



A Comparative Study of Text-Independent Speaker Identification using Statistical Features

Sherman Ong and Cheng-Hong Yang

Department of Electronic Engineering,
National Kaohsiung Institute of Technology,
Kaohsiung, Taiwan
E-mail : song@nkit.edu.tw

- [Abstract](#)
 - [Introduction](#)
 - [Theory](#)
 - [Distance measures](#)
 - [Karhunen-Loeve transformation](#)
 - Experiment
 - [Speech database](#)
 - [Experimental results](#)
 - [Methods Comparison](#)
 - [Friedman test](#)
 - [Multiple comparison approach](#)
 - [Conclusion](#)
-

Abstract

This paper concerns a comparative study on long term text-independent speaker identification using statistical features. Performances of six statistical methods are compared. Four of them are the distance measures (the City block, the Euclidean, the Weighted Euclidean, and the Mahalanobis distance measures). The other two are the Gaussian probability density estimation and the probability estimation after the Karhunen-Loeve orthogonal transformation. Comparisons are based on two statistical tests (the Friedman test and the multiple comparison approach). Experimental results show that (1) probability calculation is generally better than most distance measures, (2) the orthogonally transformed feature vectors of dimension 15 (originally 20) performs better than all the other methods, (3) the Weighted Euclidean distance measure performs

better than the other three distance measures, and (4) the Mahalanobis distance measure does not perform well. An explanation is advanced for this result in the conclusion.

Keywords: City block distance measure; Euclidean distance measure; Weighted Euclidean distance measure; Mahalanobis distance measure; Gaussian probability density estimation; Karhunen-Loeve transformation; Friedman test; Multiple comparison approach



1. Introduction

Automatic speaker recognition (SR) comprises speaker identification (SI) and speaker verification (SV). SV is the process to verify whether a speaker is who he claims to be from a given speech, whereas SI is to output the identity of the person most likely to have spoken that speech from among a known population [1]. While performance of SV is unaffected by the population size, performance of SI decreases as the population size increases [2].

SR is intrinsically a task of classification in which pattern matching is done between reference models and test patterns. A general scheme for SR is shown in Figure 1.

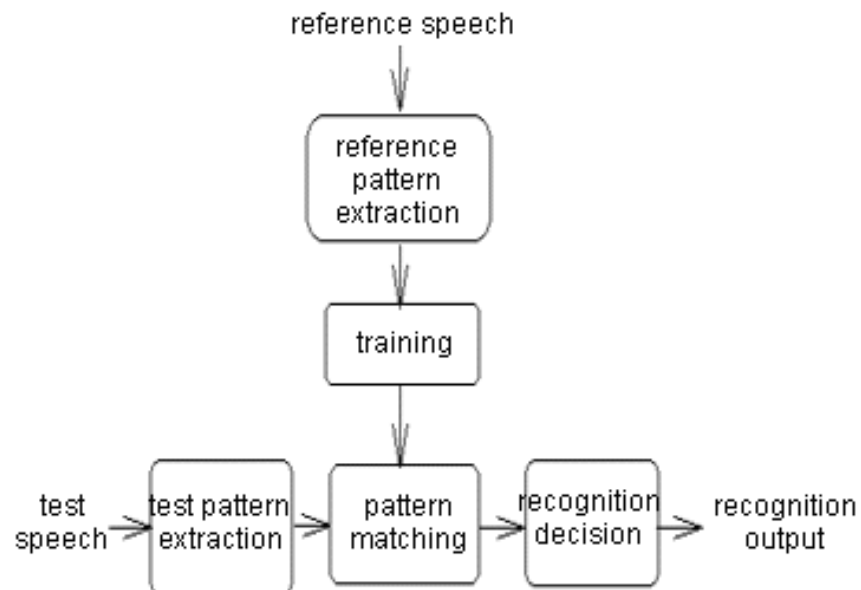


Figure 1. A general scheme for speaker recognition

Test and reference patterns (i.e., acoustic feature vectors) are extracted from speech utterances statistically or dynamically. Dynamic features are harder to get but better in accuracy [3]. Various statistical acoustic features may be used, to name a few, the linear prediction coefficients, the cepstral coefficients, the reflection coefficients, and the log area ratio coefficients [4]. At the training stage, reference models are generated (or trained) from the reference patterns by various methods, e.g., the general statistical methods, Vector Quantization [5], Hidden Markov Modelling [6], and Neural Network [7,8].

