Usefulness of Residual-based Features in Speaker Verification and Their Combination Way With Linear Prediction Coefficients

Wei-Chih Hsu Wen-Hsing Lai Wei-Ping Hong
National Kaohsiung First University of Science and Technology
Department of Computer and Communication Engineering
weichih@ccms.nkfust.edu.tw lwh@ccms.nkfust.edu.tw

Abstract- This thesis focuses on usefulness of the LPC-residue and LPC coefficients in the speaker verification system. First step, in the front-end, feature extraction get the magnitude spectrum of the speech signal from a 32ms short-time segment of speech that is pre-emphasized and processed by a mel-scale filterbank. And the output of the filterbank are then cosine-transformed to produce the cepstral coefficients. After the coefficients have gotten, they are passed to a Gaussian mixture model (GMM). The GMM is used to represent the claimed speaker's acoustic classes. GMM will produce a maximum-likelihood value. If the value is greater than a predefined threshold, the claimed speaker is accepted. The input of our proposed system have two elements; one is the original speech, and the other is the residual signal. In our study, we create a new feature vector. It is composed of the cepstral coefficients (denoted as LPCC), derived from the LPC, and the MFCC of the residual signal. We find that this new feature vector perform the best comparing to the LPCC and residual-MFCC.

keywords：Speaker Verification、Residual、LPC

I. INTRODUCTION

The speech signal conveys two important levels of information. Firstly, the speech signal communicates the words or message being spoken. Secondary, the speech signal communicates information about the identity of the speaker. Therefore speech processing comprises two regions of speech recognition: one is speech recognition and the other is speaker recognition. Relying upon the application, the region of speaker recognition is divided into two specific subdivisions: identification and verification. In identification, the objective is to decide which one of a group of known persons utters the input sound. In verification, the objective is to decide whether the speaker is who claims to be. Moreover, in either condition the speech can be text-independent or text-dependent. In this paper, the subject of text-independent speaker verification is investigated. Speaker recognition is used for many applications such as criminal investigation, automatic money transfer. Success in speaker identification depends on extracting and modeling the speaker-acoustic characteristics of the speech signal that can effectively distinguish one speaker from the other.

In many application of speaker recognition, a problem called "mismatch" must be seriously thought. Many themes have shown that the recognizer drastically reduce in performance due to environmental differences between in the condition of a mismatch between training and test situations. A usual model of the mismatch is that the training speech is recorded on purify environment and the testing speech is corrupted by noise and channel. And the robust speaker recognition tries to keep comparable performance under such distinct situations. There are three major approaches to make up for the mismatch effect, speech enhancement, model adaptation, and feature extraction. Even though the speech enhancement and model adaptation were verified helpful, the robust feature extraction is the simplest method for dealing robust problem. The principal advantage of the robust feature extraction is to use the same method in both training and testing stages. However, many techniques that have a better accuracy always make the system architecture complex. That means these techniques need more parameters and computational time to obtain higher accuracy. We believe that an alleged good method can promote a higher accuracy but sacrifice just a little more computational time.

A. Feature extraction

The first procedure of feature extraction in speaker recognition is to capture speaker-acoustic information that can effectively distinguish speakers from each other. Some methods have been proposed to extract speaker features from the speech signal. These methods are formant frequencies (Doddington,1971), pitch period (Atal,1972;Quatieri,1994), log-area ratios (Furui, 1973) , linear prediction coefficients (Atal,1974), and line spectrum pair frequencies (Liu et al.,1990) glottal flow derivative waveform (Plumpe,1999). Some investigates have found that the nasal sound provides the most discriminative for speaker recognition (Su, 1974; Naik 1990). Many various feature sets have been proposed for speaker recognition. Short-term spectral analysis is the most popular approach speech signal processing nowadays. Linear predictive cepstral coefficients (LPCC) and mel-frequency cepstral coefficient (MFCC) are two feature extraction ways based on short-term spectral analysis (Reynolds and Rose, 1985).

