New Techniques for Automatic Speaker Verification

AARON E. ROSENBERG, MEMBER, IEEE, AND MARVIN R. SAMBUR

***Abstract*—An interactive automatic speaker verification system has been augmented to include linear prediction parameters in addition to the already existing pitch and intensity analysis of sentence-long utterances. This improved system has been evaluated on a new and enlarged speaker population. A method for selecting optimum speaker-dependent features has been incorporated in this system which significantly improves its performance. The evaluation indicates that verification error rate is approximately 1 percent with respect to casual impostors and 4 percent with respect to well-trained mimics.**

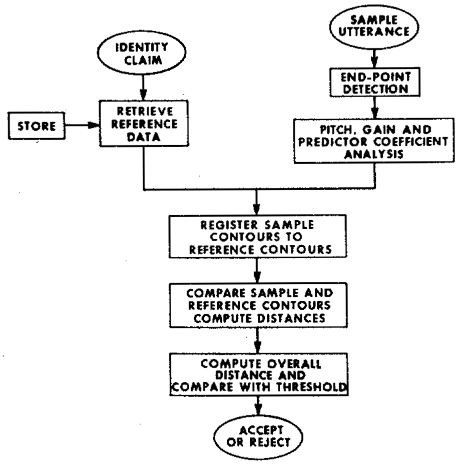
IDENTITY CLAIM
SAMPLE UTTERANCE
STORE
RETRIEVE REFERENCE DATA
END-POINT DETECTION
PITCH, GAIN AND PREDICTOR COEFFICIENT ANALYSIS
REGISTER SAMPLE CONTOURS TO REFERENCE CONTOURS
COMPARE SAMPLE AND REFERENCE CONTOURS COMPUTE DISTANCES
COMPUTE OVERALL DISTANCE AND COMPARE WITH THRESHOLD
ACCEPT OR REJECT
AUTOMATIC SPEAKER VERIFICATION OPERATIONS

Fig. 1. Block diagram showing principal automatic speaker verification operations.

INTRODUCTION

A SYSTEM for automatic speaker verification has been under investigation at Bell Laboratories [1]–[6].

In a speaker verification system, the object is to accept or reject the identity claim of a speaker by comparing a set of measurements of the speaker's utterances with a reference set of measurements of the utterance of the person whose identity is claimed.

The present system is based on an acoustic analysis of fixed, sentence-long utterances which results in a set of analyzed functions plotted versus time through the utterance. The features selected for analysis are expected to provide characteristic and distinctive information over a population of speakers and to be tolerant of the variation in spectral transmission characteristics to be expected in dialed-up telephone lines.

A block diagram indicating the principal operations of the system is shown in Fig. 1. There are two inputs to the system, the identity claim and the sample utterance. The identity claim, which may be provided by a keyed-in identification number, causes reference data corresponding to the claim to be retrieved. The second input is activated by a request to speak the sample utterance. The recording interval is scanned to find the end-points of the utterance. The utterance is then analyzed. The analysis in the present implementation provides pitch period (600 Hz low-pass), gain or intensity, and two predictor coefficients. The data points plotted as a function of time for each analysis feature are referred to as a contour. A crucial property of the system is automatic time registration of the contours of the sample utterance to the contours retrieved as part of the reference data of the claimed identity. Following registration, a set of measurements is applied to the sample contours and compared with the same set of measurements applied to the reference contours. Distances are calculated which quantify the dissimilarity between the sample and reference contour measurements. Finally, the distances are combined into an overall distance which is compared with a threshold distance value to determine whether the identity claim should be accepted or rejected.

Details concerning the analysis procedures, reference construction, contour measurements, and distance calculations can be found in Lummis' paper [6].

The differences between the present implementation and the previous implementation of the system lie in the features selected for analysis and the method of overall distance computation. In addition, a new and enlarged sample of speech utterances has been used for evaluations. The previous implementation and evaluations will be reviewed briefly to provide a basis for the present modifications.

The first implementation was carried out in a large-scale Honeywell computer [1], [2]. The analysis used was essentially the formant analysis devised by Schafer and Rabiner [7] using FFT cepstral techniques. This analysis generates five features, the first three formants, $F1$, $F2$, $F3$, pitch period, and gain. The system was evaluated using sample utterances recorded from a population of 40 male speakers. The utterance recorded was the all voiced utterance "We were away a year ago." Eight of the speakers were designated "customers," the remainder were "impostors." These were "casual impostors" in that they spoke the test sentence with no attempt to imitate customer voices. Performance was specified in terms of the equal-error criterion. That is, the decision threshold was adjusted for each customer so that the rate of customer rejection (false-alarm rate) is equal to the rate of im-

Manuscript received June 21, 1974.

