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 predition cepstral coefficient (LPCC)and mel frequency cepstral coefficints (MFCC) which both measure the short-term spectral envelope. This paper suggests a new feature for speaker recogntion based of time-frequency features using acoustic features planes. Experimental resuls demonstrate that the proposed method gives higher identifiction rate than LPCC and MFCC.

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

1 Introduction

There are many wayes to authentic pepole such as by hand written text, finger-print,iris recognition, face recogntion, voice or other features of human being. The most popular ways are face and finger print. However among these methods voice identifiction or verifiction is a traditionally convenient way which is importnat for many applications such as access contorol automatic mony transfer, telephone shopping.
Automatic speaker recogntion is the task of automatically recognition a person from a spoken pharse using speaker-specific information included in speech waves.Successful speaker recogntion 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 recogntion 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 verifications. Automatic speaker identicifation is the ues of machine to decide the unkown speaker's identity among N reference speakers while in automatic speaker verifiction the system is to verify a person's claimed identity from his/her voice. Thus, the goal of speaker identification is to develpoe feature which are unique to each speaker. Speaker recognition also can be text-indepent and text-dependent. In text-indepent, the recognition proceduue works for any text in either training or testing while in text dependent the text in both traing and testing is the same. Many different features have been investigated in speaker recogntion. Popular features are linear predition cepstral coefficient (LPCC) and mel frequency cepstral coefficints (MFCC) which both measure the short-term spectral envelope. This paper suggests a new feature for speaker recogntion based of time-frequency features using the modified two dimentional cepstrum analysis (MTDRC)[10]. The results will be compared to the LPCC and MFCC feature due to its widespad use and sucess in speaker recogntion 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 predition cepstral coefficient

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

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

$$c_1 = a_1$$

$$c_0 = 0$$

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:

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

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 bandpass 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 Discreet Cosine Transform as it shown in Figure 1. Therefore the Mel-Frequency Cepstral coefficients can be computed by using the following equation:

$$c_n = \sum_{i=1}^{L} X_i cos[n(i-0.5)\pi/L] \quad for \quad n=1,2,...,M \qquad (3)$$

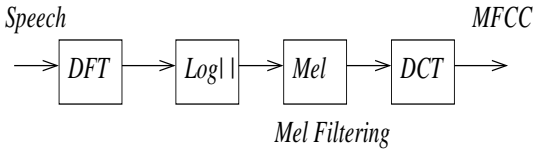

Figure 1: Computing MFCC

where X_i is the log-energy output of the i^{th} filter, L is the number of filters in the desired band-width 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]:

$$c_{n,m} = \sum_{k=1}^{M}\sum_{i=1}^{L}\beta[E(i)]^{\gamma} cos[n(i-0.5)\pi/L]e^{j2\pi km/M}$$
$$for \quad n=1,2,...,P \quad and \quad m=1,2,...,M \qquad (4)$$

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:

$$d(x_r, x_t) = \frac{1}{8}(\mu_r - \mu_t)^T \left(\frac{\mathbf{C_r} + \mathbf{C_t}}{2}\right)^{-1} (\mu_r - \mu_t) + \\ + \frac{1}{2} ln \frac{|\frac{\mathbf{C_r}+\mathbf{C_t}}{2}|}{|\mathbf{C_r}|^{\mathbf{1/2}}|\mathbf{C_t}|^{\mathbf{1/2}}} \quad (5)$$

where μ_r and μ_t are the means value of reference and test data respectively, $\mathbf{C_r}$ and $\mathbf{C_r}$ 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 $H(z) = 1 - 0.97z^{-1}$
- 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 16×256 or 32×256 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 comparsion, 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 result get poor. Seconly, it is observed from Fig.2 that the proposed method outperform conventional featurs method such as LPCC and MFCC.

Features sizes	ID rate %
20	61.43
30	65.7
40	70.31
50	74.95
60	82.24
70	80.53
80	75.91
90	72.46
100	64.50
110	60.45
120	59.01

Table 1: Evaluation results

6 Conclusion

In this paper a speaker recognition thechnique 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 performnace especially in case of speaker identification.

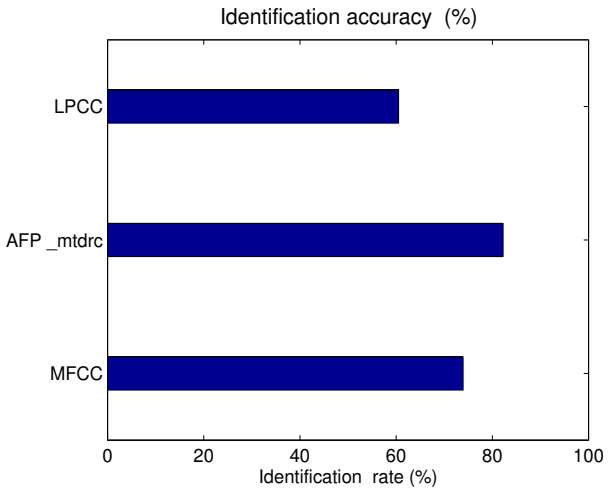

Figure 2: Identification rate (%) for diferent types feature set

7 Acknowledgement

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

References:

[1] A. Bhattachayya. On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society*, 35:99–109, 1943.

[2] E. Chilton and H. Marvi. Two-dimensional root cepstrum as a feature extraction method for speech recognition. *Electronics Letters*, 39(10):815–816, May 2003.

[3] B. Dautrich, L. Rabiner, and T.Marin. On the effects of varying filter speech parameters on isolated word recognition. *IEEE transaction on Acoustic speech and signal processing*, 31:793–807, 1983.

[4] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. NY, J. Wiley, New York, 1973.

[5] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE transaction on Acoustic speech and signal processing*, 2:254–272, 1981.

[6] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. On communication Technology*, Com.-15, February 1967.

[7] H. Marvi and E. Chilton. Application of LDA to improve the accuracy of two-dimensional root cepstrum. In *SCI2003*, volume 10, pages 328–333, Orlando, Florida, USA, 2003.

[8] H Marvi and E. Chilton. Noisy speech recognition through the TDRC analysis. In *Proceeding of the ICEE2004*, volume 1, pages 131–135, 2004.

[9] H Marvi and E. Chilton. Speaker-independent speech recognition using acoustic images based on the TDRC. In *The 18th international congress on Acoustics, ICA2004*, volume II, pages 1735–1738, Kyoto, Japan, 2004.

[10] H. Marvi and E. Chilton. Modified two-dimensional root cepstrum analysis. *IEE, Electronics Letters*, 41(5):285–286, March 2005.

[11] H Marvi and E. Chilton. Comparative study of different distance measures based on the two-dimensional cepstrum for speech recognition. In *IST2003 International symposium on telecommunication*, pages 86–90, 2003, Iran, Isfahan.

[12] L. Rabiner and B.H. Juang. *Fundamentals of speech recognition*. Prentice Hall, 1993.