Improving Performance of Speaker Identification System Using Complementary Information Fusion

Md. Sahidullah, Sandipan Chakroborty and Goutam Saha
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur, India, Kharagpur-721 302
Email: sahidullah@iitkgp.ac.in, mail2sandi@gmail.com, gsaha@ece.iitkgp.ernet.in
Telephone: +91-3222-283556/1470, FAX: +91-3222-255303

Abstract—Feature extraction plays an important role as a front-end processing block in speaker identification (SI) process. Most of the SI systems use Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), Linear Predictive Cepstral Coefficients (LPCC), as a feature for representing speech signal. Their derivations are based on short term processing of speech signal and they try to capture the vocal tract information ignoring the contribution from the vocal cord. Vocal cord cues are equally important in SI context, as the information like pitch frequency, phase in the residual signal, etc could convey important speaker specific attributes and are complementary to the information contained in spectral feature sets. In this paper we propose a novel feature set extracted from the residual signal of LP modeling. Higher-order statistical moments are used here to find the nonlinear relationship in residual signal. To get the advantages of complementarity vocal cord based decision score is fused with the vocal tract based score. The experimental results on two public databases show that fused mode system outperforms single spectral features.

Index Terms—Speaker Identification, Feature Extraction, Higher-order Statistics, Residual Signal, Complementary Feature.

I. INTRODUCTION

Speaker Identification is the process of identifying a person by his/her voice signal [1]. A state-of-the art speaker identification system requires feature extraction unit as a front end processing block followed by an efficient modeling scheme. Vocal tract information like its formant frequency, bandwidth of formant frequency etc. are supposed to be unique for human beings. The basic target of the feature extraction block is to characterize those information. On the other hand this feature extraction process represents the original speech signal into a compact format as well as emphasizing the speaker specific information. The function of the feature extraction process block is also to represent the original signal into a robust manner. Most of the speaker identification system uses Mel Frequency Cepstral coefficients (MFCC) or Linear Prediction Cepstral Coefficient (LPCC) as a feature extraction block [1]. MFCC is the modification of conventional Linear Frequency Cepstral Coefficient keeping in mind the auditory system of human being [2]. On the other hand, the LPCC is based on time domain processing of speech signal [3]. Later conventional LPCC is also modified motivated by perceptual property of human ear [4]. Like vocal tract, Vocal cord information also contains some speaker specific information [5]. Residual signal which can be obtained from the Linear Prediction (LP) analysis of speech signal contains information related to source or vocal cord. Earlier Auto-associative Neural Network (AANN), Wavelet Octave Coefficients of Residues (WOCOR), residual phase etc. were used to extract the information from residual signal. In this work we have introduced Higher-order Statistical Moments to capture the information from the residual signal. In this paper we are integrating the vocal cord information with vocal tract information to boost up the performance of speaker identification system. The log likelihood score of both the system are fused together to get the advantages of their complementarity [6], [7]. The speaker identification results on both the databases prove that combining the two systems, the performance can be improved over baseline spectral feature based systems.

This paper is organized as follows. In section II we first review the basic of linear prediction analysis followed by the proposed feature extraction technique. The speaker identification experiment with results is shown in section III. Finally, the paper is concluded in section IV.

II. FEATURE EXTRACTION FROM RESIDUAL SIGNAL

In this section we first explain the conventional method of derivation of residual signal by LP-analysis. The proposed feature extraction process is described consequently.

A. Linear Prediction Analysis and Residual Signal

In the LP model, $(n-1)$-th to $(n-p)$-th samples of the speech wave $(n, p$ are integers) are used to predict the $n$-th sample. The predicted value of the $n$-th speech sample [3] is given by

$$\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k)$$  \hspace{1cm} (1)

where $\{a(k)\}_{k=1}^{p}$ are the predictor coefficients and $s(n)$ is the $n$-th speech sample. The value of $p$ is chosen such that it could effectively capture the real and complex poles of the vocal tract in a frequency range equal to half the sampling frequency. The Prediction Coefficients (PC) are determined by
minimizing the mean square prediction error \[ E(n) = \frac{1}{N} \sum_{n=0}^{N-1} (s(n) - \hat{s}(n))^2 \] and the error is defined as

\[ s(n) = -\sum_{k=1}^{p} a(k)s(n-k) + e(n) \] (5)

The LP transfer function can be defined as,

\[ H(z) = \frac{G}{1 + \sum_{k=1}^{p} a(k)z^{-k}} = \frac{G}{A(z)} \] (6)

where \( G \) is the gain scaling factor for the present input and \( A(z) \) is the \( p \)-th order inverse filter. These LP coefficients itself can be used for speaker recognition as it contains some speaker specific information like vocal tract resonance frequencies, their bandwidths etc.