The authors are with Bell Laboratories, Murray Hill, N. J. 07974.

postor acceptance (miss rate). The average equal-error rate found by Doddington [1] was approximately 1.5 percent. Doddington reported that the distances based on formant measurements contributed relatively little to the system performance. Moreover, a large amount of processing time is consumed in the computation of formant data—approximately 200–300 times the duration of the utterance.

Lummis introduced modifications of the system which permitted the elimination of formant contours from the analysis [6]. (Specifically, time registration formerly controlled by the $F2$ contour was now controlled by the gain contour. To facilitate registration contours were subjected to 16 Hz low-pass filtering to smooth out irregularities.) Lummis evaluated system performance using the same 40-speaker sample of recorded utterances. His results confirmed Doddington's observations in that the average equal-error rate was 1 percent when all analysis data was included in the distance computations and increased slightly to 1.2 percent when formant data was omitted. The omission did not appear to lead to a significant degradation.

However, in a separate study [5], Lummis and Rosenberg showed that formant data may be significant with respect to the class of impostors that deliberately attempt to imitate customer utterances. Four professional mimics were retained to imitate the sample utterances of each of the eight customers in the original test population. After intensive training sessions these mimic utterances were recorded, analyzed, and compared with the respective customer references. When all analysis data were included, 27 percent of these utterances were accepted. Since the threshold distances remained unchanged from the evaluation with casual impostors, the customer rejection rate was still 1 percent. With the omission of formant data, the mimic acceptance rate increased to 41 percent. The experiment demonstrated that additional sophistication is required to provide a reasonably mimic-resistant system. Moreover, formant analysis data may provide features less susceptible to mimicking than those provided by pitch and gain alone.

Experience with the implementation of Doddington and Lummis indicates the desirability of omitting formant analysis. On the other hand, additional sophistication is required to maintain acceptable error rates with respect to casual impostors and reduce the rate of acceptance of mimic utterances. A second implementation of the system was initiated to explore the possibility of achieving this.

The second implementation of the automatic speaker verification system resides in a Honeywell DDP-516 laboratory computer. This implementation was initiated by Lummis [4] and completed by the present authors. It is the principal subject matter of this paper. Transferring the system to the laboratory computer provides a more convenient interactive access to the system with on-line graphical displays and printouts. It then becomes possible, for example, to record an utterance directly into the system, display its analyzed contours and compare them with the contours of stored references.

Three goals were set for this laboratory computer implementation. First, it was desired to investigate new features for analysis to supplement pitch and gain analysis and replace formant analysis. Requirements for these new features are fast and efficient computation and improvement of system performance particularly with respect to mimic utterances. Second, a more sophisticated overall distance computation was desirable which would discriminate more effectively individual customers from all other speakers. Third, the collection of a new and larger sample of speech utterances was required with which to evaluate the system to provide a more statistically reliable evaluation of its performance.

PREDICTOR COEFFICIENT ANALYSIS

The first goal set for this implementation was satisfied by means of predictor coefficient analysis. A predictor coefficient technique which Sambur applied to speaker recognition studies [8] was restructured to conform to the existing analysis techniques. In the predictive coding approach to the analysis and synthesis of a speech waveform, it is assumed that the combined effects of the glottal source, the vocal tract, and the radiation characteristic can be represented by an all-pole digital filter [9]. The input excitation in this model is assumed to be a periodic impulse train for voiced speech and random noise for unvoiced speech. These assumptions imply that the output sequence for voiced speech is given by

$$\hat{x}(n) = \sum_{k=1}^{M} -a_k\hat{x}(n-k) + \sum_{l} \delta(n-lT)$$

where $\delta(n)$ is the unit impulse response, M equals the number of poles, and T equals the period of the excitation. The a_k's are the coefficients characterizing the filter and are referred to as the predictor coefficients.

The method of predictive coding computes the most suitable all-pole representation of the speech process by a minimum mean-squared technique. This method takes advantage of the fact that the modeled speech sequence is, except at the beginning of each period, linearly predictable in terms of the past M speech samples. The a_k's are then chosen in such a way that the mean-squared difference between the predicted sequence, $\hat{x}(n)$, and the actual speech, $x(n)$, is minimized.[1]

The z transform description of the modeled digital filter

$$H(z) = 1/1 + \sum_{k=1}^{M} a_k z^{-k}$$

is employed to determine the resonance structure used to represent the combined effects of the vocal tract and glottal source. However, some decision logic is sometimes necessary to determine which poles account for the vocal tract contribution (formants) and which poles can be used to approximate the glottal source.

[1] The method used in this paper for calculating the predictor coefficients is referred to as the covariance method [9].

