The Application of Fusion Technology for Speaker Recognition

Wang He ping1 and Pan Hong xia2

College of Mechanical Engineering and Automatization, North University of China, Taiyuan, China

Summary

This paper describes the fusion technology for speaker recognition. First, these feature sets are extracted form the same speech signal, and then the correlations between the feature sets are analyzed, only the uncorrelation feature sets can be combined. The single feature set and combined feature set are compared. Experiments show that the performance of combined feature set is better than the single feature set. In the model fusion technology, different speaker recognition models sets are combined. The performance of the combined model is compared to other Models. Results show that the combined model can improve the performance of speaker recognition system.

Key words: speaker Recognition Feature fusion model fusion

1. Introduction

With the development of the Internet and communication technologies, speaker recognition plays more and more important roles. The identity of the voice of callers can be used in customer verification for bank transactions, access to bank accounts through telephones, control on the use of credit cards, and for security purposes in the army etc. Although current automatic speaker recognition systems act high performance in laboratory or clean audio conditions, the accuracy degrades rapidly in conditions where background noise or channel distortion are introduced. In order to enhance the performance of the speaker recognition, fusion technology is studied in speaker recognition. Roland Goecke studies an audio-visual automatic speech recognition based on the speech feature and face feature, [1] Georg F. Meyer and Jeffrey Mulligan combine the Mel-frequency cestrum coefficients and lip feature to enhance the system [2].

This paper deals with the fusion of the different features and models for improving speaker recognition. Each feature set is hoped to capture some aspect of the speech signal that may be missed by the other feature sets [3]. In order to provide some benefit to a system, the feature sets must have some level of uncorrelation. If two sets make the same errors on a task, the combination of two sets will not improve the capability of the system. Uncorrelation of the feature sets is a necessary condition to eliminate errors. The performance of different feature sets is compared. In the model fusion technology, different models have different recognition performance; their performance can be used to supplement each other. The performance of BP model, GMM model and VQ model can supplement each other. They are combined by fusion algorithm. Finally, the performance of the combined model is compared to other Models.

2. Feature fusion methods

2.1. Feature selection

1. MFCC and △MFCC Mel-frequency Cestrum Coefficients (MFCC) is a method that analyzes how the Fourier transform extracts frequency components of a signal in the time-domain. This method uses 12~16 absolute MFCC as well as the first and second-order derivatives of those coefficients. △MFCC are derived from the MFCC coefficients.

2. LPCC and △LPCC From each frame of speech signals, LPCC and △LPCC are derived from the LPC coefficients with the following equations:

$$c_m = a_m + \sum_{n=1}^{m-1} (n/m) c_n a_{m-n} 1 \le m \le p \qquad (1)$$

$$\Delta c_m(t) = \sum_{i=-k}^{k} c_m(t+i) \cdot i \qquad (2)$$

Where a_m are the LPC coefficients, p is the LPC analysis order.

3. Pitch The pitch extraction can be based on frequency methods, temporal methods and time- frequency methods. Frequency methods are based on the calculation of FFT, temporal methods are based on the autocorrelation function such as: AMDF, LPC, PPA, time- frequency methods are based on the wavelet transform.

2.2. Correlation analyses

In order to improve the capability of the system, the feature sets that to be combined must have some level of uncorrelation. If two feature sets make the same errors on a task, then no fusion of these feature sets will result in one that remedies these errors. Uncorrelation is a necessary condition for the feature sets to be combined. To resolve this problem, we compute the correlations between the feature sets.

$$\gamma = \frac{\frac{\sum xy}{n} - \overline{X} \bullet \overline{Y}}{\sqrt{\frac{\sum x^2}{n} - \overline{X^2}} \bullet \sqrt{\frac{\sum y^2}{n} - \overline{Y^2}}} \qquad (3)$$

Where n is the data set number; x, y are two variables to be analyzed; $\overline{X}, \overline{Y}$ are the mean value of the xy ; $\sum x^2, \sum y^2$ are the square sum of two variables; $\overline{X^2}, \overline{Y^2}$ are the square of the $\overline{X}, \overline{Y}$.

The correlation of the feature sets is list in table 1. From table. 1, we can see that LFCC is highly correlated with MFCC. This is an expected result, since both of them describe essentially the same quantity, spectral shape. Also cestrum parameters are highly correlated with each other, which can be explained by the method they are computed: △LPCC is merely a differenced version of LPCC.

From the sets, the pitch is least correlated with other feature sets. In order to prove that least correlated feature sets make least error, we select several feature sets to be combined. The fusion scores are compared.