The prediction error i.e., \( e(n) \) is called Residual Signal and it contains all the complementary information that are not contained in the PC. Its worth mentioning here that residual signal conveys vocal source cues containing fundamental frequency, pitch period etc.

**B. Statistical Moments of Residual Signal**

Residual signal which is introduced in Section II-A generally has a noise like behavior and it has flat spectral response. Though it contains vocal source information, it is very difficult to perfectly characterize it. In literature Wavelet Octave Coefficients of Residues (WOCOR) [7], Auto-associative Neural
Network (AANN) [5], residual phase [6] etc are used to extract the residual information. It is worth mentioning here that higher-order statistics have shown significant results in a number of signal processing applications [8] when the nature of the signal is non-gaussian. Higher order statistics also got attention of the researchers for retrieving information from the LP residual signals [9]. Recently, higher order cumulant of LP residual signal is investigated [10] for improving the performance of speaker identification system.

Higher order statistical moments of a signal parameterizes the shape of a function [11]. Let the distribution of random signal \( x \) be denoted by \( P(x) \), the central moment of order \( k \) of \( x \) be denoted by

\[
M_k = \int_{-\infty}^{\infty} (x - \mu)^k dP
\]

for \( k = 1, 2, 3..., \) where \( \mu \) is the mean of \( x \).

On the other hand, the characteristics function of the probability distribution of the random variable is given by,

\[
\phi_x(t) = \int_{-\infty}^{\infty} e^{itx} dP = \sum_{k=0}^{\infty} M_k (\frac{jt}{k!})
\]

From the above equation it is clear that moments \( M_k \) are coefficients for the expansion of the characteristics function. Hence, they can be treated as one set of expressible constants of a distribution. Moments can also effectively capture the randomness of residual signal of auto regressive modeling [12].

In this paper, we use higher order statistical moments of residual signal to parameterize the vocal source information. The feature derived by the proposed technique is termed as Higher Order Statistical Moment of Residual (HOSMR). The different blocks of the proposed feature extraction technique from residual are shown in fig. [2].

At first the residual signal is first normalized between the range \([-1, +1]\). Then central moment of order \( k \) of a residual signal \( e(n) \) is computed as,

\[
m_k = \frac{1}{N} \sum_{n=0}^{N-1} (e(n) - \mu)^k
\]

where, \( \mu \) is the mean of residual signal over a frame. As the range of the residual signal is normalized, the first order moment (i.e. the mean) becomes zero. The higher order moments (for \( k = 2, 3, 4...K \)) are taken as vocal source features as they represent the shape of the distribution of random signal. The lower order moments are coarse parametrization whereas the higher orders are finer representation of residual signal. In fig. [1] LP residual signal of a frame is shown as well as its higher order moments. It is clear from the picture that if the lower order moments are considered both the even and odd order values are highly differentiable.

C. Fusion of Vocal Tract and Vocal Cord Information

In this section we propose to integrate vocal tract and vocal cord parameters identifying speakers. In spite of the two approaches have significant performance difference, the way they represent speech signal is complementary to one another. Hence, it is expected that combining the advantages of both the feature will improve [13] the overall performance of speaker identification system. The block diagram of the combined system is shown in fig. [3] Spectral features and Residual features are extracted from the training data in two separate streams. Consequently, speaker modeling is performed for the respective features independently and model parameters are stored in the model database. At the time of testing same process is adopted for feature extraction. Log-likelihood of two different features are computed w.r.t. their corresponding models. Finally, the output score is weighted and combined.

We have used score level linear fusion which can be formulated as in equation (10). To get the advantages of both the system and their complementarity the score level linear fusion can be formulated as follows:

\[
LLR_{combined} = \eta LLR_{spectral} + (1 - \eta) LLR_{residual}
\]

where \( LLR_{spectral} \) and \( LLR_{residual} \) are log-likelihood ratio calculated from the spectral and residual based systems, respectively. The fusion weight is decided by the parameter \( \eta \).

IIII. SPEAKER IDENTIFICATION EXPERIMENT

A. Experimental Setup

1) Pre-processing stage: In this work, pre-processing stage is kept similar throughout different features extraction methods. It is performed using the following steps:

- Silence removal and end-point detection are done using energy threshold criterion.
- The speech signal is then pre-emphasized with 0.97 pre-emphasis factor.
- The pre-emphasized speech signal is segmented into frames of each 20ms with 50% overlapping i.e. total number of samples in each frame is \( N = 160 \), (sampling frequency \( F_s = 8KHz \)).
- In the last step of pre-processing, each frame is windowed using hamming window given equation

\[
w(n) = 0.54 - 0.46 \cos(\frac{2\pi n}{N-1})
\]

where \( N \) is the length of the window.
2) Classification & Identification stage: Gaussian Mixture Modeling (GMM) technique is used to get probabilistic model for the feature vectors of a speaker. The idea of GMM is to use weighted summation of multivariate gaussian functions to represent the probability density of feature vectors and it is given by

\[ p(x) = \sum_{i=1}^{M} p_i b_i(x) \]  

(12)

where \( x \) is a \( d \)-dimensional feature vector, \( b_i(x), i = 1, ..., M \) are the component densities and \( p_i, i = 1, ..., M \) are the mixture weights or prior of individual gaussian. Each component density is given by

\[ b_i(x) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{rac{1}{2}}} \exp \left\{ -\frac{1}{2} (x - \mu_i)^\top \Sigma_i^{-1} (x - \mu_i) \right\} \]  

(13)

with mean vector \( \mu_i \) and covariance matrix \( \Sigma_i \). The mixture weights must satisfy the constraint that \( \sum_{i=1}^{M} p_i = 1 \) and \( p_i \geq 0 \). The Gaussian Mixture Model is parameterized by the mean, covariance and mixture weights from all component densities and is denoted by

\[ \lambda = \{ p_i, \mu_i, \Sigma_i \}_{i=1}^{M} \]  

(14)

In SI, each speaker is represented by the a GMM and is referred to by his/her model \( \lambda \). The parameter of \( \lambda \) are optimized using Expectation Maximization(EM) algorithm [14]. In these experiments, the GMMs are trained with 10 iterations where clusters are initialized by vector quantization [15] algorithm.

In identification stage, the log-likelihood scores of the feature vector of the utterance under test is calculated by

\[ \log p(X|\lambda) = \sum_{t=1}^{T} p(x_t|\lambda) \]  

(15)

Where \( X = \{ x_1, x_2, ..., x_t \} \) is the feature vector of the test utterance.

In closed set SI task, an unknown utterance is identified as an utterance of a particular speaker whose model gives maximum log-likelihood. It can be written as

\[ \hat{S} = \arg \max_{1 \leq k \leq S} \sum_{t=1}^{T} p(x_t|\lambda_k) \]  

(16)

where \( \hat{S} \) is the identified speaker from speaker’s model set \( \Lambda = \{ \lambda_1, \lambda_2, ..., \lambda_S \} \) and \( S \) is the total number of speakers.

3) Databases for experiments:

**YOHO Database:** The YOHO voice verification corpus [1, 16] was collected while testing ITT’s prototype speaker
verification system in an office environment. Most subjects were from the New York City area, although there were many exceptions, including some non-native English speakers. A high-quality telephone handset (Shure XTH-383) was used to collect the speech; however, the speech was not passed through a telephone channel. There are 138 speakers (106 males and 32 females); for each speaker, there are 4 enrollment sessions of 24 utterances each and 10 test sessions of 4 utterances each. In this work, a closed set text-independent speaker identification problem is attempted where we consider all 138 speakers as client speakers. For a speaker, all the 96 (4 sessions × 24 utterances) utterances are used for developing the speaker model while for testing, 40 (10 sessions × 4 utterances) utterances are put under test. Therefore, for 138 speakers we put 138 × 40 = 5520 utterances under test and evaluated the identification accuracies.

**POLYCOST Database:** The POLYCOST database [17] was recorded as a common initiative within the COST 250 action during January-March 1996. It contains around 10 sessions recorded by 134 subjects from 14 countries. Each session consists of 14 items, two of which (MOT01 & MOT02 files) contain speech in the subject’s mother tongue. The database was collected through the European telephone network. The recording has been performed with ISDN cards on two XTL SUN platforms with an 8 kHz sampling rate. In this work, a closed set text independent speaker identification problem is addressed where only the mother tongue (MOT) files are used. Specified guideline [17] for conducting closed set speaker identification experiments is adhered to, i.e. ‘MOT02’ files from first four sessions are used to build a speaker model while ‘MOT01’ files from session five onwards are taken for testing. As with YOHO database, all speakers (131 after deletion of three speakers) in the database were registered as clients.