Regarding the general statistical methods, a reference model (or template) is formed by obtaining the statistical parameters from the reference speech data. A test pattern is compared against the reference templates at the pattern matching stage. The comparison may be conducted by probability density estimation [9] or by distance (dissimilarity) measure [10,11,12]. After comparison, the test pattern is labelled to a speaker model at the decision stage. The labelling decision is generally based on the minimum risk criterion [13].

Two modes for SR are the text-dependent and the text-independent. In the text-dependent mode, utterances of the same text are used for training and testing. Time alignment [14] is needed for this mode. In the text-independent mode, training and testing involve utterances from different texts. Much more speech utterance is needed for this mode to increase accuracy [15]. Statistical features are better for the text-independent case [1].

In this study, SI is dealt in such a way: acoustic feature vectors of reflection coefficients are statistically extracted and averaged over a long period [16]. A text-independent reference model is formed for each speaker by generating a reference template (a mean vector and a covariance matrix) from the reference feature vectors. Each test feature vector is compared against a reference model by distance measure or by probability estimation. Regarding the distance measure, four variations according to different usage of the covariance matrix [10] are studied. They are the City block (CBD), the Euclidean (ED), the Weighted Euclidean (WED), and the Mahalanobis (MD) distance measures. Regarding the

probability calculation, the Gaussian probability density estimation (GP) [9] and the probability estimation after the Karhunen-Loeve transformation (KLT) [17,18] are studied. A match occurs if a test vector is labelled to the right speaker: for distance measures, it means the intra-speaker distance is smaller than all the inter-speaker ones; while for probability density estimation, it means the intra-speaker probability is larger than the inter-speaker ones. The purpose of this is to meet the minimum risk criterion. Accuracy is then obtained by evaluating the percentage of matches. Performances of the six statistical classification methods are compared through two statistical test methods: the Friedman test [19] and the multiple comparison approach [20].

This paper is organized as follows. In Section 2, theories of the Gaussian probability distribution, the four distance measures, and the Karhunen-Loeve orthogonal transformation are introduced. In Section 3, SI experiment is discussed and experimental results are shown. Section 4 compares performances of the six methods. Section 5 presents the conclusion.



2. Theory

Distance measure is one of the classification methods, it is based on the assumption that the underlying probability has a Gaussian distribution. The Gaussian-distributed probability density function for a class (speaker) is shown as

$$p(x) = \frac{1}{(2\pi)^{N/2} |W|^{1/2}} \exp\left\{-\frac{1}{2}(x - \bar{x})^T W^{-1} (x - \bar{x})\right\} \quad (1)$$

where x is a Gaussian random vector (in column) with dimension N , \bar{x} and W are respectively the mean vector and the covariance matrix for the class model, and T means the transposition of a vector. Suppose it is desired to plot the curve surface of $p(x)$. Let $p(x) = C$, where C is a constant. Then eq. (1) reduces to

$$(x - \bar{x})^T W^{-1} (x - \bar{x}) = C' \quad (2)$$

where C' is another constant related to C in an obvious way. The quantity on the left-hand side of eq. (2) explains the property of distance measure. Due to the normal distribution, intra-speaker distance is generally smaller than inter-speaker distance. Classification is then reduced to the process of finding a speaker model nearest to a given test vector and then labelling this vector to its speaker.

2.1 Distance measures

Let $y^{(i)}$ denote the i^{th} reference speech feature vector (in column) of a certain speaker, where $1 \leq i \leq R$, R is the total number of the reference speech feature vectors for the speaker. The reference template of a certain speaker, \bar{y} , is expressed as

$$\bar{y} = \frac{1}{R} \sum_{i=1}^R y^{(i)} \quad (3)$$

and W , the covariance matrix for the speaker, is expressed as

$$W = \frac{1}{R} \sum_{i=1}^R y^{(i)} y^{(i)T} - \bar{y} \bar{y}^T \quad (4)$$

where $y^{(i)T}$ is the transposed vector (in row) of $y^{(i)}$, and y^t is that of y .

Let x denote a test column vector. The distance between x and y (both with dimension N) is used in defining the four distance measure methods which are defined below respectively :

City block distance : $d_c(x,y)$, or absolute value distance, is defined as

$$d_C(x, \bar{y}) = \sum_{i=1}^N |x_i - \bar{y}_i| \quad (5)$$

where x_i is the i^{th} component of x and y_i is that of y .

Euclidean distance : $d_E(x, y)$, is defined as

$$d_E(x, \bar{y}) = (x - \bar{y})^T (x - \bar{y}) = \sum_{i=1}^N (x_i - \bar{y}_i)^2 \quad (6)$$

where the symbols are the same as in Equation (5).

Weighted Euclidean distance : $d_W(x, y)$, is defined as

$$d_W(x, \bar{y}) = (x - \bar{y})^T D^{-1} (x - \bar{y}) \quad (7)$$

where D a diagonal matrix. Its diagonal elements are exactly the same as that of W .

Mahalanobis distance : $d_M(x, y)$, is defined as

$$d_M(x, \bar{y}) = (x - \bar{y})^T W^{-1} (x - \bar{y}) \quad (8)$$



2.2. Karhunen-Loeve transformation

The purpose of the Karhunen-Loeve transformation is to decorrelate a vector, i.e., to make a vector have pairwise uncorrelated components [17]. The Karhunen-Loeve transformation is implemented by the following steps:

1. find the autocorrelation matrix of a set of column vectors (each vector of N components), denoted as A .
2. generate the eigenvectors and their corresponding eigen values from A , denoted as u_i for the i^{th} eigenvector (in column) and its corresponding eigenvalue λ_i , and

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_N \geq 0$$

3. compose the Karhunen-Loeve transformation matrix, denoted as L_k , and

$$L_k = [u_1 \ u_2 \ u_3 \ \dots \ u_k],$$

where $1 \leq k \leq N$.

4. multiply L_k to an input column vector (say x) to obtain the orthogonally transformed output column vector (say y) i.e.,

$$y = L_k^T x \quad (9)$$

where L_k^T is the transposition of L_k . The output vector can be dimension-reduced according to the variable column size of L_k (see step 3).

3. Experiment

3.1. Speech database

The speech data was obtained the same way by [11]: the subjects were 14 male speakers. Their speeches were recorded onto tapes during various periods of the same FM radio channel. These speeches were then digitised into a personal computer at 10kHz sampling rate by 12 bit resolution. Test and reference utterance sets, for all speakers, were each one minute in duration. Pauses and silent parts were removed from each utterance to achieve higher accuracy. An edited utterance set (40 seconds in length) contained 40 segments. Each segment was about 1 second in duration and contained 40 frames (each frame 256 sample points). From each segment, a vector of 20 reflection coefficients was extracted by getting the mean of the 40 vectors of reflection coefficients in that segment. There were 40 feature vectors of 20 reflection coefficients for either the test or the reference set. Other processing specifications were 98% first order pre-emphasis, non-overlapping 250 points Hamming window, and analysis filter of order 20.



3.2. Experimental results

For each speaker, a mean value and a covariance matrix were generated from the 40 records of the reference set. The two parameters were the basis for either the Gaussian probability estimation or the distance measures.

To calculate the Gaussian probability, each speaker had a reference model expressed by eq. (1). A test vector was applied to every reference model and labelled to the model with largest probability. Accuracy was estimated on percentage of matches.