Predictive Coding and Speaker Verification

In the previous study cited [8] it was shown that predictive coding was a great aid in providing a fast and reliable means of measuring speaker sensitive features of the speech waveform. In that study, predictive coding parameters were used to determine steady-state formant and bandwidth data during the production of an ensemble of speech events. Since predictor coefficients specify the resonance structure of the analyzed speech frame, it is natural to ask whether the conversion from coefficients to formants and bandwidths is necessary to obtain speaker characterizing features. Although formant and bandwidth data are intuitively appealing, there is no *a priori* reason to believe that these data should characterize a speaker any better than a set of predictor parameters. Moreover, formant and bandwidth data are obtained from the coefficients by a time consuming, nonerror free polynomial root finding process. (The root finding process requires almost the same amount of computation time as finding the coefficients alone.) In addition, a nonerror free decision logic program is required to correctly label the appropriate formants and glottal "poles."

To determine the speaker verification effectiveness of predictor coefficients as opposed to that of formant and bandwidth data, a verification experiment was conducted. This experiment was carried out using the available system [6] with first predictor coefficients as specimen contours and then formants and bandwidths as specimen contours.

The evaluation was initially performed for a 12-predictor coefficient analysis of the utterance "We were away a year ago" for four speakers selected at random from the set of 22 speakers. The coefficients were computed in the same fashion as Atal and Hanauer [9]. The speech was low passed at 4000 Hz and then sampled at 10 kHz. The predictive coding analysis was performed over consecutive frames of 100 digital samples. Because of a desire to limit the storage requirements of a verification system, the testing was confined to comparing the error rates of three contours from the ensemble of 12 possible coefficient contours to the error rates of three contours from the ensemble of formant and bandwidth contours.

Before the actual comparison between coefficient contours and formant and bandwidth contours, the "best"[2] three contours from each of the groups were determined. The error rates for all combinations of three contours from the set $F_1,F_2,F_3,F_4,B_1,B_2,B_3,B_4$ were evaluated, and the results demonstrate that the formant information is significantly more valuable than bandwidth information. For this experiment, F_1,F_2,F_4 were nominally the "best" contours, but the other combinations of formant contours were almost as good.

Before ascertaining the relative merits of the coefficient contours, it was decided to plot the various contours for each speaker for several repetitions of the test utterance. Fig. 2 shows the first eight smoothed predictor contours of a 12th-order analysis for a typical speaker. These plots reveal an unexpected amount of redundancy in the coefficients.

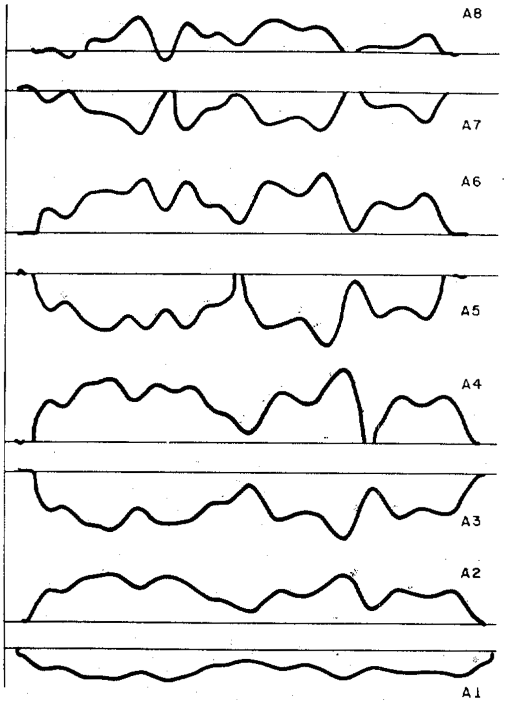

Fig. 2. Eight smoothed predictor coefficient contours for the utterance "We were away a year ago."

A description of this redundancy is given by Sambur [10]. His results show that there is an extremely high negative correlation between adjacent coefficients and that widely spaced predictor coefficients are reasonably uncorrelated with one another. An eigenvalue analysis of the covariance matrix of the predictor coefficients was used by Sambur to further demonstrate the redundant nature of the coefficients by showing that the 12-predictor coefficients can be adequately represented in a space significantly less than 12 dimensions.

The high correlation between adjacent predictor coefficients suggests that all the contours are not needed to obtain good speaker verification results, and that the best strategy for choosing the coefficient contours is to select contours that are widely spaced in numbered ordering (for example, a_2,a_7,a_{12}). The experimental results confirm these observations. The error rate for all 12 contours was not appreciably better than the error rates for the "best" set of three contours. In addition the choice of widely spaced contours was significantly better than choosing three consecutively numbered coefficient contours.