Table 1. Correlation of the feature set

Feature	MFCC	△MFCC	LPCC	△ LPCC	Pitch
MFCC		0.77	0.88	0.71	0.35
△ MFCC			0.73	0.69	0.21
LPCC				0.85	0.27
△LPCC					0.21

2.3. Feature fusion algorithm

Feature fusion is widely concerned in the domain of pattern recognition recently. Its efficiency is mainly determined by algorithm and measuring criterion of feature fusion. The fusion of multiple types of feature vectors is not a simple process of merging different feature subsets. There are numerous ways to combine the multiple types of feature vectors. In this paper, the linear opinion pool method is used for combining the feature sets; the linear opinion pool method computes the final feature set as a weighted sum of the feature sets:

$$F_{linear}(x) = \sum_{i=1}^{n} w_i f_i(x) \qquad (4)$$

Where F_{linear} is the fused feature set, w_i are weights, $f_i(x)$ is the primary set.

The combined feature sets by fusion algorithm include LPCC+ △ LPCC, MFCC+ △ MFCC, LPCC+ △ LPCC+Pitch, MFCC+△MFCC+Pitch. The structure is shown in figure 1.

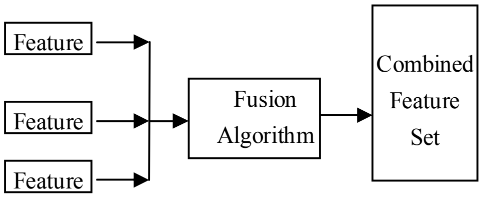

Fig.1.The structure of feature fusion

3. Multi –classifier fusion methods

3.1The structure and algorithm of the Multi-classifier fusion

The structures of multi –classifier machine have tree structure and parallel structure. The parallel multi –classifier machine is the popular method in speaker recognition. In this method, each classifier machine is designed. Then, each classifier machine gives each result. Finally, the fusion algorithm is used to combine each result and decide the output. The structure of multi-classifier machine is shown in figure 2.

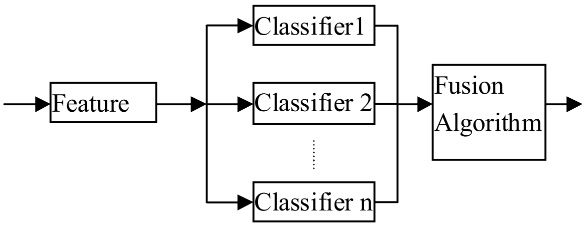

Fig.2 The parallel structure of Multi –classifier

When feature set Z has n class $G = \{\omega_1, \omega_2 \cdots, \omega_N\}$ to choose and m classifier machines $F = \{x_1, x_2, \cdots, x_M\}$ to use, Each classifier machine's output is $P(\omega_n \mid x_m)$ .The all outputs $P(\omega_n \mid x_m), n = 1,2,3.....$ are combined by multi-classifier fusion algorithm, and the class which the feature set Z belong to is decided. There are numerous ways to be chosen as the multi-classifier fusion algorithm, just as feature set fusion, here the linear pool theory is chosen as the fusion algorithm.

3.2 The speaker recognition models

In speaker recognition, the based recognition models have the artificial neural network theory, Gaussian Mixture Model and Vector Quantization. These models have different recognition performance; their performance can be used to supplement each other. The recognition rate can be improved by Multi –classifier. In this paper, The BP neural network theory, Gaussian Mixture Model and Vector Quantization are used as based speaker recognition models to the Multi –classifier fusion. The linear opinion pool is used to the Multi –classifier fusion. The linear opinion pool synthesizes the output of different recognition model. The combined speaker recognition model is shown in figure .3.

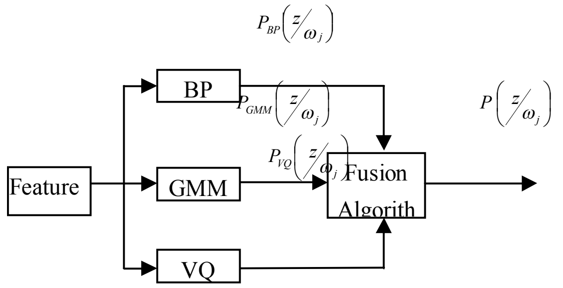
$P_{BP}\left(z/\omega_j\right)$
BP
$P_{GMM}\left(z/\omega_j\right)$
$P\left(z/\omega_j\right)$
$P_{VQ}\left(z/\omega_j\right)$
Feature
GMM
Fusion Algorith
VQ