To process the Karhunen-Loeve transformation, all the reference vectors of the 14 speakers (totally $14 \times 40 = 560$) were combined to form a set. From this set the Karhunen-Loeve transformation matrix was generated. Every vector, in either the test or the reference set, was transformed. Then, the remaining process was the same as that in the paragraph above. Performances of the orthogonally transformed feature vectors with various dimensions (originally 20) were estimated. The result is shown in Figure 2, where performance reaches summit as the dimension size reaches 15, but it then decays slightly as the dimension size increases. The best one (i.e., with dimension 15) was selected and then joined into the comparison of the performances of other methods.

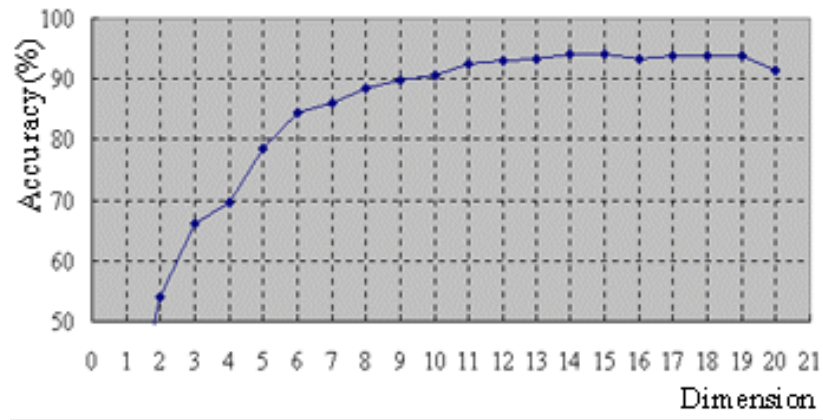


Figure 2. Accuracy versus dimension size for Karhunen-Loeve orthogonal transformation

To do the distance measure, a test vector was applied to each of the reference models to calculate distance through each one of the eqs. (5), (6), (7), and (8). Labelling, matches counting, and accuracy were done as mentioned in Section 1.

Table 1 shows the number of matches for the 14 speakers by the six methods. Each number in the table represents the number of matches out of 40. Accuracy is then equal to the number divided by 40. Accuracies for the 14 speakers are shown in Figure 3.

Table 1. Number of matches (out of 40) for the 14 speakers by the six methods.

Method	<u>Reference speakers</u>													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
CBD	37	39	33	33	38	28	36	30	39	32	28	27	39	40
ED	33	39	26	26	37	31	37	24	38	32	26	23	39	40
WED	38	40	35	39	40	33	37	33	40	34	27	29	39	40
MD	35	40	39	37	37	36	39	36	33	38	32	39	39	40

GP	34	40	29	38	38	35	39	36	33	38	33	39	40	40
KLT	38	40	35	38	39	36	38	38	38	39	33	37	39	40

CBD -- City block distance measure
 ED -- Euclidean distance measure
 WED -- Weighted Euclidean distance measure
 MD -- Mahalanobis distance measure
 GP -- Gaussian probability density estimation
 KLT -- probability estimation after Karhunen-Loeve transformation

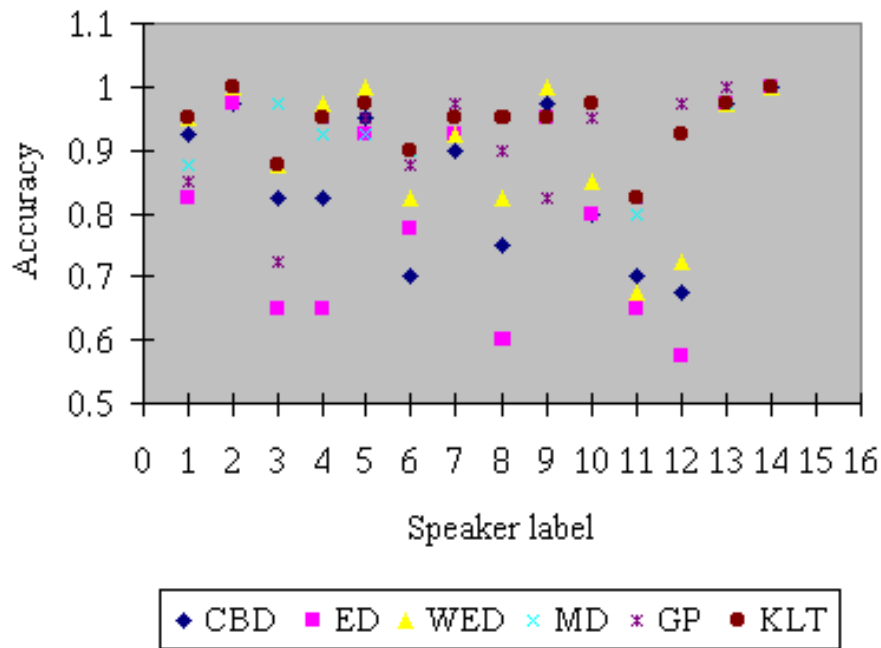


Figure 3. Accuracy for the 14 speakers with the six methods



4. Methods Comparison

The six measure methods were compared by applying the accuracy data in Table 1 to the software package [20]. Two statistical tests, the Friedman test and the multiple comparison approach, were used for the comparison.

4.1. Friedman test

The Friedman test is a nonparametric counterpart of the parametric two-way analysis of variance test and was used to test if the medians of the methods were totally matched when the distribution of the underlying population was not specified. The hypothesis being tested was that all the methods had equal median total matches, and the alternative hypothesis was that all

methods did not have equal median total matches.

Let R_{ij} be the rank (from 1 to k) assigned to method j on problem i . It will equal 1 if it is the lowest value among the methods. In the case of ties, average ranks are used. The test statistic is defined as following:

$$T_f = \frac{(n-1) \left(B_f - \frac{nk(k+1)^2}{4} \right)}{A_f - B_f}$$

where

$$A_f = \sum_{i=1}^n \sum_{j=1}^k R_{ij}^2$$

$$B_f = \frac{1}{n} \sum_{j=1}^k R_j^2$$

The null hypothesis is rejected at the α significance level if the value of the test statistic exceeds the $1-\alpha$ quantile of the F-distribution with $k-1$ and $(n-1)(k-1)$ degrees of freedom. We illustrate the Friedman test on the data in Table 1 ($n = 14$ and $k = 6$). The calculated value of $T_f 7.01$ is greater than the critical value of $F_{.05}^F(5, 65) = 2.37$. We rejected the null hypothesis that all the methods had the same median total matches at the .05 significance level.



4.2. Multiple comparison approach

The multiple comparison approach was used to determine which method had significantly different median total matches. Methods i and j are considered different if the following inequality is satisfied:

$$|R_i - R_j| > t(\alpha/2) \sqrt{2n(A_f - B_f) / ((n-1)(k-1))} \quad \square$$

where R_i , R_j , A_f , and B_f are given previously, and $t(\alpha/2)$ is a critical value on the t-table using $(n-1)(k-1)$ degrees of freedom ($\alpha/2 = P(t_{(n-1)(k-1)} > t(\alpha/2))$). The total matches of the six methods were ordered in an array, and the rank was assigned to each corresponding value as its order. The rank sums of KLT, GP, WED, MD, CBD, and ED were respectively 65.0, 57.0, 55.5, 55.5, 36.0, and 25.0; if the rank sums of any two methods were greater than 12.62 units apart (with $\alpha = .05$), they might be regarded as having unequal

medians total matched. Therefore, it concluded that KLT, GP, WED, and MD might be regarded as superior to CBD and ED. There were no other significant differences.



5. Conclusion

In this study, six methods for long term text-independent speaker identification using statistical features were compared. Four of them were distance measures (i.e., the City block, the Euclidean, the Weighted Euclidean, and the Mahalanobis). The other two were the Gaussian probability estimation and the probability estimation after the Karhunen-Loeve orthogonal transformation. The experimental conditions under which the comparisons were made are stated in Section 3.

The orthogonally reduced feature vectors of dimension 15 (originally 20) performed better than any other dimension as well as any other methods. This might be reasoned that the least significant orthogonal parameters would be indicative of the recording media rather than the utterance itself [18]. Omitting them could improve performance.

The method by the Gaussian probability estimation performed better than most distance measures (only the Weighted Euclidean distance measure was excepted). This might be explained as following: distance measures are simplified models derived from the Gaussian distribution model.

Under the condition of this experiment, the six methods can be grouped in two. KLT, GP, WED, and MD might be regarded as superior to CBD and ED. There were no other significant differences.



References

- [1] D. O'Shaughnessy (1986). "Speaker recognition", *IEEE ASSP Magazine*, pp. 4-17.
- [2] G.R. Doddington (1985). "Speaker recognition - identifying people by their voices", *Proc. IEEE*, Vol. 73, No. 11, pp. 1651-1664.
- [3] S. Furui (1986). "Comparison of speaker recognition methods using statistical features and dynamic features" *IEEE Trans. ASSP*, Vol. 29, No. 3, pp. 342-350.
- [4] M. Shridhar and N. Mohankrishnan (1982). "Text-independent speaker recognition: a review and some new results," *Speech Commun.*, Vol. 1, Nos. 3-4, pp. 257-267.
- [5] F. Soong, A. Rosenberg, L. Rabiner, and B.-H. Juang (1985). "A vector quantization approach to speaker recognition", *ICASSP-85*, pp. 387-390.

- [6] J.M. Naik (1990). "Speaker verification: a tutorial", *IEEE Communications Magazine*, pp. 42-48.
- [7] L. Rudasi and S.A. Zahorian (1991). "Text-independent talker identification with neural networks", ICASSP-91, pp. 389-392.
- [8] H. Hattori (1993). "Text-independent speaker recognition using neural networks", *IEICE Trans. Inf. & Syst.*, vol. E76-D, No. 3, pp. 345-351.
- [9] R. Schwartz, S. Roucos, and M. Berouti (1982). "The application of probability density estimation to text-independent speaker identification", ICASSP-82, pp. 1649-1652.
- [10] R.E. Wohlford, E.H. Wrench, and B.P. Landell (1980). "A comparison of four techniques for automatic speaker recognition", ICASSP-80, pp. 908-911.
- [11] S. Ong and M.P. Moody (1995). "Confidence analysis for text-independent speaker identification using statistical feature averaging", *Applied Sig. Process.*, Vol. 1, No. 3, pp. 166-175.
- [12] R.P. Ramachandran, M.S. Zilovic, and R.J. Mammone (1995). "A comparative study of robust linear predictive analysis methods with applications to speaker identification", *IEEE Trans. Speech Audio Process.*, Vol. 3, No. 2, pp. 117-125.
- [13] C. Basztura (1991). "Experiments of automatic speaker recognition in open sets", *Speech Commun.*, Vol. 10, No. 2, pp. 117-127.
- [14] F. Itakura (1975). "Minimum prediction residual principle applied to speech recognition", *IEEE Trans. ASSP*, Vol. 23, pp. 67-72.
- [15] K.P. Li and G.W. Hughes (1974). "Talker differences as they appear in correlation matrices of continuous speech spectra", *J. Acoust. Soc. Am.*, Vol. 55, pp. 833-837.
- [16] J.D. Markel, B.T. Oshika, and A.H. Gray (1977). "Long-term feature averaging for speaker recognition", *IEEE Trans. ASSP*, Vol. 25, pp. 330-337.
- [17] A. Gersho and R.M. Gray (1992). "Vector quantization and signal processing", Chapter 8, Kluwer Academic Publishers, Boston.
- [18] M.R. Sambur (1976). "Speaker recognition using orthogonal linear prediction", *IEEE Trans. ASSP*, Vol. 24, No. 4.
- [19] W.I. Conover (1980). "Practical nonparametric statistics", 2nd ed., John Wiley and Sons, New York.
- [20] Yang, Cheng-Hong, Shyang-Lung Lin, BinChiang Cheng (1997), An Interactive Statistical Comparison System for Routing Problems, *International Journal of Computer and*

Engineering Management.



[AU Intranet](#), Assumption University, Thailand
Tel.3004543 ext.1315, 3004886