An Approach to Speaker Identification Using Acoustic-Feature Planes

Hossein Marvi
Shahrood University of Technology
Faculty of Electrical Engineering and Robotic
Shahrood
Iran

Abstract: Speaker recognition is the use of a machine to recognise a person from a spoken phrase. Different features have been investigated in speaker recognition system. Popular features are linear prediction cepstral coefficient (LPCC) and mel frequency cepstral coefficients (MFCC) which both measure the short-term spectral envelope. This paper suggests a new feature for speaker recognition based on time-frequency features using acoustic features planes. Experimental results demonstrate that the proposed method gives higher identification rate than LPCC and MFCC.

Key–Words: Speaker recognition, LPCC, MFCC, MTDRC

1 Introduction

There are many ways to authenticate people such as by handwritten text, fingerprint, iris recognition, face recognition, voice or other features of human beings. The most popular ways are face and fingerprint. However, among these methods voice identification or verification is a traditionally convenient way which is important for many applications such as access control, automatic money transfer, telephone shopping. Automatic speaker recognition is the task of automatically recognising a person from a spoken phrase using speaker-specific information included in speech waves. Successful speaker recognition is critically dependent on obtaining good speaker models from the training data. The problem arises at two points: extraction of features to represent the signal and building the recognition model. The feature extraction is the most critical component of the system and also much more difficult part to be designed than other. Speaker recognition can be classified into two categories: speaker identification and speaker verification. Automatic speaker identification is the use of machine to decide the unknown speaker’s identity among N reference speakers while in automatic speaker verification the system is to verify a person’s claimed identity from his/her voice. Thus, the goal of speaker identification is to develop feature which are unique to each speaker. Speaker recognition also can be text-independent and text-dependent. In text-independent, the recognition procedure works for any text in either training or testing while in text dependent the text in both training and testing is the same. Many different features have been investigated in speaker recognition. Popular features are linear prediction cepstral coefficient (LPCC) and mel frequency cepstral coefficients (MFCC) which both measure the short-term spectral envelope. This paper suggests a new feature for speaker recognition based on time-frequency features using the modified two dimensional cepstrum analysis (MTDRC)[10]. The results will be compared to the LPCC and MFCC feature due to its widespread use and success in speaker recognition application. This paper is structured as follows. The main conventional features for speaker recognition are described in section 2. Acoustic-feature planes using MTDRC analysis is given in section 3. Section 4 deals with classification process. Experimental results are presented in section 5 and finally a conclusion is given in section 6.
2 Conventional features

2.1 Linear prediction cepstral coefficients

One of the most powerful speech analysis techniques is Linear prediction coefficient (LPCC). The important of this method lies both in its ability to provide extremely accurate estimates of speech parameters and its relative speed of computation. The LPCC can be determined as follows [5]:

\[ c_k = \sum_{i=1}^{k-1} (1-i/k) a_i c_{k-i} + a_i \quad \text{for} \quad 1 < k < p \quad (1) \]

\[ c_1 = a_1 \]

\[ c_0 = 0 \]

where \( c_i \) and \( a_i \) are the \( i^{th} \)-order cepstral and linear prediction coefficients respectively. Note, that while there are a finite number of LPC coefficients, the number of cepstrum coefficients is infinite therefore for \( k > p \) we will have:

\[ c_k = \sum_{i=1}^{p} (1-i/k) a_i c_{k-i} \quad \text{for} \quad k > p \quad (2) \]

Normally, in cepstral representation the number of coefficients which is used is around \( 3p/2 \) [12] where \( p \) is the number of LPC coefficients.

2.2 Mel-Frequency cepstral coefficients

Mel-Frequency Cepstral coefficients are mainly applied to Speech and speaker recognition. Perhaps this method is the most popular of feature extraction techniques used for speaker recognition today. MFCC are based on the known variation of the human ear’s critical bandwidths with frequency. Mel-scale frequency is distributed linearly in the low range but logarithmically in the high range corresponding to the physiological characteristics of human ear.

These coefficients can be obtained by simulating critical-band filtering with a set of triangular band-pass filters. The filters are spaced linearly in the range 0 to 1000Hz. The Mel-Frequency cepstrum coefficients can be calculated from the log filter-bank using the Discrete Cosine Transform as it shown in Figure 1. Therefore the Mel-Frequency Cepstral coefficients can be computed by using the following equation:

\[ c_n = \sum_{i=1}^{L} X_i \cos \left[ n \left( i - 0.5 \right) \pi / L \right] \quad \text{for} \quad n = 1, 2, ..., M \]

\[ \beta = \sqrt{2/L} \]

\[ c_{n,m} = \sum_{k=1}^{M} \sum_{i=1}^{L} \beta |E(i)|^\gamma \cos \left[ n \left( i - 0.5 \right) \pi / L \right] e^{j2\pi km/M} \]

\[ \text{for} \quad n = 1, 2, ..., P \quad \text{and} \quad m = 1, 2, ..., M \]

where \( X_i \) is the log-energy output of the \( i^{th} \) filter, \( L \) is the number of filters in the desired bandwidth and \( M \) is the total number of coefficients required. The effects of varying filter bank parameters are discussed in[3].