The error rate comparison between the best three formant and bandwidth contours ($F_1\,F_2\,F_4$) and the best three predictor coefficient contours (a_2,a_7,a_{12}) showed that the predictor coefficients were slightly more informative than the formants. Since only four speakers were examined in this experiment, it is unfair to conclude that the predictor coefficients were a "better" source of verification

[2] "Best" is defined as that set of contours that has the lowest average error rate over the four speakers. The error rate is an equal error rate as discussed previously.

parameters than formants. However, in view of the fact that the computation time involved in extracting the formants is almost double the time requirement for coefficients, the predictor coefficient seems a more effective choice for verification parameter.

Order of Linear Predictor Analysis

For formant analysis, it is necessary to use at least 12-predictor coefficients to obtain accurate results. However, speech synthesis studies have shown that the quality of the speech is only slightly degraded when the number of predictors specified is reduced to approximately eight coefficients [9]. In addition, it has also been observed that the overall shape of a speech spectrum specified by a set of predictor coefficients is not noticeably perturbed when a lesser order analysis is used [11]. Since the computation time varies in an approximate linear fashion to the order of the predictor analysis [9], it is worthwhile to investigate the error rate dependence on this order.

To test the effects of changes in the order of the predictor analysis on verification error rates, the experiment discussed above was repeated for an eighth order analysis. The error rate determination for all combinations of three contours from the eight possible contours again showed that widely-spaced coefficients in number yielded the lowest error rates. The error rate for the best contours from the 12th-order analysis was not significantly better than the error rate for the best contours from the 8th-order analysis. This result indicates that very little verification potential is lost by using only an eighth order analysis.

Number of Frames Per Second

Since the predictor coefficients plots showed that the variation rate with time of the coefficients was similar to that of formants (about 16 Hz), it was decided to test whether the verification potential of the predictors would be severely compromised by calculating the coefficients once every 200 points instead of every 100 points. (In both cases the coefficients are computed over a 100-point frame and the sampling rate is fixed at 10 000 points per second.) For an eighth-order analysis, the error was not diminished by calculating the coefficients for every other 100 point frame. Thus the effective storage requirements and the computation time can be cut in half without significantly reducing the verification error rates.

FINAL SELECTION OF FEATURES

The preliminary analysis indicated that the best verification system using predictor coefficients was one that employed only a few widely spaced predictor coefficients. An eighth-order analysis with a frame length of 10 ms and a 10 ms interval between frames was used. This system offers an optimum compromise between storage requirements, error performance and computation time. The processing time required is approximately 11 times real time.

The error performance results and a statistical analysis of these results showed that the coefficients a_2, a_4, and a_8 provided the most effective contours. For the final selection of features, the number of predictor coefficient contours was further reduced from three to two, a_4 and a_8, yielding a considerable saving in storage requirements with no significant degradation in performance. These predictor coefficient contours were added to the pitch and gain contours to complete the feature analysis. A typical set of contours is shown in Fig. 3.

In the original implementation pitch analysis was a product of the overall formant analysis. With the omission of formant analysis it is necessary to provide a new method of pitch extraction. The technique selected is that of Gold and Rabiner [13]. It provides an effective and efficient pitch analysis operating on the speech signal in the time domain. The analysis has been modified to provide gain as a by-product. The processing time required for pitch and gain analysis (including digital filtering of the input signal) is 25–30 times the duration of the input utterance.

OVERALL DISTANCE COMPUTATION

The second specified goal was to provide a more effective overall distance computation. As mentioned in the Introduction the verification process consists of a feature analysis of a test utterance followed by time registration of the resulting feature contours to the reference contours of a speaker whose identity is claimed. Then dissimilarity measures or distances are extracted which quantify the differences between the specimen and reference contours. A detailed description of the type of distance computation carried out is given in Lummis' paper [6]. The simplest type of overall distance computation is a simple average of the individual distances. That is, if comparison between the reference and specimen contours results in a set of N individual distances $d_1,d_2,\cdots,d_n$, then the overall distance D is given by

$$D = 1/N \sum_{n=1}^{N} d_n.$$

If D is less than some threshold distance D_T then the decision is made to accept the identity claim. Otherwise the claim is rejected. Since, the selection of contour measurements and corresponding distances was based to a large extent on intuition the selection is not expected to be uniformly optimal. Therefore it is desirable to find a set of weights $w_1,w_2,\cdots,w_n$ such that

$$D = \sum_{n=1}^{N} w_n d_n$$

which best desciminate the class of customer utterances from the class of utterances spoken by all other people.

In general, the set of weights $\{w_n\}$ and threshold, D_T, will be different for each customer using the system. In order to determine the weights and the threshold it is necessary to have a training set (design set, learning set) of distances from both customer and impostor utterances.