4) **Score Calculation:** In closed-set speaker identification problem, identification accuracy as defined in [18] and given by the equation [17] is followed.

\[
\text{Percentage of identification accuracy (PIA)} = \frac{\text{No. of utterance correctly identified}}{\text{Total no. of utterance under test}} \times 100 \quad (17)
\]

**B. Speaker Identification Experiments and Results**

The performance of speaker identification system based on the proposed HOSMR feature is evaluated on both the databases. The order of LP is kept at 17 and 6 residual moments are taken to characterize the residual information. We have conducted experiment based on GMM based classifier for different model order. The identification results are shown in Table I. The identification performance is very low because the vocal cord parameters are not the only cues for identifying speakers but it has some inherent contribution in recognition. At the same time it contains information which are not contained in spectral feature. The combined performance of both the system is to be observed. We have conducted SI experiment using two major kinds of baseline features, some are based on LP analysis (LPCC and PLPCC) and others (LFCC and MFCC) are based on filterbank analysis. The feature dimension is set at 19 for all kinds of features for better comparison. In LP based systems 19 filters are used for all-pole modeling of speech signals. On the other hand 20 filters are used for filterbank based system and 19 coefficients are taken for extracting Linear Frequency Cepstral Coefficients (LFCC) and MFCC after discarding the first co-efficient which represents dc component. The detail description are available in [19], [20]. The derivation LP based features can be found in [1], [4], [21].

The performance of baseline SI systems and fused systems for different features and different model orders are shown in Table I and Table III for POLYCOST and YOHO databases respectively. In this experiment, we take equal evidence from the two systems and set the value of $\eta$ to 0.5. The results for the conventional spectral features follows the results shown in [22]. The POLYCOST database consists of speech signals collected over telephone channel. The improvement for this database is significant over the YOHO which is micro-phonic. The experimental results shows significant performance improvement for SI system compare to only spectral systems for various model order.

### Table I

<table>
<thead>
<tr>
<th>Database</th>
<th>Model Order</th>
<th>Identification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLYCOST</td>
<td>2</td>
<td>19.4960</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>21.6180</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>19.0981</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>22.3138</td>
</tr>
<tr>
<td>YOHO</td>
<td>2</td>
<td>16.8841</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>18.2246</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>15.1268</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>18.2246</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>21.2138</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>21.9565</td>
</tr>
</tbody>
</table>

**IV. Conclusion**

The objective of this paper is to propose a new technique to improve the performance of conventional speaker identification system which are based on spectral features representing only vocal tract information. Higher-order statistical moment of residual signal is derived and treated as a parameter carrying vocal cord information. The log likelihood of both the system are fused together. The experimental results on two popular speech corpus prove that significant improvement can be obtained in combined SI system.

**References**


<table>
<thead>
<tr>
<th>Feature</th>
<th>Model Order</th>
<th>Baseline System</th>
<th>Fused System</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LPCC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>63.5279</td>
<td>71.4854</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>74.5358</td>
<td>72.9122</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>50.5714</td>
<td>59.1876</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>78.3408</td>
<td>82.8391</td>
<td></td>
</tr>
<tr>
<td><strong>PLPCC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>62.9673</td>
<td>70.6462</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>72.2812</td>
<td>75.2968</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>75.0063</td>
<td>77.3210</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>78.3820</td>
<td>80.5040</td>
<td></td>
</tr>
<tr>
<td><strong>LFCC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>63.9257</td>
<td>69.7613</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>72.9443</td>
<td>76.1273</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>77.8515</td>
<td>79.4430</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>77.8515</td>
<td>79.2756</td>
<td></td>
</tr>
<tr>
<td><strong>MFCC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>80.9420</td>
<td>84.7101</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>88.9855</td>
<td>91.0870</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>93.8499</td>
<td>94.7826</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>96.5299</td>
<td>97.1014</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>96.5299</td>
<td>97.1014</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>96.5299</td>
<td>97.1014</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Model Order</th>
<th>Baseline System</th>
<th>Fused System</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LPCC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>63.9257</td>
<td>69.7613</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>72.9443</td>
<td>76.1273</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>77.8515</td>
<td>79.4430</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>77.8515</td>
<td>79.2756</td>
<td></td>
</tr>
</tbody>
</table>