0-7695-3084-2/07 $25.00 © 2007 IEEE
DOI 10.1109/ISM.Workshops.2007.49

B. Speaker Model

Beforetime techniques for speaker recognition are based on long-term statistics of various spectral features, e.g., the mean and variance of spectral. These techniques are intuitive and easy to be implemented for text-independent speaker recognition. Since Hidden Markov Model (HMM) has been successfully applied in speech recognition, numerous researchers strive to use HMM in speaker recognition. Reynolds and Rose proposed one-state HMM, called Gaussian mixture model (GMM) [1]. In our paper, we use the GMM to evaluate our proposed feature extraction method. We find the residual signal includes more speaker-acoustic information than linear prediction coefficients. In the following section. we will discuss how to extract the residual signal, LPCC and MFCC. Then the way of combining residual signal's MFCC and the original signal's LPCC will be discussed.

C. Residual signal

In linear prediction analysis, the sample $s(n)$ is estimated as linear weighted sum of the past p samples. The predicted sample $\tilde{s}(n)$ is given by

$$\tilde{s}(n) = -\sum_{j=1}^{p} a_j s(n-j) \qquad (1)$$

, whree p is the order of prediction, and , $a_j, j = 1,2,...,p$ is a set of linear prediction coefficients (LPC). The LPC are obtained by minimizing the mean-squared error (MMSE) between the predicted sample value and the actual sample value over the analysis frame. The error between the actual value $s(n)$ and the predicted value $\tilde{s}(n)$ is given by

$$e(n) = s(n) - \tilde{s}(n) = s(n) + \sum_{j=1}^{p} a_j s(n-j) \qquad (2)$$

The error is called the linear prediction residual signal of the speech signal. The linear prediction residual signal contains mainly information about the excitation source.

II. EXPERIMENTAL EVALUATION

In this section, we make a speech synthesis experiment that combines the speaker A's residual signal and the speaker B's linear prediction coefficients. Some notations are listed below.

$$e_A(n) + LPC_A \rightarrow S_{AA}(n)$$
$$e_A(n) + LPC_B \rightarrow S_{AB}(n)$$
$$e_B(n) + LPC_A \rightarrow S_{BA}(n)$$

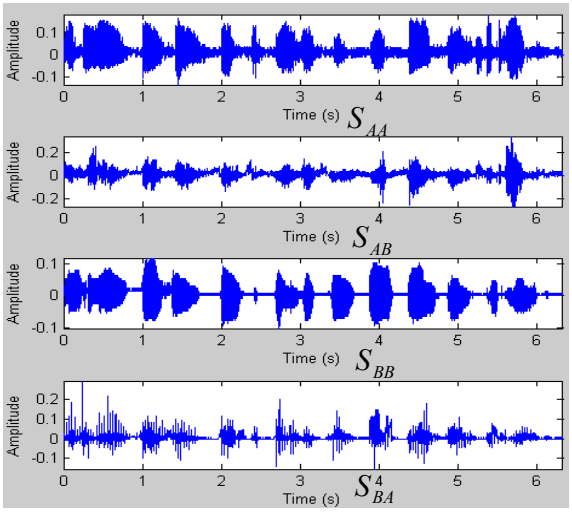

Fig.1. original signal vs. synthetical signal.

$e_A(n)$ denotes the residual signal of the speaker A, and LPC_B denotes linear prediction coefficients of the speaker B. The number of tested sound files for Speaker A and B are both 10. For each sound, residual signal and linear prediction coefficients is computed. Then, we evaluate the result from subjective and objective point of view. By subjective viewpoint, the synthesis signal $S_{AB}(n)$ tends to speaker A when be listened to, and $S_{BA}(n)$ tends to speaker B. From the waveforms shown in Fig.1, we could observe the difference of $S_{AA}(n)$ and $S_{AB}(n)$, the amplitude of waveform of $S_{AB}(n)$ is more weaker than the waveform of $S_{AA}(n)$. From the objective evaluation, we compute the mean square errors of ($S_{AA}(n), S_{AB}(n)$)and $S_{AA}(n), S_{BA}(n)$ .The MSE of $S_{AA}(n), S_{AB}(n)$ is lower than the MSE of $S_{AA}(n), S_{BA}(n)$. The above two results show $e(n)$ is predominant in the synthetical signal, meaning the $e(n)$ includes more speaker's acoustic information than the linear prediction coefficients.

