ON THE CONFIDENCE LEVEL OF SPEAKER IDENTIFICATION USING STATISTICAL MEASURES ON REFLECTION COEFFICIENTS

Miles P. Moody, Sherman Ong

Signal Processing Research Centre
School of Electrical and Electronic Systems Engineering
Queensland University of Technology

ABSTRACT - In this paper, we consider the confidence level for text-independent speaker identification. The confidence level is obtained from an analysis of identification accuracy versus weighted Euclidean distance between the reference template and the test vector of reflection coefficients extracted from segmented speech samples. The acceptance rate (the proportion of the intra-class within a limited distance) is dependent on the distance, the setting of a threshold level is then a trade-off between the accuracy and the acceptance rate. For carefully selected samples with 90% acceptance rate (for one minute for each test and each reference respectively), around 94% accuracy is achieved for a population of 14 and with 40% acceptance rate for the same population, about 99% accuracy is achieved.

INTRODUCTION

The processing principle for speaker identification (SI) is similar to that for speaker verification and speech recognition. The scheme adopted here for SI is shown in Figure 1.

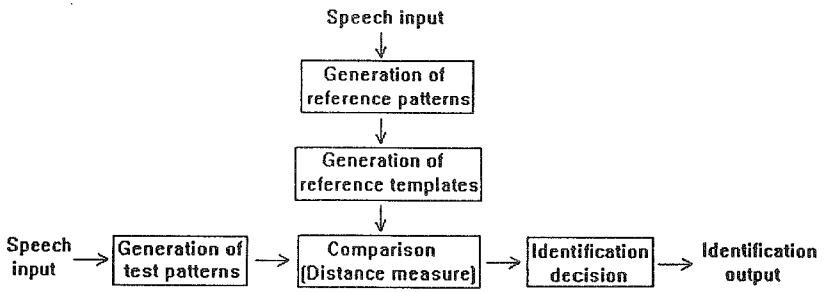

Figure 1. The scheme for speaker identification

Speech samples are text-independent and segmented. Both the test and the reference pattern vectors are generated from individual samples by extracting a set of reflection coefficients as the acoustic feature. Each segment generates one pattern vector. Reflection coefficients, defined as a sequence of ratios of the discontinuity of the cross-sectional area of the vocal tract, have produced good results for SI (Markel & Davis, 1979) (Shridhar & Mohankrishnan, 1982). Reference patterns are statistically averaged to create for each speaker a template and an inverse covariance matrix. Test patterns are compared with any reference template using the method of a weighted Euclidean distance measure (Atal 1976). Each comparison generates one distance. Distances are categorised into two classes, that is, intra-class and inter-class. Intra-class represents the set of distances obtained from comparing the texts of the same speaker (intra-speaker); and inter-class from different speakers (inter-speakers). If in a segment the intra-class distance is smaller than all the inter-class distances, it is called a match, otherwise it is a mismatch. The identification decision is to choose the speaker whose test patterns are closest in distance to the reference template, in other words, to choose the speaker whose segmented test speech has the most matches. The output is expressed by the confusion matrix or by the plot of matching distance.

In this paper, we are considering the confidence level (CL) associated with matching a test speaker with a line-up of reference speakers known to contain the test speakers (i.e. including the intra-speakers). A future paper will investigate the case where the presence of the test speaker in the

reference line-up is unknown. CL is obtained from an analysis of accuracy versus distance. Following is a statement to express CL more clearly: *Given a limited distance for acceptance, how much is the accuracy?* CL is hence a function of a limited distance. Alternatively we could state the problem as: *Given an acceptance rate by limiting the distance, how much is the accuracy?*

MATHEMATICAL MODELLING

First, given the following parameters:

N : the population.
S : total number of segments.
R : the order of the set of reflection coefficients.
$X_{ik}(Y_{ik})$: the test (reference) pattern column vector of the k^{th} segment of speaker i, $1 \le i \le N, 1 \le k \le S$.

Both X_{ik} and Y_{ik} vectors comprise the set of reflection coefficients, $|X_{ik}| = |Y_{ik}| = R$.

Let y_i denote the reference template of speaker i, then

$$y_i = \frac{1}{S}\sum_{k=1}^{S} Y_{ik}, \quad |y_i| = R,$$

and the covariance matrix for speaker i is W_i, W_i is a symmetric $R \times R$ matrix,

$$W_i = \frac{1}{S}\sum_{k=1}^{S} Y_{ik} Y_{ik}^T - y_i y_i^T$$

Now, let d_{ijk} denote the weighted Euclidean distance between the reference template of speaker j and the test pattern of the k^{th} segment of speaker i, then

$$d_{ijk} = \sqrt{(X_{ik} - Y_j)^T W_j^{-1} (X_{ik} - Y_j)}$$

The intra-class distance for speaker i is apparently d_{iik}.