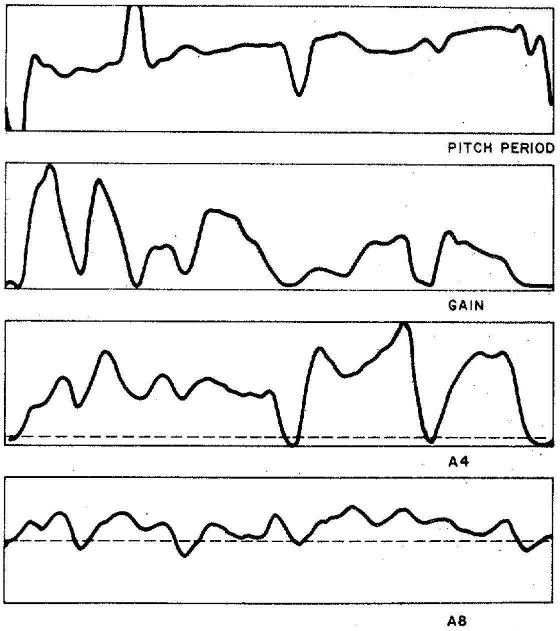

Fig. 3. An example of a set of smoothed contours showing the four features selected for analysis.

If the probability distribution functions underlying the distances obtained from both customer and impostor utterances are known it may be possible to specify functional forms for calculating the weights which optimally discriminate between customer and impostor utterances. Such functions are known as discriminant functions. For example, if it is known that the underlying distributions of the distances are Gaussian then it can be shown that optimal weights are a quadratic function of the component distances in the training set. The computations require calculation of the sample means and covariances of the distances in the training sets. An investigation has shown that the use of such quadratic discriminant functions can effectively reduce error rates, but are somewhat impractical. The impracticality lies in the sample size required to obtain reliable estimates of covariance matrices. This difficulty is related to the "curse of dimensionality" found in many pattern-recognition problems [12].

A simpler, completely nonparametric, procedure starts with the selection of a measure which quantifies how effectively the set of customer distances are discriminated from the set of impostor distances in the training set for any given set of weights $\{w_n\}$. Then with suitable constraints on the range of $\{w_n\}$, a systematic search through this range is carried out to find a set of weights which maximizes the effectiveness. A simple measurement of effectiveness is simply the fraction of correct decisions resulting from a given set of weights. A simplifying constraint on the range of $\{w_n\}$ is to allow them to assume only the values 0 or 1. Each set of $\{w_n\}$ then imposes a selection of a subset of distance components from the original set. For this reason the overall distance computation becomes a distance selection. Finally, a systematic procedure must be adopted to search for the most effective selection. If there are N distance components in the original set there are the binomial factor $\binom{N}{M}$ possible distance subsets with M components. For pitch and gain alone N is 13; with predictor coefficients N becomes 23. For such values of N an exhaustive search for every possible value of M is impractical. Some restricted search procedures were investigated which fix the size of the selected distance subset to say 5 components. However, the procedure we finally selected is a "knock-out" technique which allows the size of the final subset to be a variable [8]. Effectiveness was defined as the empirical equal-error probability.

"Knock-Out" Strategy

The "knock-out" technique begins by evaluating the effectiveness of each of the N distance-component subsets of order $N - 1$. The component not included in the most effective subset is eliminated or "knocked-out" from further consideration. The procedure continues by considering the effectiveness of $N - 1$ distance-component subsets of order $N - 2$ of the most effective subset of order $N - 1$. The iteration proceeds through the evaluation of a single-component subset. In a typical evaluation effectiveness increases monotonically with each successive iteration until a maximum is obtained for a subset of order M, $1 < M < N$, after which effectiveness monotonically decreases. It is this subset of order M which is chosen as the most effective subset of distance components for the specified customer.

The improvement in error rate obtained by use of this simple distance selection procedure over the use of an average of the whole set of distances will be shown in the results.

SAMPLE OF TEST UTTERANCES

The third specified goal of this implementation of the automatic speaker verification system was to increase the reliability of the estimate of system performance by obtaining a larger sample of test utterances.

To this end, 22 native American male speakers were recruited to provide a series of 50 recordings of a series of six sentences. The recordings were made in a sound booth using an electret microphone inserted in a telephone hand-set connected to an Ampex AG-350 tape recorder. The recordings were passed through a 4000 Hz low-pass filter with a 48 dB/octave roll-off, sampled at a 10 kHz rate, and digitized prior to analysis. At most two recording sessions per customer per day were held, one in the morning and one in the afternoon. The recordings were made over a period of approximately two months. In addition to the customer recordings, 55 additional native American male speakers provided one recording session each. These recordings were designated impostor utterances. Of the six sentences recorded, three were all-voiced and three both voiced and unvoiced. Just two sentences were digitized and analyzed for experimentation. These were the all-voiced sentences "We were away a year ago" (used in the previous evaluation) and "I know when my lawyer is due." It will be seen in the results that the system performance improves considerably when two sentences are tested instead of one.

RESULTS

As indicated in the Introduction, the performance of the system has been specified by calculating the mean over all customer references of the estimated equal-error rates. For each customer reference the overall distance to the reference of each of a set of test customer utterances and a set of test impostor utterances is calculated. A plot is constructed which shows the percentage of test customer overall distances greater than the overall distance variable plotted along the abscissa and another which shows the percentage of impostor overall distances which are less than this same overall distance variable. These plots are empirical cumulative distribution functions with respect to the overall distance variable. The intersection of these plots specifies an estimate of the "equal-error" rate, E_T, together with an accept/reject threshold, D_T. An example is shown in Fig. 4. For each of the 22 customers in the sample, two references were constructed from two distinct sets of 10 utterances. These 10 utterances were selected uniformly over the entire span of 50 utterances for each customer. The remaining 40 utterances in each case were the designated customer test utterances. Also in each case there were 55 impostor test utterances. For each customer reference then the estimate of equal-error rate is determined from the overall distances of 95 utterances. Since there are 22 customers and 2 references per customer the mean "equal-error" rate is the average of 44 individual "equal-error" rates. Two sentences were processed and results are shown for each sentence individually and for the two sentences combined. The first sentence, sentence A, is the all-voiced utterance "We were away a year ago." This is the same sentence used in evaluations of previous implementations of the speaker verification system. The second sentence, sentence B, is another all-voiced utterance but includes nasals, "I know when my lawyer is due."

Table I summarizes the results of the evaluation. The results are presented for pitch and gain measurements alone and for pitch and gain measurements supplemented by measurements from 2 predictor coefficient contours. The equal-error rate obtained with pitch and gain measurements of sentence A is 5.6 percent. The very same measurements applied to the 8 customer, 32 impostor set of recordings used in evaluations of previous implementations (denoted "old sample") results in an equal-error rate of 2.2 percent. This value is to be compared with the equal-error rate of 1.2 percent obtained by Lummis with comparable measurements and the same speaker set. Considering the transfer from a large-scale computer to a laboratory computer and some programming variations these figures are in basic agreement. There is however a considerable discrepancy in error rate between the new, 77-speaker set and the old 32 speaker set. We believe the newer estimate to be more representative, if only because of the larger number of speakers and the larger number of utterances per customer, 50. For a second sentence, sentence B, "I know when my lawyer is due" the error rate is the same order of magnitude, 7.2 percent. By calculating a combined overall distance for both sentences, the average of the distance components for both sentences, the error rate falls to 3.9 percent. This indicates that the inclusion of a second sentence for verification provides a considerable amount of information that is independent of the first sentence.

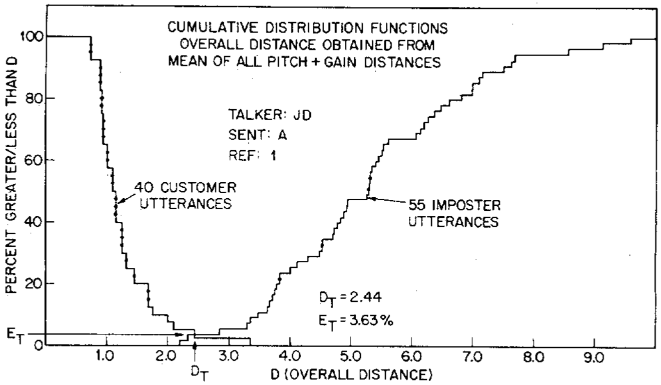

Fig. 4. Empirical cumulative distribution functions of the overall distances obtained from customer utterances and impostor utterances compared with a particular reference. The intersection of the plots determines the equal-error estimate and corresponding overall distance threshold.

TABLE I
Error Rates

		PITCH + GAIN	PITCH + GAIN + 2 PREDICTOR COEFFICIENTS
ALL DISTANCES	SENT. A	5.6	4.0
	SENT. B	7.2	4.2
	SENT. A + SENT. B	3.9	2.5
SELECTED DISTANCES	SENT. A	2.5	0.9
	SENT. B	3.3	1.7
	CHOICE OF SENT. WITH LEAST ERROR	1.6	0.4
ALL DISTANCES	OLD SAMPLE (SENT. A)	2.2	1.8
	MIMIC ACCEPTANCE	29.5	21.5
SELECTED DISTANCES	OLD SAMPLE (SENT. A)	1.0	0.
	MIMIC ACCEPTANCE	15.6	4.1

Similar results are observed when predictor coefficient measurements are included. In each case the error rate is less than the corresponding error rate with pitch and gain measurements alone. For the case of the two sentences combined the error rate is slightly more than half the error rate with pitch and gain measurements alone.

Improvements in performance are even greater with the use of the distance selection algorithm which results in speaker-dependent measurements. As described previously, a subset of the available measurements is selected which best discriminates the utterances of each customer from the impostor set of utterances. The criterion for best discrimination is least error rate. The resulting overall error rates are in each case less than half the corresponding error rates for the combination of distances from all measurements. The performance figure given for the combination of both sentences is obtained by averaging the smallest of the two error rates for each customer and reference. This is equivalent to an additional selection process in which the sentence providing the least error rate is assigned to each customer. The result is again an error rate which is approximately half the error rates for the sentences considered separately. Histograms showing the distribution of individual error rates for four of the conditions listed in Table I are shown in Fig. 5.

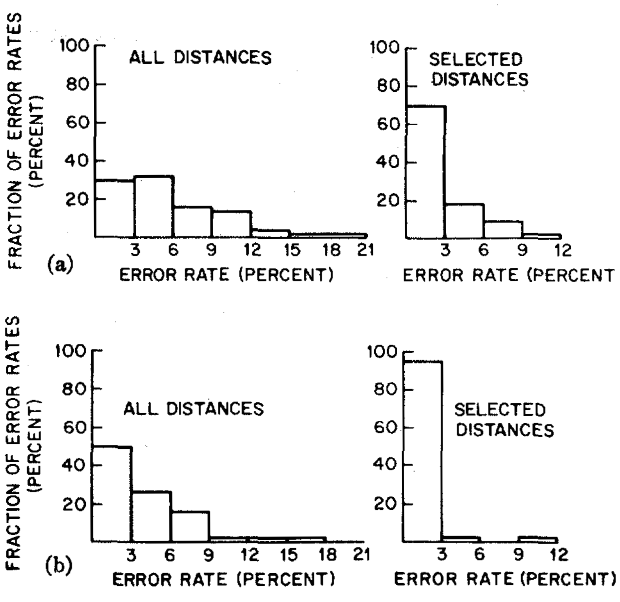

Fig. 5. Histograms showing the distribution of individual error rates for sentence A. (a) The distribution for pitch and gain measurements alone. (b) The distribution for pitch and gain and 2 predictor coefficients.

Finally, consideration is given to a reprocessing of the recordings of four professional mimics who provided imitations of each of the eight "customers" in the original speaker set. Mimic performance is given relative to the rate of acceptance of their utterances in terms of the equal-error accept/reject thresholds obtained for the original customer-casual impostor population. For example, for pitch and gain measurements alone, combining all distances for each customer, mimic utterances are accepted at a rate of 29.5 percent. The corresponding equal-error rate of 2.2 percent is the customer rejection rate associated with this rate of mimic acceptance. With the addition of predictor coefficient measurements, reduction of mimic acceptance to 21.5 percent is obtained with the equal-error rate reduced to 1.8 percent.

A possibility for further reducing the mimic acceptance rate is to base equal-error accept/reject threshold on the population of customers and mimics rather than customers and casual impostors. This will have the effect of reducing mimic acceptance at the expense of increasing the rejection rate of customer utterances. The use of selected distances for each customer not only reduces the equal-error rates but provides significant reductions in the mimic acceptance rate. With pitch and gain measurements mimic acceptance is reduced from 29.5 to 15.6 percent. With pitch, gain, and predictor coefficient measurements mimic acceptance is reduced five-fold from 21.5 to 4.1 percent with 0 equal-error. It should be noted again that the distance selection algorithm is based on the population of customers and casual impostors. A further reduction can be expected if mimics are substituted for casual impostors.

CONCLUSION

An improved system for automatic speaker verification has been implemented on a laboratory computer. The improvements over previous implementations include the addition of predictor coefficient measurements to the existing pitch and gain analysis and the use of speaker-dependent distances at the decision stage of the process. In addition, a larger data base in terms of the number of speakers, the number of utterances and the number of utterance replications has been used to evaluate the performance of the system. The resulting average equal-error rate is approximately 1 percent for a single utterance. This error can be reduced to approximately 0.5 percent with the use of two utterances. The corresponding rate of mimic acceptance for a single utterance is approximately 4 percent.

Further investigations, current or projected, include analysis of female voices, analysis over telephone lines, a large-scale evaluation over telephone lines permitting direct customer access and on-line response, and specialized hardware processing to improve response times.

ACKNOWLEDGMENT

The authors wish to express their appreciation to Judy Johnson, Maura Johnson, and B. Donnelly for their help in collecting and processing the speech samples, to D. Bock for setting up and maintaining the equipment, and to L. Rabiner for assistance in implementing the pitch detector.

REFERENCES

[1] G. R. Doddington, "A computer method of speaker verification," Ph.D. dissertation, Dep. Elec. Eng., Univ. Wisconsin, 1970.

[2] ——, "A method of speaker verification," *J. Acoust. Soc. Amer.*, vol. 49, p. 139(A), 1971.

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

[4] ——, "Implementation of an on-line speaker verification scheme," *J. Acoust. Soc. Amer.*, vol. 52, p. 181(A), 1972.

[5] R. C. Lummis and A. E. Rosenberg, "Test of an automatic speaker verification method with intensively trained mimics," *J. Acoust. Soc. Amer.*, vol. 51, p. 131(A), 1972.

[6] R. C. Lummis, "Speaker verification by computer using speech intensity for temporal registration," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 80–89, Apr. 1973.

[7] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634–648, 1970.

[8] M. R. Sambur, "Speaker recognition and verification using linear prediction analysis," Ph.D. dissertation, Dep. Elec. Eng., Mass. Inst. Technol., Sept. 1972; also M.I.T. Res. Lab., Electron. Quart. Prog. Rep. 108, pp. 261–268, Jan. 1973.

[9] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.* vol. 50, pp. 637–655, 1971.

[10] M. R. Sambur, "An efficient LPC vocoder," in preparation.
[11] J. D. Markel, *Formant Trajectory Estimation from a Linear Least-Squares Inverse Filter Formulation* (SCRL Monograph No. 7). Speech Communication Res. Lab., Santa Barbara, Calif., Oct. 1971.
[12] W. S. Meisel, *Computer-Oriented Approaches To Pattern Recognition.* New York: Academic, 1972, pp. 13–15.
[13] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 442–448, 1969.

Selection of Acoustic Features for Speaker Identification

MARVIN R. SAMBUR

Abstract—The aim of this study was to determine a set of acoustic features in the speech signal that are effective for the identification of a speaker. The investigation examined a large number of theoretically attractive features. The analysis technique of linear prediction was incorporated to examine features that were previously ignored because their measurement was either too time consuming or not easily amenable to automatic measurement.

A novel probability of error criterion was used to determine the the relative merits of the features. The experimental data base was collected over a $3\frac{1}{2}$ year period and afforded the oportunity to investigate the variation over time of the measurements. The measurements that were found to be the most important were the value of the second resonance (around 1000 Hz) in /n/, the value of the third or fourth resonance (1700–2000 Hz) in /m/ the values of the second, third and fourth formant frequencies in vowels, and the average fundamental frequency of the speaker.

A speaker identification experiment using only the best five features was performed. The test data consisted of the multisession data of 11 speakers, and the test data was kept independent of the design data. One error was made in the identification of these speakers for 320 separate identification experiments.

I. INTRODUCTION

A CRUCIAL ingredient in the success of any pattern recognition system is the selection of features that efficiently characterize the patterns of interest. This paper reports on a study [1] undertaken to determine a set of acoustic features in the speech signal that are effective for the identification of a speaker.

The investigation was conducted by first determining an initial set of acoustic parameters which, on the basis of theoretical considerations and past experimental work [2]–[4], might be suitable candidates for indicating the unique properties of a speaker's vocal apparatus, as well as some aspects of his learned pattern of speaking. The initial selection of features was also made to take advantage of the speech-analysis technique of linear prediction [5]. This analysis technique provided a quick and convenient measurement of many theoretically important speaker characterizing properties that have not been incorporated in recognition schemes because of the inefficient methods available for their measurement. These parameters include formant bandwidths, glottal source "poles," and the pole locations during the production of nasals and strident consonants. The initial list of speaker characterizing features also included the formant structure of vowels, the duration of certain speech events, the dynamic behavior of the formant contours, and various aspects of the pitch contour throughout an utterance. In all, a total of 92 features was examined in the study.

To determine the relative merits of the features, a novel probability of error approach was devised. This method evaluates the features in accordance with their relative contribution to the performance of a given speaker recognition scheme. The experimental data used in the evaluation were collected over a $3\frac{1}{2}$ year period and afforded the opportunity to investigate the variation over time of the features. The goal of this evaluation was an ordered list of speaker characterizing measurements that could guide an individual in selecting features to incorporate in a speaker recognition system. Before considering the merits of the features investigated, we shall discuss the method used in this paper to evaluate the measurements.

II. FEATURE EVALUATION

The problems associated with evaluating the relative effectiveness of a set of features can be best understood within the concept of a feature space. If N features are to be measured in the recorded speech of a talker, every replication of the experimental data by the speaker can be represented as a point in what is termed an N dimensional feature space. Fig. 1 illustrates an example of a two dimensional feature space representation of the measured data points for a number of talkers. The statistical nature of the ensemble of measured points for each individual can be considered to be governed by some underlining multidimensional probability distribution that is hopefully different for each speaker. The ability of a set of measure-

Manuscript received June 10, 1974.
The author is with the Bell Laboratories, Murray Hill, N. J. 07974.