Abstract

This paper proposes a new feature vector—Mel Frequency Principal Coefficient(MFPC), applied to speaker recognition. It is derived by performing Principal Component Analysis on the Mel Scale Spectrum Vector. Compared with conventional Mel Frequency Cepstrum Coefficient, MFPC efficiently exploited the correlation information among different frequency channels. These correlations, which is mainly caused by the vocal tract resonance, have been found to vary consistently from one speaker to another. And we select these feature coefficients according to their Fisher Ratio, which will guarantee the largest discriminability between classes in the given dimensionality. Finally, we implement a text-independent speaker recognition system. It uses Vector Quantization to design codebooks of given reference speakers. The experiment results demonstrate that our proposed feature vector has characteristics of compactness, large discriminability and low redundancy.

Key Words Mel Frequency Principal Coefficient (MFPC), Principle Component Analysis (PCA), Vector Quantization (VQ), Fisher Ratio

1. Introduction

In the past 30 years, with the evolution of information technology, the capability of automatically identifying the individuals is becoming more and more essential. Personal identification by his voice has been widely studied because of its convenience and acceptability. Nowadays, speaker recognition system is commercialized and applied in many areas, such as entering critical place, the voice login of electronic bank, etc. But there are still many problems to solve in this technology. The most essential one among them is to find a suitable feature to discriminate the speaker.

Research shows that the coefficients based on frequency domain are efficient in speaker recognition. Observing someone's two utterances of the same content in different time, we find that they vary greatly in the wave but are quite similar in the frequency domain. Most of the feature representations that are often used in the literature of speaker recognition such as Mel Frequency Cepstrum (MFCC)[1,3], Linear Prediction Coefficient Ceptrum (LPCCEP)[3], Delta cepstrum[3] etc, are all short-term spectral based features. Among them, the MFCC is undoubtedly the most widely used and successful feature. It reduces the redundancy among the spectrum parameters by summing the energy of each channel. But the simple summation doesn’t consider high-order or even second-order statistics in them. In fact such information is important in speaker recognition. For example, significant degree of correlation exists between the spectrum at different frequencies. These correlations, which are mainly caused by the vocal tract resonance, have been found to vary consistently from one speaker to another[2].

So here we propose the Mel Frequency Principal Coefficients(MFPC) as the feature vector. It is derived by performing PCA on the Short-term Spectrum and is selected according their Fisher Ratio. A text-independent voice recognition subsystem based on Vector Quantization (VQ) of MFPC vectors
is built. The experimental results show the outperformance of our feature vector in the task of speaker recognition.

The paper is organized as follow. In section 2, we introduce the Principal Component Analysis (PCA). Section 3 presents the derivation of Mel Frequency Principal Coefficient and in section 4 we describes the selecting of coefficients according to Fisher Ratio. The architecture of our speaker recognition system and its experimental results are in section 5. Finally the summary and conclusion are given in section 6.

2. Principal Component Analysis

Principal Component Analysis (PCA), also termed Karhunen-Loeve transform, is a kind of linear orthogonal transform that can remove the correlation among the vector components. And it is an optimal dimension reducing method in aspect of mean square reconstruction error. So PCA is widely used in areas such as low bit rate speech coding, feature extraction etc.

Let \( \mathbf{x} = (x_1, x_2, \cdots, x_N)' \) denote the \( N \)-dimensional input vector. The mean vector of \( \mathbf{x} \) can be estimated from \( L \) samples of such vectors by

\[
\mathbf{m}_x = \frac{1}{L} \sum_{l=1}^{L} \mathbf{x}_l
\]

and its covariance matrix by

\[
\mathbf{R}_x = E\{\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)'\} = \frac{1}{L} \sum_{l=1}^{L} \mathbf{x}_l \mathbf{x}_l' - \mathbf{m}_x \mathbf{m}_x'
\]

The covariance matrix is \( N \times N \), real, and symmetric. Its diagonal elements are variances of the individual feature measurements, while the off-diagonal elements are their covariances. Now let matrix \( \mathbf{A} \) define a linear transformation that generates a new vector \( \mathbf{y} \) from \( \mathbf{x} \) by

\[
\mathbf{y} = \mathbf{A}(\mathbf{x} - \mathbf{m}_x)
\]

where \( \mathbf{A} \) contains the eigenvectors of \( \mathbf{R}_x \) in its rows. For convenience, we arrange the rows in order of decreasing magnitude of corresponding eigenvalues.