Now, we define p(a,b) as

$$p(a,b) = \begin{cases} 1 & \text{if } a \le b \\ 0 & \text{else} \end{cases}$$

The number of all Intra-class with limited distance d can be denoted as

$$\text{Intra}(d) = \sum_{i=1}^{N}\sum_{k=1}^{S} p(d_{iik}, d)$$

A match within limited distance d for the k^{th} segment of speaker i can be denoted as

$$m_{ik}(d) = \prod_{j=1}^{N} p(d_{iik}, d_{ijk})$$

The number of total matches with limited distance d is

$$T_m(d) = \sum_{i=1}^{N}\sum_{k=1}^{S} m_{ik}(d)$$

Now, CL can be shown as a function of d,

$$CL_d(d) = \frac{T_m(d)}{\text{Intra}(d)}$$

The acceptance rate with limited distance d can be denoted as

$$Ar(d) = \frac{\text{Intra}(d)}{N \times S}$$

Threrfore CL can be shown as a function of the acceptance rate,

$$CL_A(A) = CL_d(d \mid A_r(d) = A)$$

EXPERIMENTS & RESULTS

The experiment was conducted using the signal processing package ILS (Interactive Laboratory System) (Moody & Prandolini, 1990). The subjects were 14 male newsreporters. Their speeches were recorded on the radio during various periods of broadcasting. The test and the reference texts were each 1 minute long, with 10 kHz sampling frequency and 12 bit resolution. Inside each text, the silent parts were edited out to achieve higher accuracy. An edited speech text was compressed to about 50 seconds. Then each text was segmented into 40 segments. Each segment contained 46 frames and each frame contained 256 sample points, which implied each segment length was about 1.18 seconds. Reflection coefficients for both the test and the reference patterns were extracted. There were 40 records (vectors) from each text. Each record was composed of 20 reflection coefficients. Other processing specifications were first order pre-emphasis percentage (98), windowing (Hamming), and analysis filter order (20). For each subject, all 40 records in the reference pattern were averaged to generate a template and an inverse covariance matrix. Then comparisons between all the test patterns (14×40 records) and the set of 14 templates generated a confusion matrix and a large set (14×560) of weighted Euclidean distances from which a plot of matching distance could be drawn. Finally the accuracy and the CL were analysed. The basic parameters in the preceding section were associated with the experimental data as N=14, S=40, and R=20.

The above experiment was repeated for three different data sets. The best result was achieved if the speech texts were carefully selected and edited. Table 1 is the plot of matching distance of the best result from among the three experiments. The abscissa represents distance value. Each row contains the 40 distances for a speaker. " ∘ " represents a match and "+" a mismatch. Table 2 is the confusion matrix from the data of Table 1. Each row contains the 40 comparisons for a speaker. The number of comparisons that cause shortest distance to any template appears on the corresponding column. The number of intra-speaker matches for any speaker hence falls on the diagonal.

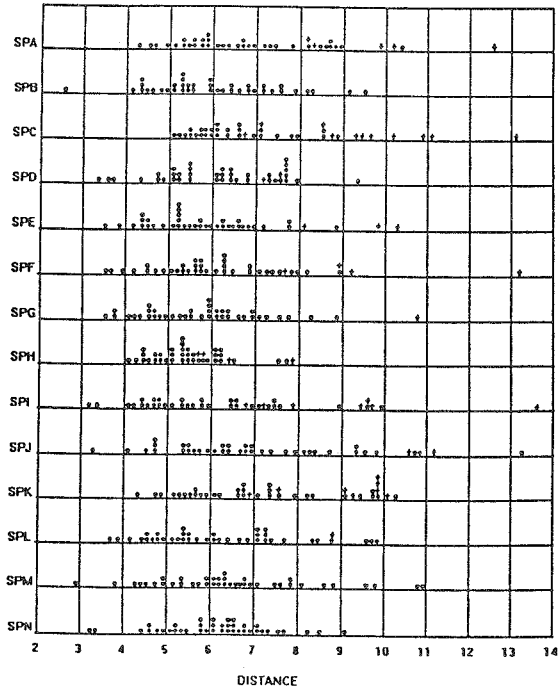

Table 1. The plot of matching distance for the best result

CONFUSION MATRIX - REFERENCE DATA (COLUMNS) VS TEST DATA (ROWS)

	SPA	SPB	SPC	SPD	SPE	SPF	SPG	SPH	SPI	SPJ	SPK	SPL	SPM	SPN
SPA	35		4							1				
SPB		40												
SPC			29					1	1		1	8		
SPD				38						1		1		
SPE					37					1		2		
SPF						36	1		2			1		
SPG							39		1					
SPH			4					36						
SPI							2		33	5				
SPJ		1								38				1
SPK						3		2	1		32	2		
SPL		1										39		
SPM													40	
SPN														40

Table 2. The confusion matrix for the best case.

