices of mimics, respectively.

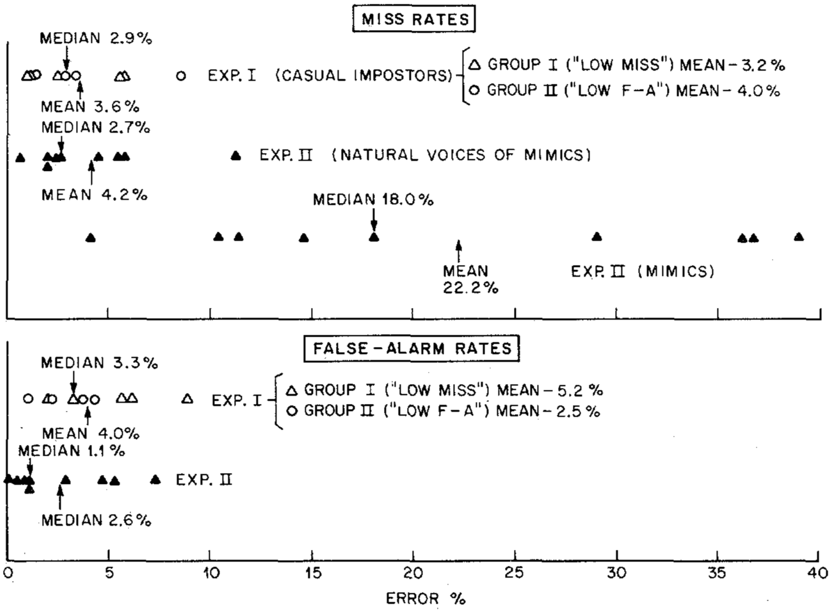

Fig. 1. Miss rates and false-alarm rates for individual listeners.

vidual listeners with the lowest miss rate and lowest false-alarm rate. The mean miss rate and false-alarm rate are both of the order of 4 percent. The median rates are slightly less suggesting that the consensus is towards a better group performance than indicated by the means. The best individual miss rates and false-alarm rates are of the order of 1 percent but are somewhat compensated by less impressive performances in opposing categories. Individual miss rates and false-alarm rates are plotted in Fig. 1 together with overall means and medians. Group I ("low miss") and Group II ("low false-alarm") individuals are shown by different symbols. There is some difference in performance between the two groups in the direction of the intended bias. For example, the mean miss rate for Group I is 3.2 percent and for Group II it is 4.0 percent; the mean false-alarm rate for Group I is 5.2 percent and for Group II, 2.5 percent. The separation between groups is more distinct with respect to false alarm rate. Not included in the calculation of miss rates is the special customer–impostor combination of identical twins mentioned previously. For this combination the impostor was accepted at an average rate of 96 percent. That is, the twin impostor was nearly totally accepted as his brother.[2]

Aside from special combinations of voices such as identical twins or other close relatives it is of interest

[2] In contrast, the automatic system was able to distinguish the brother without error. The difference in performance may be attributed to the fact that the twin impostor was a casual impostor in the sense that he did not mimic the inflection or intonation pattern of his brother. Although listeners may have been overwhelmed by the overall identical quality of the voices the automatic system found sufficient differences in the intonation pattern to reject the impostor.

for speaker verification studies to know the chances of obtaining pairs of significantly confusable voices in an arbitrary population of, say, adult male speakers. We can offer an estimate of these chances with respect to verification by listeners for our sample of speakers by examining the accept-impostor rates for each individual combination of customer and impostor. This information is transformed into a histogram in Fig. 2. This shows the percent of all voice pairs, or customer-impostor combinations, for which a particular range of miss rate (5 percent) was elicited. Again the combination of twins is omitted. We find that 90 percent of the combinations are associated with miss rates of 10 percent or less. Nevertheless there are two or three combinations out of this population with miss rates of the order of 40 percent or more indicating significant confusability.