III. RESIDUAL SIGNAL for MFCC

A. Database Description

The experiments were primarily conducted using the MAT2000 speech database . The MAT2000 database is a collection of conversational Mandarin speech from 10 male and 10 female speakers. All speech signals were sampled at 16 KHz. Each speaker utters approximately 30 seconds. The speech was recorded from a high-quality microphone. Before the experiments, we segment the speech to 8 seconds as a set used for training data and testing data. Twenty sets are used as training data and testing data ,respectively.

B. Constructing Gaussian Mixture Speaker Model

The individual component Gaussians in a speaker-dependent GMM is interpreted to represent some acoustic classes. The acoustic classes mirror some speaker-dependent vocal tract structures that are helpful for modeling speaker verification [3][4]. Furthermore, a Gaussian mixture density is proven to provide the long-term distribution of utterances for a speaker. This section will describe the type of the GMM [2] used in this paper as a speaker identity for text-independent speaker verification. The Gaussian mixture speaker model and its parameterization will be explained.

A Gaussian mixture density is a weighted sum of M Component Gaussian densities and given by the equation:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} c_i b_i(\vec{x}) \qquad (3)$$

,where $\vec{x}$ is a D-dimensional random vector, $\{b_i(\vec{x})$, $i = 1,...,M\}$ are the component densities and $\{c_i$, $i = 1,...,M\}$ are the mixture weights. Each component density is a D-dimensional Gaussian function of the equation (4)

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_i \mid^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i)\right\} \qquad (4)$$

with mean vector $\vec{\mu}_i$ and covariance matrix $\sum_i$;. The mixture weights must satisfy the confinement that the sum of c_i is equal to one. A whole GMM is parameterized by the mean vectors, covariance matrices and mixture weights, listing in the foloowing:

$$\lambda = \{c_i, \vec{\mu}_i, \Sigma_i\} \qquad i = 1,...,M \qquad (5)$$

Each speaker is represented by a GMM and is referred to model. The GMM has several different kinds depending on the selection of covariance matrices. It could have one covariance matrix per Gaussian density component, as indicated in equation (4). Or all Gaussian density components in common use single one covariance matrix. In this paper, the diagonal covariance matrices are mainly used for each Gaussian component.

One important issue for employing Gaussian mixture densities as a representation of speaker verification is that the individual component density of a multi-dimension density, like the GMM, is to model some underlying set of acoustic classes. It is rational to assume that the acoustic classes corresponding to a speaker's sound can be described by acoustic classes representing some broad phonetic incidents, like vowels, nasals, or fricatives. These acoustic classes mirror some general speaker vocal tract structures that are helpful in verifying a speaker. The distribution shape of the i th acoustic class can be illustrated by a mean $\vec{\mu}_i$ and a matrix $\sum_i$. These two GMM parameters should be estimated based on the prior information for a speaker.

Once the characteristic feature vectors for the training set have been obtained, they are clustered into M classes according vector quantization (VQ) scheme. The GMM parameters are estimated by using the Expectation Maximization (EM) training algorithm. On every EM iteration of agreement, the following formulas are used, which guarantee a monotonic increase in the model's likelihood value.

For mixture weights, we have

$$\vec{c}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \qquad (6)$$

;for means,

$$\vec{\mu}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \cdot \vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} \qquad (7)$$

;for variances,

$$\vec{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \cdot \vec{x}_t^2}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} - \vec{\mu}_i^2 \qquad (8)$$

E. Cohort Normalization

The cohort normalization has been proven to be efficient in improving speaker verification performance [6][7]. The ordinary approach is to apply a likelihood ratio test to input test utterance *T* using the claimed speaker model λ_c:

$$L(T) = \frac{p(\lambda_c \mid T)}{p(\lambda_{\bar{c}} \mid T)} \qquad (9)$$

Assuming equal prior probabilities, the likelihood ratio in the log domain becomes

$$\log(L(T)) = \log p(T \mid \lambda_c) - \log p(\lambda_{\bar{c}} \mid T) \qquad (10)$$

,where $\lambda_{\bar{c}}$ is representing a cohort model derived from all other possible speakers model. The likelihood $p(T \mid \lambda_{\bar{c}})$ is approximated by averaging the likelihoods of the background speaker models. The background speaker set contains *K* speakers who are acoustically the K closest to the claimant. Cohort normalization involves computing the log-likelihood equation (11) and performing normalization equation (12).