3 Acoustic-feature planes

An acoustic-feature planes for any speech utterance can be generated by applying the MTDRC analysis to an utterance. This causes an one dimensional signal transfer to an acoustic-features planes which represents the characteristic of any utterance more obviously. These characters can be used as some proper features to represent any utterance and suitable for recognition. The MTDRC analysis can be defined as [10]:

\[ \sum_{k=1}^{M} \sum_{i=1}^{L} \beta |E(i)|^\gamma \cos \left[ n \left( i - 0.5 \right) \pi / L \right] e^{j2\pi km/M} \]

where \( P \leq N \) is the number of coefficients which are used in each frame and \( M \) is the number of frames, \( E(i) \) is the energy of \( i^{th} \) filter, \( L \) is the number of filters in the desired bandwidth and \( \beta = \sqrt{2/L} \). These new MTDRC coefficients are represented in matrix form. As with the original TDRC [2] [9] [11], analysis of the results show that the coefficients located at the lower index portion of the MTDRC matrix are more significant than elsewhere. Thus, a sub-matrix is selected from the lower dimensions of the original matrix.

4 Classification process

To classify the pattern features which are derived from the the acoustic-featur planes (AFP-mtdrc) in the testing phase, a system based on Bhattacharya [6] [1] criteria is used. The Bhattacharya distance outperforms Mahalanobis distance which is more commonly used.
for classification. It is a better but more complicated distance measure for comparing two pattern \( x_r \) and \( x_t \) and can be defined as follows:

\[
d(x_r, x_t) = \frac{1}{8} (\mu_r - \mu_t)^T \left( \frac{C_r + C_t}{2} \right)^{-1} (\mu_r - \mu_t) + \frac{1}{2} \ln \frac{|C_r + C_t|}{|C_r|^{1/2} |C_t|^{1/2}}
\]

(5)

where \( \mu_r \) and \( \mu_t \) are the means value of reference and test data respectively, \( C_r \) and \( C_t \) are the covariance matrices obtained from the feature parameters of reference and test data.

5 Experiments

5.1 Experimental conditions

The experiment have been done over TIMIT corpus. In all 100 speaker have been selected from the TIMIT corpus as training set. The experimental condition are listed as follows:

- Pre-emphasis with \( H(z) = 1 - 0.97 z^{-1} \)
- Frame length 30 ms, Hamming windowed.
- Frame rate 10 ms.
- Remove DC offset.
- Feature set I: Linear predition cepstral coefficient
- Feature set II: Mel-Frequency Cepstral coefficients
- Feature set III: Acoustic feature planes based on MTRDC analysis

5.2 Experimental results

A series of experiments have been carried out to evaluate the performance of suggested method. Each utterance is considered to have a fixed number of frames depending on its length. Each frame size is set to 256 samples and the frames are spaced by 128 samples. To avoid the influence of signal energy, energy normalisation has been applied before any further processing. The total size of the feature matrix would be \( 16 \times 256 \) or \( 32 \times 256 \) etc. but only a sub-set is used to represent an audio signal. Since lots of the used features usually contain no relevant information reducing the feature dimension is a sensible approach towards improving the performance of a classification system. Thus, the linear discriminant analysis (LDA) [4] has also been applied to the AFP-mtdrc features to achieve extra dimensionality reduction while preserving as much of the class discriminatory information and improving the classification accuracy [7] [8].

The percentage of correctly identification has been used as the evaluation criterin. The result for different features size are shown in table 1. For comparison, conventional LPCC and mel cepstral coefficient are computed with same frame rate and equal parameters. These resules are given in Fig.2. Based on the experiments, several observations can be made. Firstly the optimum size of features vector is 60 features. After that when the feature vector size is increased resul get poor. Seconly, it is observed from Fig.2 that the proposed method outperform conventional featurs method such as LPCC and MFCC.

<table>
<thead>
<tr>
<th>Features sizes</th>
<th>ID rate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>61.43</td>
</tr>
<tr>
<td>30</td>
<td>65.7</td>
</tr>
<tr>
<td>40</td>
<td>70.31</td>
</tr>
<tr>
<td>50</td>
<td>74.95</td>
</tr>
<tr>
<td>60</td>
<td>82.24</td>
</tr>
<tr>
<td>70</td>
<td>80.53</td>
</tr>
<tr>
<td>80</td>
<td>75.91</td>
</tr>
<tr>
<td>90</td>
<td>72.46</td>
</tr>
<tr>
<td>100</td>
<td>64.50</td>
</tr>
<tr>
<td>110</td>
<td>60.45</td>
</tr>
<tr>
<td>120</td>
<td>59.01</td>
</tr>
</tbody>
</table>

Table 1: Evaluation results

6 Conclusion

In this paper a speaker recognition technique has been presented that used acoustic features planes based on modified two-dimensional root cepstrum analysis. The performance is found to be superior to the LPCC performance especially in case of speaker identification.
7 Acknowledgement

The author wish to thank the Shahrood University of Technology, for financial support of this work.

References:


