An Investigation Of Speaker Verification Accuracy Using Fundamental Frequency And Duration As Distinguishing Features

Robert M. Ward* John N. Gowdy

Electrical & Computer Engineering Department, Clemson University

ABSTRACT

A method of automatically verifying a person's identity based upon measurements of features taken from a sample of speech is presented. The design goal was to develop a simple, efficiently implemented system, compared with existing systems, while retaining as much accuracy as possible. Parameters extracted from the utterance being analyzed are fundamental frequency of vocal cord vibration and duration of voicing. Mahalanobis and Euclidean distances are used to compare the sets of features. An accuracy of 100% was obtained for one individual with the Mahalanobis distance measure; several other individuals had performance accuracies over 90%. On the whole, accuracy rates achieved in the simulation were about 80%.

INTRODUCTION

Speaker recognition refers to the task of identifying a person based on some aspect or aspects of his voice. Machines have performed this task well in several studies [1-7]. There are two categories of speaker recognition by machine: automatic speaker identification (ASI) and automatic speaker verification (ASV). In ASI, a machine classifies a voice as matching one of many stored in its database. In ASV, the machine must make a single decision as to whether the speaker is who he claims to be or an impostor. A recognition system can be text- independent or text-dependent.

In ASV studies, a machine is first trained with certain distinguishing features extracted from the voices of people it will later recognize. Papers by Doddington [8], Rosenberg [9], and Atal [10] present reviews of research in this area. Features are stored as elements of n-dimensional vectors. To test the system, a speaker claims an identity and utters a word or phrase. The utterance is analyzed, and the resulting vector is compared with the reference vector for the claimed identity. If the distance between the two vectors is within a certain threshold, the speaker is verified. If not, he is rejected as an imposter.

This paper presents the results of a study of text-dependent ASV. The study tested the accuracy of a simple speaker verification system that employs measures of fundamental frequency of vocal cord vibration (or its inverse, pitch period) and duration as distinguishing features.

BACKGROUND MATERIAL

The vocal tract can be viewed as an acoustical tube with the glottis at one end and the lips at the other. This tube can be modelled as a filter which acts on an input excitation and produces an output speech wave. Unvoiced sounds result from aperiodic, noiselike excitations of the vocal tract at points of constriction. Plosive sounds are produced when the vocal tract becomes completely blocked at some point. Pressure builds up behind the blockage, and when it is abruptly released, a burst of sound is produced. Voiced sounds such as vowels result when air expelled from the lungs is alternately blocked and released by the rapid opening and closing of the vocal cords. Speech produced is periodic with a frequency of oscillation corresponding to the fundamental frequency, the rate at which the vocal cords vibrate. This frequency is determined by the size of the vocal cords, the tension with which they are held, and the pressure of the air leaving the lungs [10]. Average values for adult males and females are 132 Hz and 223 Hz, respectively [12]. Average pitch values can be used as features to separate males from females [12] ; however, an average pitch value over an utterance may not be enough to discriminate well between individuals within these groups. Instead of averaging pitch measurements into a single value, several measurements can be recorded as a function of time in an utterance. Successful recognition has been achieved in this manner [1].

In a simplified model of the speech process, one of two inputs excites a vocal tract filter with an impulse response which includes the effects of the vocal cords, vocal tract, and radiation [11]. Voiced speech results when the input is a periodic pulse train. Unvoiced

CH2714-4/89/0000/0390$01.00 © 1989 IEEE

speech results from exciting the filter with noise. Movement of articulators during speech production is modelled by changing the frequency response of the filter [12]. For most sounds, vocal tract shape changes slowly with time [10] allowing output speech to be represented over a short interval as the convolution of the input excitation with the vocal tract impulse response. In the frequency domain

$$X(\Omega) = E(\Omega)V(\Omega) \quad (1)$$

where $X(\Omega)$, $E(\Omega)$, and $V(\Omega)$ are the Fourier transforms of the output speech, excitation signal, and vocal tract response, respectively.

Cepstral analysis is a method of measuring fundamental frequency which takes advantage of the simplified speech production model. In order to extract information about fundamental frequency from the spectrum $X(\Omega)$, the effects of the excitation and vocal tract must be separated. The product in equation 1 can be transformed into a sum by taking a logarithm,

$$\log [X(\Omega)] = \log [E(\Omega)] + \log [V(\Omega)]. \quad (2)$$

The log spectrum thus obtained can then be analyzed spectrally by performing an inverse Fourier transform, resulting in the cepstrum,

$$\hat{x}(t) = \hat{e}(t) + \hat{v}(t) \quad (3)$$

where

$$\hat{x}(t) = F^{-1} \{\log [X(\Omega)]\},$$

$$\hat{e}(t) = F^{-1} \{\log[E(\Omega)]\},$$

$$\hat{v}(t) = F^{-1} \{\log[V(\Omega)]\}.$$

The independent variable in the cepstrum is called quefrency and has the units of time. Therefore, cepstral analysis generates the amplitudes of the various quefrency components present in the spectrum. While vocal tract effects taper off at low quefrency (3 to 4 ms), components due to periodic excitation generally occur in the range from 4 to 15 ms [12]. The signal x(t) must be high-time windowed beginning at 4 ms. The windowed cepstrum exhibits a sharp peak at a quefrency corresponding to the fundamental frequency, which can be detected by a peak-picking algorithm [13]. A peak between 4 and 15 ms corresponds to a frequency between 66 and 250 Hz. This range includes most fundamental frequencies encountered in normal speech. Cepstra of unvoiced speech do not exhibit pronounced peaks because there is no fundamental frequency present.

Once extracted from speech, features are stored as elements of vectors, one vector per utterance. Each vector describes a point in n-dimensional space. If a set of features contains good discriminators, utterances by the same person will tend to generate a cluster of points in this space and different speakers will have widely separated clusters. A decision rule calculates the distance between a test point and a reference point or points, accepting or rejecting the test utterance as coming from the claimed identity.

Euclidean and Mahalanobis distances have been used in speaker recognition systems. Euclidean distance is described as

$$g_i(\underline{x}) = [(\underline{x}-\underline{\mu}_i)^T(\underline{x}-\underline{\mu}_i)] \quad (4)$$

where $\underline{x}$ is a test vector and $\underline{\mu}_i$ is a mean reference vector for speaker i. If the value of $g_i(\underline{x})$ is below a certain threshold, the speaker is verified to be speaker i. Mahalanobis distance is a similar measure but is weighted by the inverse of a covariance matrix W computed using reference vectors,

$$g_i(\underline{x}) = [(\underline{x}-\underline{\mu}_i)^T\ W^{-1}\ (\underline{x}-\underline{\mu}_i)] \quad (5)$$

By including the covariance matrix in the calculation the Mahalanobis distance takes into account the usefulness of the individual features extracted during training.

The database used in this study was a subset of a database originally produced by Texas Instruments and obtained from the National Bureau of Standards. It consisted of 140 utterances of the word 'stop' spoken by seven male and seven female speakers. This word was chosen over others in the original database primarily because the stop consonants before and after the vowel allow fairly reliable and easy segmentation.

PROCEDURE

The first step in extracting features was to identify the endpoints of the utterance from the background noise. An algorithm proposed by Rabiner and Sambur was modified and used for this task [14]. Examination of the speech waveforms showed that the endpoints selected by this algorithm were almost always correct to within 10 ms. Another similar algorithm chose the endpoint of the voiced interval.

With the boundaries of voicing determined, the waveform is next windowed and analyzed to find fundamental frequency values. A Hamming window of width 40.96 ms was used to window the waveform at the beginning, middle, and end of voicing. A Hamming window was chosen because the side lobes taper off sharply in the frequency domain. This minimizes the effects due to including a non-integral number of pitch periods in the analysis window [15,16]. The window

must be wide enough to include at least two pitch periods for the cepstral analysis to be effective. Since the maximum expected period is 15 ms the window must be at least 30 ms wide. The cepstral analysis routine required that the window width be a power of two. The data was sampled at 12.5 kHz, so the best choice was 512 samples, giving 40.96 ms.

The windowed waveform is no longer simply the convolution of the excitation and the vocal tract response. It is now

$$x'(t) = [e(t) * v(t)]\, w(t). \qquad (6)$$

This problem is accounted for by assuming equation 6 can be approximated by

$$x'(t) = e_1(t) * v(t),$$

where

$$e_1(t) = e(t) * w(t).$$

This equation is valid if the window function varies slowly as compared with the impulse response of the vocal tract [11]. The cepstrum now becomes

$$\hat{x}(t) = \hat{e}_1(t) + \hat{v}(t).$$

Periodicity should be maintained by the windowed pulse train $e_1(t)$.

Cepstral analysis is performed first on a region beginning 5.12 ms after the estimated starting point of the voiced region. The delay allows time for the vocal cords to begin oscillating once the preceding consonant has been articulated. The resulting cepstral data is checked for a peak value between 50 and 188 samples (4 to 15 ms) and the sample number having the highest amplitude is recorded. Values of the cepstrum are then squared to emphasize the peak and are written to a file that can plotted. One such plot is shown in Figure 1. Pitch values are then calculated for an interval centered on the midpoint of the voiced region and one beginning 51.2 ms before the ending point.

The peak-picking algorithm assumes that the cepstrum has only one peak. The pitch could conceivably change within the analysis window, however. If this occurs, there could be more than one peak [13]. The peak with the largest amplitude is chosen to correspond to the dominant pitch period within the interval. For one speaker, a larger peak occurred in the cepstrum at a value that was twice the actual ptich period. The peak-picking algorithm was modified to correct this. If the measured pitch occurred at a sample number higher than 100, an interval of 10 samples centered on half that number was checked. If that interval contained a value within 50% of the peak value, then the sample exhibiting this value was chosen as representing the true pitch period. For utterances in which the true pitch period was at 100 samples or higher, the value at half the number of samples was at least a power of 10 less and was not mistakenly chosen as the pitch period.

With the feature extraction completed, distance calculations between feature vectors were performed. Three types of calculations were made: three dimensional Euclidean distance, four-dimensional Euclidean distance and four-dimensional Mahalanobis distance. Instead of testing utterances one at a time, all of the feature vectors for each individual speaker could be compared with a reference and the distances between them displayed. The number of errors could then be calculated for several different thresholds by scanning the resulting distance values.

For the Euclidean distance calculations, three speakers from the male and the female group (considered separately) were chosen as customers. Vectors corresponding to the first, middle and last repetitions of the word for each customer were chosen as references. These words were spoken the farthest apart in time during the recording interval and would hopefully exhibit more of the speaker's natural variations than utterances spoken in succession. The reference vectors take the place of the means $\underline{\mu}i$ in equation 4. The decision rule implied by this measure assigns an utterance to the claimed identity if its distance from one (or more) of the three reference vectors is below a certain threshold. Each trio of reference vectors was compared to vectors for the seven remaining utterances from the corresponding customer's data, plus files from the other six speakers in the group. Theoretical performance was then calculated for several threshold values. The three-dimensional vector contained the three pitch measurements; the four-dimensional vector also included the measure of duration of voicing.

For the Malalanobis distance calculations, the feature vectors were separated into two groups: a training set and a test set. Five vectors from each of the members of a group were used to calculate mean vectors for each customer and a single covariance matrix for the group. The three pitch measurements and the duration of voicing were used. Mean vectors were calculated by averaging corresponding elements. A matrix was computed for each training utterance, defined by

$$Vij = [\underline{x}ij - \underline{\mu}j]\,[\underline{x}ij - \underline{\mu}j]^T,$$

where $\underline{x}ij$ denotes the i th training utterance from speaker j and $\underline{\mu}j$ denotes the mean vector for speaker j. The resulting thirty-five matrices for each group were

averaged element by element to form the group's covariance matrix. Mahalanobis distances based on equation 5 were calculated between each of the test vectors and each customer's mean vector. The data was then analyzed as with Euclidean distance and the number of errors were computed for several threshold values.

RESULTS

Choosing the threshold that gives the best performance in a speaker verification system is not a simple task. There are two possible sources of error: false rejections and false acceptances. A threshold must optimize performance in some way based on both error rates. Performance of two types of optimal thresholds were examined on an individual and overall basis for both groups. One such threshold examined was the lowest point at which at least half of the customer utterances were accepted. A small inconvenience for customers might be rewarded with fewer impostors being accepted. Another point examined was the lowest threshold at which the two error rates were most equal.

It was found that performance varied widely between individuals; some speakers were more easily distinguished than others. Table 1 gives error rates for male speakers RLD, WMF, and REH for selected threshold values using the three-dimensional Euclidean distance. The table illustrates the tradeoff between the two error rates. For a threshold of five only 0.58% of the impostor utterances were falsely accepted. However, there was also 100% customer rejection at this threshold. The false rejection rate was as low as 14.3% for a threshold of 30, but the corresponding false acceptance rate was almost 70%. Table 2 summarizes the best and worst accuracy rates achieved with each distance measure for the two types of threshold points examined.

Although perfect performance was achieved for an individual speaker at one point, it was the exception rather than the rule. As Table 2 shows, the performance of the male group as a whole was significantly improved by adding the duration feature to the distance calculations. The female group's performance was not, however. The best individual performance was achieved using the Mahalanobis distance [17].

SUMMARY

These results show that perfect recognition performance is possible, at least on an individual basis. Accuracy rates above 90% at rejecting impostors were achieved for several individuals at the thresholds where at least half of the customer utterances were accepted. Good performance was not consistent among all the individuals in a group, however. As a result, overall accuracy rates do not compare well with those reported in previous studies. The fact that performance varied between individuals suggests a limitation of this study. Only 10 utterances from each speaker were included in the database. Many more could be needed to sufficiently characterize the speech of some individuals. The utterances should also be recorded in several sessions separted by at least a day to include more of the nature variations in each individual speaker's voice. Testing the system with a larger group of speakers would provide a better approximation to the expected real-world performance.

In conclusion, the results of attempting to verify a speaker's identity based on only three pitch measurements and a measurement of duration of voicing are not nearly as exceptional as results reported using a larger number of pitch measurements recorded as a function of time in an utterance. The accuracy rates achieved in this study might be good enough for some low risk applications, such as children's toys, but not for those applications for which an error would be more costly, such as in banking.

REFERENCES

[1] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours," J. Acoust. Soc. Amer., vol.52, pp. 1687- 1697,1972.

[2] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," J. Acoust. Soc. Amer., vol. 55, pp. 1304-1312, 1974.

[3] S. K. Das and W. S. Mohn, "A Scheme for Speech Processing in Automatic Speaker Verification," IEEE Trans. Audio Electroacoust., vol. AU-19, pp. 32-43, 1971.

[4] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Trans. Acoust., Speech, Signal Processing vol. ASSP-29, pp. 254-272, 1981.

[5] S. Furui, "Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features," IEEE Trans. Acoust.,Speech, Signal Processing, vol. ASSP-29, pp. 342-350, 1981.

[6] J. E. Luck, "Automatic Speaker Verification Using Cepstral Measurements," J. Acoust. Soc. Amer., vol. 46, pp. 1026-1032,1969.

[7] R. C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans. Audio Electroacoust., pp. 80-89,1973.

[8] G. R. Doddington, "Speaker Recognition-- Identi-

fying People by their Voices," Proc.IEEE, vol.73, pp. 1651-1664, 1985.

[9] A. E. Rosenberg, "Automatic Speaker Verification: A Review," Proc. IEEE, vol. 64, pp. 475-487,1976.

[10] B. S. Atal, "Automatic Recognition of Speakers from their Voices," Proc.IEEE., vol. 64, pp. 460-475, 1976.

[11] W. Verhelst and O. Steenhaut, "On Short-Time Cepstra of Voiced Speech," Proc. ICASSP-88, vol. I, pp. 311-314, 1988.

[12] D. O'Shaughnessy, Speech Communication: Human and Machine, Addison-Wesley, New York, 1987.

[13] A. M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Amer., vol. 41,293-309, 1967.

[14] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," Bell Sys. Tech. J., vol. 54, pp. 297-315, 1975.

[15] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[16] 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.

[17] R. M. Ward, An Investigation Of Speaker Verification Accuracy Using Fundamental Frequency And Duration As Distinguishing Features, Master of Science Report, Electrical & Computer Engineering Department, Clemson University, Clemson, SC, 1988.

TABLE 1

ERROR RATES FOR SELECTED THRESHOLD VALUES FOR 3-D EUCLIDEAN DISTANCE (MALE GROUP)

Threshold	False Rejection Rate (%)				False Acceptance Rate (%)			
	RLD	WMF	REH	Total	RLD	WMF	REH	Total
5	100	100	100	100	0	1.75	0	0.58
9	71.4	85.7	28.6	61.9	3.51	14	0	5.85
10	71.4	85.7	28.6	61.9	3.51	14	3.51	7.02
16	42.9	85.7	14.3	47.6	10.5	33.3	14	19.3
20	28.9	85.7	14.3	42.9	26.3	47.4	42.1	38.6
21	14.3	85.7	14.3	38.1	26.3	49.1	49.1	41.5
25	14.3	57.1	0	23.8	38.6	57.9	61.4	52.6
29	0	42.9	0	14.3	61.4	68.4	71.9	67.3

TABLE 2
ACCURACY RATES

	Individual				Overall	
	Best		Worst			
	50% pt.	Equal %	50% pt.	Equal %	50% pt.	Equal %
3-D Euclidean						
Male	100%	86%	32%	43%	81%	60%
Female	94%	88%	73%	60%	92%	81%
4-D Euclidean						
Male	98%	86%	54%	55%	80%	74%
Female	92%	81%	73%	72%	80%	77%
Mahalanobis						
Male	100%	100%	78%	60%	93%	80%
Female	100%	80%	86%	80%	95%	75%

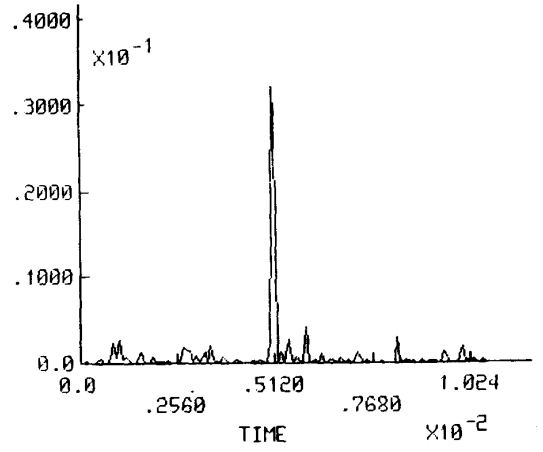

Figure 1. A Plot of the Squared Cepstrum Function of a Voiced Segment of Speech Exhibiting the Peak Corresponding to the Speaker's Pitch Period

* Robert Ward is now with IBM Corp., Kingston, NY.