$$\log(p(X \mid \lambda_{\bar{c}})) = \frac{1}{K} \sum_{k=1}^{K} \log\{p(X \mid \lambda_k)\} \qquad (11)$$

$$p_{norm}(T \mid \lambda_i) = \frac{p(X \mid \lambda_i)}{\frac{1}{K}\sum_{k=1}^{K} p(X \mid \lambda_k)} \qquad (12)$$

IV. EXPERIMENTS

A. Performance Comparing For Different Number of Mixtures

The first experiment compares the equal error rate [5] of the Gaussian mixture speaker model for different number of mixtures for MFCC feature extracting methods. The experiment results are shown in Table I. We could find the best result available when the number of mixture is equal to 32 and MFCC is extracted form original $s(n)$. However, this experiment also reveals that the residual signal e(n) still contains the rich speaker information since its MFCC's verification performance is close to that of s(n)'s MFCC.

TABLE I EER Comparison

Input Signal	Feature Vector	Number of Mixture				
		4	8	16	32	64
$s(n)$	MFCC	20	18.8	15.3	15.1	17.5
$e(n)$		36.6	32	23.1	22.8	22.5

TABLE II

Input Signal	Feature Vector	Number of Mixture				
		4	8	16	32	64
$s(n)$	LPCC	29	29.3	28.3	31.2	31.4
$e(n)$		53.1	54	51.7	55.6	56.1

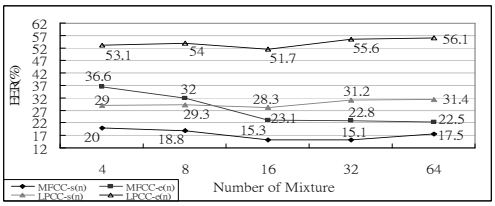

Fig.1. EER for different number of mixture for different input signal and feature extraction methods

In the next experiment, we want to observe the performance of LPCC. Table II summaries the results. It could find that the best EER is at the mixture number of 16. Comparing Table II to Table I, we have the following two observations. First, the performance of LPCC is inferior to that of MFCC. The second observation is that the speaker information hiding in the residual signal e(n) can not be extracted using LPCC feature extraction method. Fig.1. summaries the results of Table I and Table II. From the curve for s(n)'s MFCC, we see that its verification performance is almost alike for 16 and 32 mixture numbers. As a result, we choose the 16 as the mixture number in other experiments.

B. Performance Comparing For Different Element Number of Feature Vector

In the subsequent experiments, we want to see the performance for different element number of feature vectors. In the Table III and Fig2 are shown the experiment results. And Table IV shows the results when the cohort normalization is applied. For element number of 32, the EER is 15.3%, which is the best among all results. And the cohort normalization does improve the performance, about 32%.

TABLE III

Input Signal	Feature Vector	Number of feature vector				
		8	12	14	16	32
$s(n)$	MFCC	25.4	17.1	19.6	16.7	15.3
$e(n)$		28.7	31.6	26.9	29.1	23.1

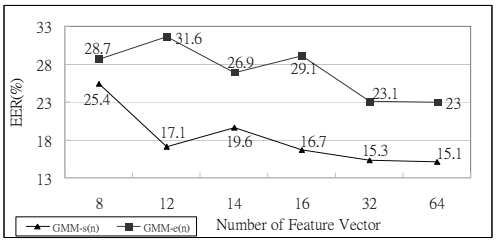

Fig.2. EER for different number of feature vector and input signal.

TABLE IV

Input Signal	Feature Vector	Normalization
		EER (%)
$s(n)$	MFCC	10.6
$e(n)$		17.8

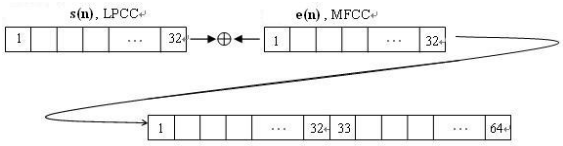

Fig.3. Connect the feature vector

C. Combination of Feature Vector

In these paper, we propose a new method of combining the e(n)'s MFCC and s(n)'s LPCC. The first combination way is depicted in Fig. 3. The $s(n)$'s LPCC is combined with $e(n)$'s MFCC dircetly.