One other result that can be noted is the effect of training or familiarity on listener performance. As noted in the introduction, aside from a short familiarization session, there was no training of the listeners prior to the collection of data, so that any effects of training or familiarization are imbedded in the data. It was found that after completion of 25 percent of the listening sessions the average miss rate was approximately 6 percent. This fell to approximately 3.5 percent after completion of 50 percent of the test and remained at about that level to the final completion of the test. No observable trend in performance level with regard to false-alarm rate is indicated.

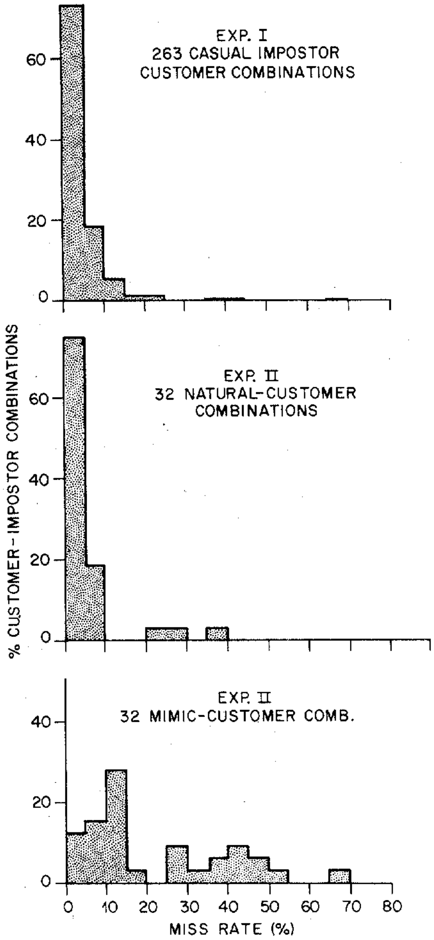

Fig. 2. Miss-rate distributions of impostor-customer combinations.

Experiment II

In this experiment as in the previous one the comparison utterances were always customer utterances. The challenge utterances were drawn with equal likelihood from three classes: customer utterances distinct from those used as comparison utterances, natural voice utterances from four professional mimics, and intentional mimic utterances from these mimics. The following challenge-comparison combinations were admitted: each customer challenge was paired with a comparison from the same customer for a total of eight combinations; each mimic natural voice was paired with each customer for a total of 4 × 8 = 32 combinations; finally each intentional mimic utterance was paired with the customer which was intended to be mimicked again for a total of 4 × 8 = 32 combinations. To balance the number of presentations four times as many judgments were collected from the customer-customer class of combinations than from the other two classes of combinations. There were three complete sets of challenge utterances each paired with four sets of customer comparisons. Each listener provided 3072 judgments in 32-, 10-, or 12-min listening sessions.

The panel of listeners in this experiment consisted of seven girls and two boys of high school or early college age. These listeners were paid for taking the test and were told that continued participation was conditioned on their maintaining acceptable levels of attendance and performance. (Three early participants were eliminated because of substandard performance and replacements obtained.)

The instructions given to the listeners were simply to decide carefully whether the challenge and comparison utterances were from the same or different speakers. They were not given any other information concerning the nature of the test.

Results

As in Experiment I the results are expressed in terms of miss rate and false-alarm rate. In addition we will distinguish miss rate with respect to natural voices of mimics and miss rate with respect to intentional mimic utterances. The mimic natural voice-customer combinations are comparable with the casual impostor-customer combinations of Experiment I.