Fig.3 .The combined speaker recognition model

$P_{BP}(Z \mid \omega_j)$ is probability that the feature set Z belong to some class ω_j. The basic probability assignments can be calculated according to the linear pool theory. In this article, we can calculate the last probability that the feature set Z belong to the class by following equations:

$$\begin{pmatrix} \sum_{m=1}^{3} w_m P^Z(\omega_1 \mid x_m) \\ \sum_{m=1}^{3} w_m P^Z(\omega_2 \mid x_m) \\ \vdots \\ \sum_{m=1}^{3} w_m P^Z(\omega_N \mid xm) \end{pmatrix} = w_1 \begin{pmatrix} P^Z(\omega_1 \mid x_{BP}) \\ P^Z(\omega_2 \mid x_{BP}) \\ \vdots \\ P^Z(\omega_N \mid xB_P) \end{pmatrix} + w_2 \begin{pmatrix} P^Z(\omega_1 \mid x_{GMM}) \\ P^Z(\omega_2 \mid x_{GMM}) \\ \vdots \\ P^Z(\omega_N \mid x_{GMM}) \end{pmatrix} + w_3 \begin{pmatrix} P^Z(\omega_1 \mid x_{VQ}) \\ P^Z(\omega_2 \mid x_{VQ}) \\ \vdots \\ P^Z(\omega_N \mid x_{VQ}) \end{pmatrix} \tag{5}$$

$P^Z(\omega_n \mid x_{BP})$, $P^Z(\omega_n \mid x_{GMM})$ and $P^Z(\omega_n \mid x_{VQ})$ are get from the BP classifier, GMM classifier and VQ classifier. They are the probability that the feature set Z belong to the class $\omega_n (n = 1,2,\cdots N)$. The maximum $\sum_{m=1}^{3} w_m P^Z(\omega_n \mid x_m)$ is the class that feature set Z belong to. Where w_i is the weight, here

$$w_1 + w_2 + w_3 = 1 \quad (6)$$

4. Experiments and results

4.1 Database

The experiments are conducted using a subset of KING speech database. The KING database is a collection of conversation speech from 51 male speakers. For each speaker there are 10 conversations. The speech from a session is recorded from a high-quality microphone locally and is transmitted over a long distance telephone link.

4.2 Feature set fusion experiments

The Gaussian Mixture Model is often used to text-independent speech recognition. Here The Gaussian Mixture Model is used to evaluate the performance of the feature set. First, the performance of each feature set alone is evaluated. Then the performance of the combined feature sets is evaluated. The performance of individual set and combined feature sets is shown in Table 2.

Table.2. Performance of different feature sets

Feature set	Error rate %
LPCC	10.9
△LPCC	9.3
MFCC	9.8
△MFCC	8.7
Pitch	9.5
LPCC+△LPCC	7.6
MFCC+△MFCC	6.8
LPCC+△LPCC+Pitch	4.4
MFCC+△MFCC+Pitch	2.8

The results show that the different combined feature sets can improved the performance of the system. The least correlation between different feature sets combined; the performance is more improved by combined feature sets.

4.3 Multi –classifier fusion experiments

In this experiment, the performance of BP neural network model, Gaussian Mixture Model, Vector Quantization model and combined model is compared. Six people are chosen as the recognition object. The correct recognition rates of different models are shown in table.3.

Table .3. Performance of different models

Model Rate%	BP	GMM	VQ	Combined Model
1	76.21	88.75	78.07	98.62
2	73.15	90.32	78.82	99.78
3	68.23	91.56	79.60	97.68
4	67.65	95.95	77.86	99.58
5	68.98	93.68	78.51	99.65
6	66.82	89.88	78.32	98.23
Average	70.17	91.69	78.49	98.92

The results show that the combined model improves the performance of the recognition system. The performance of the combined model is better than the other recognition models.

5. Conclusions

We proposed two methods for text-independent speaker recognition in this paper. First, feature fusion method is evaluated for text-independent speaker recognition. An analysis of the correlation between different features sets is provided. The different features sets are combined to evaluate the performance of the system. The results show that the least correlation between different features sets combined. The performance is more improved by combined feature sets. Then, model fusion technology is researched. BP, GMM and VQ model are combined. Experiments show that the combined model improves the performance of the recognition system.

6. References

[1].Roland Goecke,Gerasimos Potamianos,nosiy audio feature enhancement using audio-visual speech data,IEEE2002 2025-2028.

[2] K.R.Farrell, R.P.Ramachandran, An analysis of data fusion method for speaker verification,IEEE International Conference on Acoustic, Speech and signal Processing, Seattle,Washing,May 1998,pp,1129-1132.

[3].Y.LU,F.Yanaoka,Fuzzy integration of classification results, Pattern Recognition 30(1997)1877-1891.

[4]. L.Xu ,A.Krzyzak,C.Y.Suen, Methods of combining multiple classifiers and their applications to hand-written character recognition,IEEE Trans.Systems Man Cybernet.23(1992) 18-43