In the Table V are shown the results.

We attempt to change the number of feature vector, and observe the EER. We could see, when LPCC used 14 parameters and MFCC used 16 parameters, the best EER could be obtained. It also can be seen that LPCC(32)+ MFCC(32) performs worse. A possible reason is as follows. The elements in the longer feature vectors have correlation between each other but we assume the covariance matrix between them is diagonal. This causes the poor EER.

TABLE V

Feature vector(number)	EER(%)
LPCC(32)+ MFCC(32)	26.7
LPCC(32)+ MFCC(12)	25
LPCC(16)+ MFCC(16)	16.7
LPCC(16)+ MFCC(12)	15
LPCC(14)+ MFCC(16)	11.7
LPCC(12)+ MFCC(12)	13.3
LPCC(12)+ MFCC(16)	15

TABLE VI

Method		LPCC(14)	
		Correct	**Incorrect**
MFCC (16)	**Correct**	m_{11}=310	m_{12}=131
	Incorrect	m_{21}=104	m_{22}=55

The EER of LPCC(14)+ MFCC(16) is worth paying attention to. Its EER is equal to 6.6%, which is better than $s(n)$'s MFCC by 4% comparing the results of this experiment to those of Table III. In the following, we will discuss why this kind combination could make the EER outperform s(n)'s MFCC. In Table VI, the number of files for correct and incorrect verification are recorded.

In order to see the correlation between e(n)'s MFCC and s(n)'s LPCC, we conduct the χ^2 (chi-square) test, which is a common way to judge the degree of correlation within a confidence level [8]. Let α be the fraction of the total errors made by the LPCC(14) but not for MFCC(16). And let β be the fraction of total errors made by the MFCC(16) but not for LPCC(14). Using the formula described in reference [8], from the data in Table VI, we have α = 0.704 and β = 0.654. Equation (13) and (14) are the expected values of m_{21} and m_{12}, respectively. And he test criterion is defined equation (15).

$$e_{12} = (m_{12} + m_{22}) \cdot \alpha \qquad (13)$$

$$e_{21} = (m_{21} + m_{22}) \cdot \beta \qquad (14)$$

$$\chi^2 = \frac{[m_{12} - e_{12}]^2}{(m_{12} + m_{22}) \cdot \alpha} + \frac{[m_{21} - e_{21}]^2}{(m_{21} + m_{22}) \cdot \beta} \qquad (15)$$

Replacing the value of m_{21} and m_{12} in Table VI, we get equation (16).

$$\chi^2 = \frac{[131 - 186\alpha]^2}{186\alpha} + \frac{[104 - 159\beta]^2}{159\beta} \qquad (16)$$

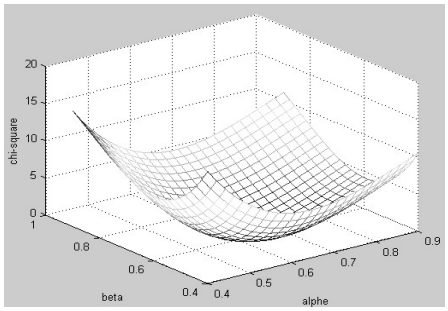

Fig.4. χ^2 function of α and β.

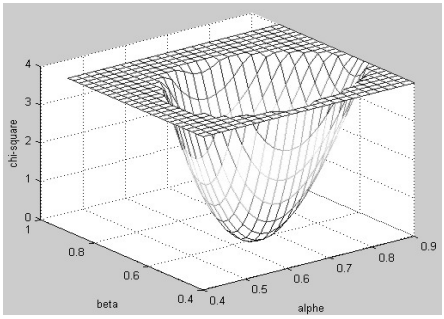

Fig.5. χ^2 values limited to $\chi^2 < 4$.

Fig. 4. shows three dimensional plot of χ^2 as a function of α and β. In Fig. 5. we test the criterion at 5% level of significance, and we plotted the portion of the surface which lies below $\chi^2 = 4$ plane in Fig.6. We could conclude from Fig. 6 that the null hypothesis is accepted for vales of α that are at least 0.48 and for values of β that are at least 0.44. Fig. 5. implies errors made by the two feature vectors are highly uncorrelated. This explains why the combined feature vector outperform s(n)'s MFCC.