The results are summarized in the bottom half of Table I (labeled Experiment II). Mean and median error rates across the nine listeners are shown as well as error rates for the individuals with the lowest miss rate with respect to intentional mimicking, lowest miss rate with respect to natural voices of mimics and lowest false-alarm rate, respectively. First, note that the mean false-alarm rate and miss rate for naturals are of the order of 3 or 4 percent which is comparable to the mean false-alarm rate and mean miss rate (with respect to casual impostors) obtained in Experiment I. However the mean miss rate with respect to intentional mimicking is significantly higher, on the order of 20 percent, indicating a considerable degree of success on the part of the mimics at imitating customer voices. Again, as in Experiment I, the listener median error rates are consistently lower than listener means. As for the lowest individual error rates in each category, again, as in Experiment I, they seem to be compensated by less impressive performance in remaining categories. Especially noteworthy, however, is the performance of the individual with the lowest miss rate with respect to intentional mimicking. This error rate is comparable with the listener mean miss rate for natural voice mimics or casual impostors! Her overall performance was the best of all listeners. Individual error rates are plotted in Fig. 1 with listener means and medians indicated by arrows. What is particularly observed in this figure is the clustering of individual error rates for each performance category with one prominent exception. This exception is the category of miss rate with respect to intentional mimicking. For this category the performance range across listeners is quite wide, extending from about 4–40 percent. Apparently, individual attitudes towards the somewhat peculiar mimicking voices are quite variable.

As in Experiment I, we can provide histograms obtained from a tabulation of miss rates for each customer–impostor combination that show a "confusability" distribution of the voices in our samples. These are shown in Fig. 2. For the mimic natural voice–customer combinations we find as in Experiment I that about 90 percent of the combinations have miss rates less than 10 percent with just two or three combinations in the range of 20–40 percent miss rates. The shapes of the distributions for both experiments are similar with perhaps a somewhat greater chance for extreme error rates in Experiment I where the population has considerably more voice pairs.

For intentional mimic–customer combinations the distribution is quite different. Only about 30 percent of the combinations have miss rates less than 10 percent while a significant fraction, nearly 25 percent, have miss rates greater than 40 percent. Thus the composition of this population of voice pairs is quite different from the arbitrary populations.

It is of interest to make some note of the skills of individual mimics and conversely the susceptibility of individual customers to mimicking. The skill of a mimic is indicated by averaging the listener miss rates over all the impostor-customer combinations which include intentional mimic utterances of that particular mimic. Ranked by this procedure, the best mimic is associated with a 37.5 percent miss rate. The others are associated with miss rates of 25.3, 19.3, and 6.6 percent, respectively. Note that the worst mimic performs at a miss-rate level which is not significantly better than the casual impostor or natural voice miss rate.

The susceptibility of individual customers to mimicking is assessed by averaging listener miss rates for customer–impostor combinations which include intentional mimic utterances over each customer. We find one highly susceptible customer, associated with a miss rate on the order of 40 percent, five customers moderately susceptible, with associated miss rates ranging from 15–25 percent, and two customers only slightly susceptible with associated miss rates of the order of 8 percent.

At the completion of the main series of test sessions listeners were interviewed individually to try to assess their attitudes and impressions of the test and the test material. It was reported that with experience most listeners achieved a fair amount of familiarity with the voices and could distinguish four, five, or six of the most distinctive voices. They also reported that some of the voices sounded peculiar or unreal but none guessed that there was any attempt at mimicking. On the contrary, one or two listeners felt that there was some attempt at disguising voices. At these interviews the listeners were informed of the exact nature of the experiment. In particular, they were told that one third of the voice pairs included a mimic. Subsequently a small auxiliary test series, 384 judgments per listener, was conducted to determine to what extent this information affects listener performance. As might be expected there was a dramatic decrease in the miss rate with respect to intentional mimicking: the listener median dropped from 18.0 to 7.0 percent. In addition there was an overall shift to decreased miss rates at the expense of increased false-alarm rates. This shift was most pronounced with individuals whose performance in the main test series was weaker than average.