The transformed feature vector, \( \mathbf{y} \), is zero mean and its covariance matrix is related to that of \( \mathbf{x} \) by

\[
\mathbf{R}_y = \mathbf{AR}_x \mathbf{A}'
\]

Since the rows of \( \mathbf{A} \) are eigenvectors of \( \mathbf{R}_x \), \( \mathbf{R}_y \) is a diagonal matrix having the eigenvalues of \( \mathbf{R}_x \) along its diagonal. Thus

\[
\mathbf{R}_y = \mathbf{AR}_x \mathbf{A}' = \begin{bmatrix} \lambda_1 & 0 \\ & \ddots \\ 0 & \lambda_N \end{bmatrix}
\]

and the \( \lambda_k \) are the eigenvalues of \( \mathbf{R}_y \) as well.

Because the off-diagonal elements of \( \mathbf{R}_y \) are zero, the variables of \( \mathbf{y} \) are uncorrelated. Furthermore, each \( \lambda_k \) is the variance of \( y_k \), the \( k \)th transformed variable.

We can reduce the dimensionality of the output vector by ignoring one or more of the eigenvectors that have small eigenvalues. Let \( \mathbf{B} \) be the \( M \times N \) matrix formed by discarding the lower \( N - M \) rows of \( \mathbf{A} \), and assume, for simplicity, that \( \mathbf{m}_x = \mathbf{0} \). Then the transformed vectors are smaller than \( \mathbf{x} \) and are given by

\[
\hat{\mathbf{y}} = \mathbf{Bx}
\]

but the \( \mathbf{x} \) can still be reconstructed(approximately) by

\[
\hat{x} = \mathbf{B}'\hat{\mathbf{y}}
\]

The mean square error of this approximation is

\[
MSE = \sum_{k=M+1}^{N} \lambda_k
\]

that is, simply the sum of the eigenvalues corresponding to the discarded eigenvectors. Normally the smallest eigenvalues can be ignored without introduction significant error.

3. Extracting the Mel Frequency Principal Coefficients

3.1 The preprocess of speech signal

To realize the speaker recognition, we should preprocess the speech signal above all. And the first
stage of the preprocess is the coarse segmentation of the incoming audio. The purpose of this segmentation is to choose out segments of audio which are likely to contain speech. Here we used a simple and efficient event detector, constructed by thresholding total energy and incorporating constraints on event length and surrounding pause. These constraints are encoded in a finite state machine.

The resulting series of audio clips are arranged together to pass a preemphasis filter. The transfer function of the filter is

\[ H(z) = \frac{1}{1 - 0.95z^{-1}} \]  

(9)

It has two effects: i) enhance the higher frequencies part of the signal, ii) remove the signal's DC offset. The preemphasized signal, \( x(n), 1 \leq n \leq N \), is subdivided into \( T \) frames \( y_t(n), 1 \leq t \leq T \) by an Hamming window \( h_t(n) \)

\[ y_t(n) = x(n) \cdot h_t(n) \quad 1 \leq t \leq T = \frac{N}{S} \]  

(10)

\[ h_t(n) = 0.54 - 0.46 \cdot \cos \left( \frac{2\pi(n-tS)}{L} \right) \]  

(11)

In the equation above, \( L \) represents the length of the frame in samples and \( S \) is the frame displacement also expressed in sample number. In our system, \( L \) and \( S \) are chosen to correspond to 25.625ms and 10ms respectively. So as to the signal sampled in 16kHz, \( L = 410, S = 160 \).