CONFIDENCE LEVEL ANALYSIS

Figure 2-(a)(b)(c) contains three curves, one for each data set. Figure 2-(a) shows CL versus distance. We see that a specified accuracy can be expected if the distance is limited to a certain level. Figure 2-(b) shows the relationship between the acceptance rate and the distance, this enables the generation of Figure 2-(c). From Figure 2-(c), we see that the setting of a threshold level is a trade-off between the accuracy and the acceptance rate. For the best case, with 90% acceptance rate, about 94% accuracy is achieved; and with 40% acceptance rate, about 99% accuracy is achieved. Even in the worst case, about 94% accuracy can be expected if we limit the distance to 9, or equivalently 80% acceptance rate.

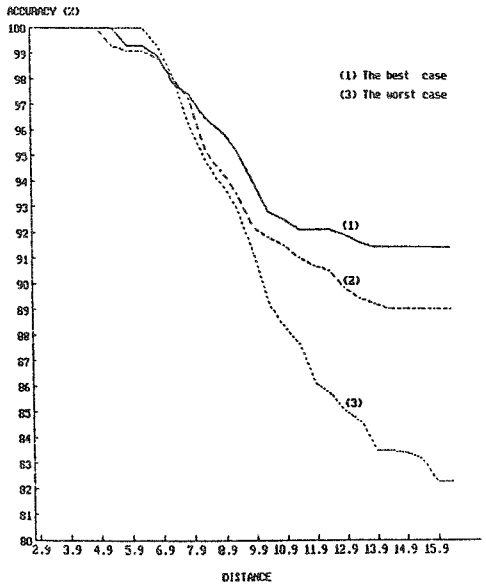

Figure 2-(a). Accuracy versus Distance

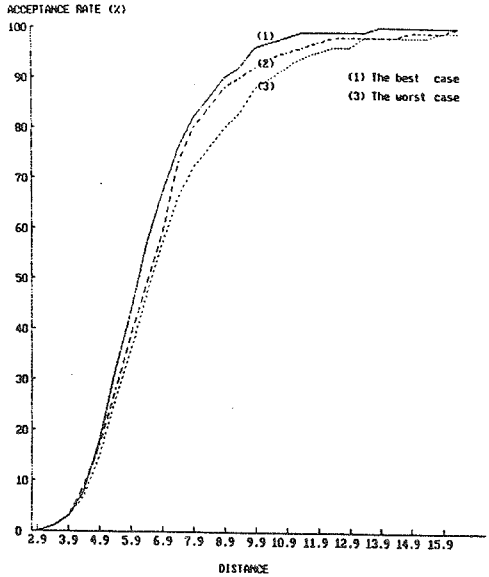

Figure 2-(b). Acceptance rate versus Distance

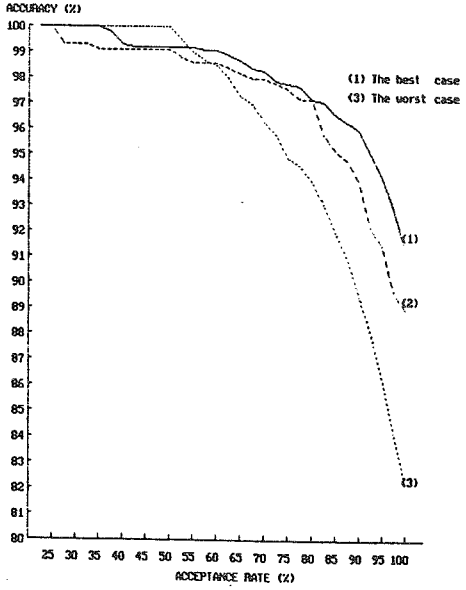

Figure 2-(c). Accuracy versus Acceptance rate

CONCLUSION & WORK IN PROGRESS

From this research, we have come to the following conclusion:
The shown results give a good measure of confidence in matching a speaker with a line-up of speakers which is known to contain the test speaker for distance measures (based on the parameters described) limited to 7 for acceptance, CL close to 100% can be obtained with acceptance rate exceeding 60%.

Future work in progress includes the following:
(1) The case where it is uncertain whether the reference line-up contains the test speaker must be investigated. This will give us a measure of the reliability of matching.
(2) Not only the reflection coefficients, but also the other acoustic features, such as LPC, log area ratio coefficients, and cepstrum, etc., can be investigated in CL analysis.
(3) Hidden Markov Modelling and Artificial Neural Networks are two new methods for SI. The idea of CL might be still suitable if we could find out some kind of parameter conceptually similar to distance.

ACKNOWLEDGEMENT

We are very grateful to Jim W. Newman for help with an earlier version and the operation of ILS.

REFERENCES

Atal, B.S. (1976) "Automatic Recognition of Speakers from Their Voices," Proc. of IEEE, Vol. 64, No. 4, 460-475.

Markel, J.D. & Davis, S.B. (1979) "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base," IEEE Trans. ASSP, Vol. ASSP-27, 74-82.

Moody, M.P. & Prandolini, R. (1990) "Speaker Recognition using ILS," Proceedings of Speech Science Technology Conference, Melbourne.

Shridhar, M. & Mohankrishnan N. (1982) "Text-Independent Speaker Recognition: A Review and some New Results," Speech Commun., Vol. 1, Nos. 3-4, 257-267.