Discussion

This discussion centers on some points of comparison between the performance of listeners and the performance of the automatic system. Statistics for testing an unknown sample on the automatic system are derived from measurements on known source populations of customers and casual impostors. The most current performance figures for the automatic system with respect to these source populations are 1 percent miss rate, (accept casual impostor), and no rejection of customers (error-free false-alarm rate). With respect to the utterances of the professional mimics as a test population the system performed with an average miss rate 14 percent. This performance is significantly better than the average performance of listeners where, as we recall, the comparable performances figures are the order of 3 or 4 percent for rejection of customers and acceptance of casual impostors and 22 percent for the acceptance of intentional mimic utterances.

However, the performance of the best listeners compares favorably with or even exceeds the performance of the automatic system. For example one listener accepted mimic utterances at a rate of only 4 percent. Although this particular listener was an exception, we have noted superior performance levels in one category are often accompanied by mediocre performance in other categories.

References

[1] G. R. Doddington, "A method of speaker verification," *J. Acoust. Soc. Amer.*, vol. 49, p. 139 (A), 1971.

[2] R. C. Lummis, "Real-time technique for speaker verification by computer," *J. Acoust. Soc. Amer.*, vol. 50, p. 106 (A), 1971.

Application of Sequential Decoding for Converting Phonetic to Graphic Representation in Automatic Recognition of Continuous Speech (ARCS)

C. C. TAPPERT, N. REX DIXON, and
ARTHUR S. RABINOWITZ

Abstract—Following segmentation and phonetic classification in automatic recognition of continuous speech (ARCS), it is necessary to provide methods for linguistic decoding. In this work a graph search procedure, based on the Fano algorithm, is used to convert machine-contaminated phonetic descriptions of speaker performance into standard orthography. The information utilized by the decoder consists of a syntax, a lexicon containing transcription variation for each word, and performance-based statistics from acoustic analysis. The latter contain information related to automatic segmentation and classification accuracy and certainty (anchor-point) data. A distinction is made between speaker- and machine-dependent corruption of phonetic input strings. Preliminary results are presented and discussed, together with some considerations for evaluation.

Introduction

The work described concerns phoneme-to-grapheme translation, i.e., the conversion of noisy phonetic sequences into standard orthographical representations of speaker performance.

The severity of the translation problem will be illustrated by an example. A broad phonetic transcription of the sequence "won't you" by a rather precise talker of general American English (GAE) might look like [wount ju], if the talker paused between words. When spoken in normal continuous fashion, the transcription could be [wountʃu]. Already several difficulties are encountered. These include the following: 1) word boundaries may not be apparent, and 2) phonetic fusion phenomena often alter the pronunciation of words from that expected on the basis of a single word production ([j] → [tʃ]). For a somewhat careless talker of GAE the phoneme string might look like [wɔnʃə], illustrating plausible: 1) ideolectical variation [ɔ] for [oʊ], 2) vowel reduction [ə] for [u], and 3) talker-dependent omission [t]. A more extreme example of speaker-dependent corruption is [dId ju it] → [dž it]. Such transcription variability is well described in the literature, e.g., Heffner [1].

Thus far it has been assumed that the transcription represented precisely what the talker produced. However, any automatic procedure for obtaining phonetic transcriptions will certainly produce errors. For the preceding transcriptions, the errors resulting from machine segmentation and classification can be considered to be of three types: substitution, omission, and insertion. For example, adding considerable machine noise to the last transcription might produce something like [wʌzət], where [ʌ] is substituted for [ɔ], [n] is omitted, [z] is substituted for [ʃ], and [t] is inserted. Certainly, decoding of such transcription is no easy task. The last transcription given would more plausibly have come from "was it" than from "won't you." The general goal of this work is the investigation of those techniques which appear to

Manuscript received April 30, 1972.

C. C. Tappert and N. R. Dixon were with the IBM Corporation, Research Triangle Park, N.C. 27709. They are now with the IBM T. J. Watson Research Center, Yorktown Heights, N.Y. 10598.

A. S. Rabinowitz was with the IBM Corporation, Research Triangle Park, N.C. 27709. He is now with the Research and Technology Division, North American Rockwell Electronics Group, Anaheim, Calif.