3.2 Extracting MFPC from speech frames

The Extraction of the MFPC from a set of speech frame \( y_t(n), 1 \leq t \leq T \) also contains three steps:

(1) Compute the power spectrum of each short-term frame

This is achieved by Discrete Fourier Transforming (DFT) of \( y_t(n), 1 \leq t \leq T \). Because there are only 410 sample points in \( y_t(n) \), we add zeros to make it contain 512 points. Then we compute the 512-points DFT of it:

\[ \tilde{f}(k) = \sum_{n=0}^{511} y_t(n) e^{-\frac{2\pi ink}{512}} \]  

(12)

Due to the conjugate symmetry of the spectrum, the power spectrum vector is in the size of 256:

\[ F = [F_1, \ldots, F_K]^T \]  

(13)

\[ F_k = [\tilde{f}(k)]^2, 0 \leq k < K = 256 \]

(2) Change the power spectrum \( F \) to Mel scale

The frequency scale is divided into 24 channels which are equal in Mel frequency scale. Then we use the triangular filters to get the energy in each channel:

\[ E_j = \sum_{k=0}^{K-1} \phi_j(k) F_k, \quad 0 \leq j < J \]  

(14)

where \( J \) equals 24, is the number of triangular filters \( \phi_j \) used. These filters cover the 0Hz to Nyquist frequency and satisfy the following constraint:

\[ \sum_{k=0}^{K-1} \phi_j(k) = 1; \quad \forall j \]  

(15)

The distribution of these filters is shown in Fig. 1. And we get the \( J \) dimensional Mel frequency energy vector \( E = [E_1, \ldots, E_J]^T \).

(3) Derive the MFPC from Mel frequency energy vector

By performing PCA on the Mel frequency energy vector \( E \), we get the MFPC vector \( C \):

\[ C = U'(E - \bar{E}) \]  

(16)

where \( U \) is the matrix containing the eigenvectors of \( R_E \) in its columns.
allocation in the frequency domain

4. Feature Selecting According to Fisher Ratio

There are a few different views on selecting those output components of PCA. For example, someone take the opinion that those components corresponding to the larger eigenvalues are suitable for classification; someone believe that those corresponding to smaller eigenvalues are better; others present that the selecting of components should according to the order. Here we propose the method of selecting the feature variables according to their class discriminability. Thus we achieve the goal of reduce the dimension of feature vector while having the best classification performance.

In pattern recognition, the class discriminability of a feature coefficient can be measured by Fisher Ratio[4]:

$$ r_{Fisher} = \frac{\sigma_{between}}{\sigma_{within}} $$

where $r_{Fisher}$ is called Fisher Ratio of a feature coefficient. The larger the Fisher Ratio of the coefficient is, the better the class discriminability of that coefficient. $\sigma_{within}$ is the within-class scatter(variability) for a coefficient $c_k$.

$$ \sigma_{within} = \sum_{i=1}^{c} \left[ \frac{1}{n_i} \sum_{\omega_{i0}} (c_k^{(i)} - m_k^{(i)})^2 \right] $$

In the condition that there have $c$ classes $\omega_{i}, 1 \leq i \leq c$, and the sample number of each class is $n_i$, equation(18) gives the expression of the $\sigma_{within}$ of $k$ th MFPC $c_k$. $m_k^{(i)}$ denotes the mean value of $c_k$ in the $i$ th class. $\sigma_{between}$ is the between-class scatter:

$$ \sigma_{between} = \sum_{i=1}^{c} (m_k^{(i)} - m_k)^2 $$

where $m_k$ is the mean value of $c_k$ under the total sample set.

Fig. 2 shows the Fisher Ratio of the 24 MFPC variables derived in Section 3. From Fig. 2 we can see that the coefficient corresponding to larger eigenvalue dose not have the larger class discriminability. So we choose the 12 coefficients having the largest Fisher Ratio. Compared to other selecting method, it has the best classifying performance when the dimensionalities are same. The experimental results in Section 5 also shows this point.

5. System Structure and Experiment

We realize a text-independent speaker recognition system on the base of MFPC vector. The rear part of the system uses the Vector Quantization(VQ) to clustering the samples. The block diagram of the system is Fig. 3. It can be divided into the training module and the recognizing one. And the VQ part in the training module uses the LBG algorithm.

![Fig. 2 Fisher Ratio of the 24 Mel Scale Principal Coefficients](image)
person, totally 240 speech clips, as the train set. The others are taken as the test set. Each speech segment contains about \( \frac{30000}{160} \approx 180 \) frames, from each of which we get a 256 dimensional spectrum vector \( \mathbf{F} \). On this set, we test the recognition rate of different kinds feature vector and different dimensional of vector.

Table. 1 gives the correct rate of feature vector derived by selecting coefficients according to the larger of eigenvalues and the rate of vector achieved using the Fisher Ratio feature selecting method in same dimensionality.

Table. 2 gives the correct comparison among MFPC vector we proposed and the same size LPC vector and MFCC vector.(using the same VQ method and classifier)

Fig. 4 shows the relationship of the vector size and its correct rate, under two different feature selecting method—by eigenvalue and by Fisher Ratio.

From the above experimental result, we can see that when choosing the feature coefficient according to the Fisher Ratio, the vector has the higher recognition rate while in the same size. And when its size increased, the enhancement of the rate is more robust. Compared to conventional feature vector, the MFPC presented here has an impressive performance in speaker recognition. That shows it has the property of compact, better discriminability and little redundancy.

![The block diagram of our system](image)

**Fig. 3** The block diagram of our system

**Table 1** The correct rate of the feature in different size and derived by distinct method(FR—Fisher Ratio)

<table>
<thead>
<tr>
<th>Size</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select by eigenvalue</td>
<td>30.14%</td>
<td>46.58%</td>
<td>89.04%</td>
<td>93.15%</td>
<td>78.08%</td>
<td>97.26%</td>
<td>71.23%</td>
<td>98.63%</td>
<td>75.34%</td>
</tr>
<tr>
<td>Select by larger FR</td>
<td>49.31%</td>
<td>78.08%</td>
<td>78.08%</td>
<td>93.15%</td>
<td>98.63%</td>
<td>98.63%</td>
<td>96.53%</td>
<td>98.63%</td>
<td>98.63%</td>
</tr>
<tr>
<td>Select by eigenvalue</td>
<td>83.56%</td>
<td>73.97%</td>
<td>98.63%</td>
<td>67.12%</td>
<td>72.60%</td>
<td>79.45%</td>
<td>97.26%</td>
<td>63.01%</td>
<td>89.04%</td>
</tr>
<tr>
<td>Select by larger FR</td>
<td>98.63%</td>
<td>95.42%</td>
<td>98.63%</td>
<td>98.63%</td>
<td>98.63%</td>
<td>98.63%</td>
<td>98.63%</td>
<td>98.63%</td>
<td>98.63%</td>
</tr>
<tr>
<td>Dimensionality</td>
<td>6</td>
<td>12</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFPC vector</td>
<td>98.63%</td>
<td>98.63%</td>
<td>98.85%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC vector</td>
<td>90.54%</td>
<td>95.62%</td>
<td>98.63%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPC vector</td>
<td>87.31%</td>
<td>93.67%</td>
<td>96.54%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 4** Relationship of the vector size and correct rate. The circle points and dashed line shows the relation when selecting coefficients according to eigenvalue and the star points and solid line means selecting according to Fisher Ratio.

### 6. Conclusion

In this paper, we propose a new speaker recognition feature vector—Mel Frequency Principal Coefficients (MFPC) vector. It is based on the short-term spectrum of speech signal. And it has the following advantages: i) little redundancy information in the feature; ii) better class discriminant when in the same size; iii) the compact of this representation. We applied it to a practical speaker recognition system and got a standout performance.

We must point out that there are still many issues to be solved in the speaker identification area. And our future work will have the purpose of further improving the global efficiency and adaptation of MFPC vector and the hardware realization of the system.

### Reference