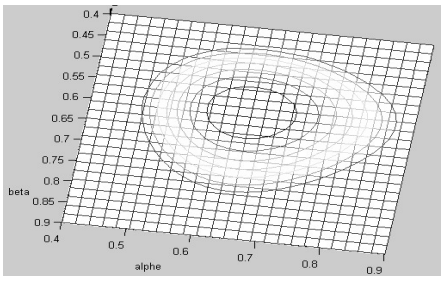

Fig.6. Projection of the plane χ^2.

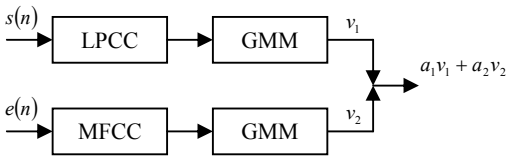

Fig.7. combined the log-likelihood value

D. Mixture Decomposition discrimination

Now that the combination of e(n)'s MFCC and s(n)'s LPCC performs well, in this paper, we also propose another combination way. This time the combination is employed in the likelihood domain, which is called the mixture decomposition discrimination method. The log-likelihood of $s(n)$'s LPCC is combined with $e(n)$'s MFCC. Fig. 7. depicts the combination procedure. The number of feature vector elements is 32. v_1 、 v_2 represent the log-likelihood value of two path, respectively. And a_1 、 a_2 represents the weighting values, which be trained based on training data. The training procedure could be divides into four step. First, input all testing utterances of speaker A to Fig. 7 , and a group of (v_1 、 v_2) = $\{\vec{v}_k\}$,is obtained. Second, input all anti-speaker's utterances of speaker A to Fig.7, $\{\vec{v}'_k\}$. Third, let s_k represent the covariance matrix of $\{\vec{v}_k\}$, s'_k represent the covariance matrix of $\{\vec{v}'_k\}$. $\bar{v}_k$ represent the mean of $\{\vec{v}_k\}$, and $\bar{v}'_k$ represent the mean of $\{\vec{v}'_k\}$. And these values are estimated. Fourth, we obtained a vector a =(a_1 、 a_2), which is the weightings for speaker A. The computation equations are listed in equation (17) t0 (20).

$$w = s_k + s'_k \tag{17}$$

$$d = \bar{v}_k - \bar{v}'_k \tag{18}$$

$$a = w^{-1} \cdot d \tag{19}$$

$$a = (a_1, a_2)^T \tag{20}$$

Using $a_1 v_1 + a_2 v_2$ as a parameter to decide whether tha claimed speaker will be accepted or not. The EER is reduced to 5.5% comparing to that of the combination of LPCC(14)+ MFCC(16). This proves the mixture decomposition discrimination perform the best.

V.CONCLUSION

This thesis has introduced and evaluated the use of Gaussian mixture speaker models for robust text-independent speaker verification. We focus on combining s(n)'s LPCC with e(n)'s MFCC. Two combination ways are proposed which are direct combining and mixture decomposition discrimination. Their performances both outperform the s(n)'s MFCC.

REFRENCES

[1] D. A. Reynolds, et al., "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing 10*,2000, pp.19-41

[2] D. A. Reynolds, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. On Speech and Audio Processing*, Vol.3, No.1 , Jan. 1995, pp.72-83

[3] S. Furui, F. Itakura, and S. Saito, "Talker recognition by longtime averaged speech spectrum," Electron., Commun., in Japan, vol. 55-A, no.10, pp.54-61, 9172.

[4] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. On Acoustic Speech and Signal Processing*, vol. ASSP-29, no.2, pp. 254-272,1981.

[5] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detecyion task performance," in *Proceedings of European Conference on Speech Communication and Technology*, 1997, pp. 1895-1898

[6] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing 10*, 2000, pp. 42-54.

[7] L. Ferrer, et al., "SRI's 2004 NIST Speaker Recognition Evaluation System," in *Proc. ICASSP 2005*, Philadelphia, pp.173-176,March 2005

[8] R. G. D. Steel and J. H. Torrie, *Principles and Procedures of Statistics*, New York: McGraw-Hill, 1960.