The Algorithm for the Special Speaker Recognition on the Application of Yunnan Province

Ya Wu[1, a *], Min Chen[1,b,*] , ShuLe Zhang[1,c] and HongYi Guo[1,d]

[1]Information Security college, Yunnan Police Office Academy, Kunming, Yunnan, China

[a]shiwangshibiran@qq.com, [b]forestcm@163.com, [c]1192824576@qq.com

[d]614497679@qq.com

* Corresponding Author: Min Chen, forestcm@163.com

Keywords: Speaker Recognition; EMD; Cepstrum of linear predictive coding;

Abstract. In this paper, the speaker recognition algorithm for the special dialect is proposed. The research object of this algorithm aims to the dialect from Yunnan Province. It is different from the previous algorithms on the sight that the Empirical Mode Decomposition is employed to separate the audio signal into details and the cepstrum of linear predictive coding is used as the feature description to help recognition process. Then the proposed algorithm is used to recognize three dialect from Yunnan different areas and the results indicate that the proposed algorithm is better than the similar previous algorithms.

Introduction

The Speaker Recognition System (SRS) aims to recognize the person who give the speech by using the voice feature held in these speech. In the recognizing process, some special features, such as frequency feature, energy and so on, are captured. Then these features are clustered respectively into their corresponding class in order to recognize the nearest class which these features belonged to. Then the alignment will be taken place to judge which person holds these features. The result if SRS is the identification of this people.

In recent years, SRS technology are developed rapidly, furthermore, some systems have been used for some practical application. However, there are some flaws needed to be considered. One of them is that which one or more features should be used for recognition. For this problem, the predecessors gave the different choice for expressing the speech more accuracy. In [1], the basic features the speeches contains are discussed. Meanwhile, in [2], the long term features are also used for the speaker recognition to ensure the accuracy of the identification. These researches aimed to explore the effective features for SRS, but only some basic features are obtained. In [3], the cepstrum of linear predictive coding(LPCC) is used to express the effective feature and it is used to improve the recognition accuracy. This is one effective feature, which lead a lot of linear optimization algorithms can be used for the implement of SRS.

On the other hand, the methods used for identification are also improved. In [4], the statistical features and the corresponding statistical methods are presented to improve the speaker recognition algorithm. Then many similar algorithms are presented, in [5,6], different statistic methods are discussed. However, although these algorithms can increase the identification accuracy but in some time the results will fall into difficulty to recognize some speeches polluted by various extra audio signals. Actually, before the feature capturing, these speech signals should be separated into some more basic signals. The Empirical Mode Decomposition (EMD) is one easy method to separate the signal into different sub-signals and its' efficiency is testified in [7]. In our work, EMD is employed to capture those special sub-signals which are contained in the special speech. Meanwhile, the novel distance measure proposed in [8], the increment of the description length is used to help the implement of clustering.

In this paper, we discuss the strategy of the construction of the recognition algorithm for the speeches from Yunnan Province. By using EMD for separating, some databases are constructed for different areas of Yunnan. Then the weighting method based on the Bayesian average is proposed. The experiments and results are listed at last and the conclusion is given.

Method

Empirical Mode Decomposition. The empirical mode decomposition is one of effective separating methods for the signal decomposition. Despite its' empirical mode, a large scale of researchers use it for signal separating. Based on the data itself, EMD can separate those no-linear signals. The objective of EMD is to make the signal being separated into some IMFs. Each two IMFs satisfy two properties [9]. In EMD process, each signal is separated into some sub-signals $c_i(t)$ and by weighting, these $c_i(t)$ are weighted to form one resulted signal. The process of EMD is given in Fig. 1.

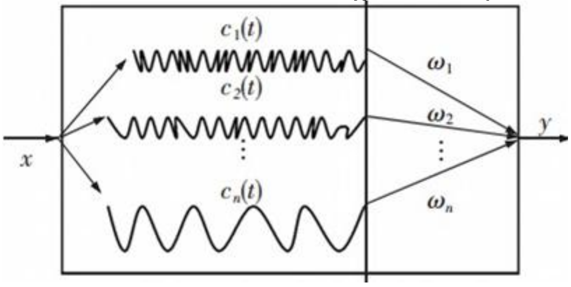

Fig. 1 The process of EMD

In our work, EMD is used firstly to separate the audio signal into some sub-signals and the features of each sub-signal are captured. In this case, the feature of the speech can be described by using the feature matrix instead of the feature vectors, which is formed as (1)

$$\begin{bmatrix} \theta_{1,1},\dots,\theta_{1,k} \\ \dots\theta_{i,j}\dots \\ \theta_{s,1},\dots,\theta_{s,k} \end{bmatrix} \tag{1}$$

Where $\theta_{i,j}$ denotes the j_{th} feature of the sub-signal $c_i(t)$. In order to simplify the recognition process, each row of (1) is actually mapped into a value in practice, which is described as (2)

$$\begin{bmatrix} \theta_{1,1},\dots,\theta_{1,k} \\ \dots\theta_{i,j}\dots \\ \theta_{s,1},\dots,\theta_{s,k} \end{bmatrix} \to \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_s \end{bmatrix} \tag{2}$$

Where each mapped feature δ_i denotes the significant feature of the sub-signal $c_i(t)$. s and k denote the total number of sub-signals and the number of useful features each sub-signal contains.

However, in practice, achieving the mapping result directly is difficult. One essential method to obtain the feature vector is linear predictive coding and its' cepstrum can be used as the description of the feature. In next sub section, the cepstrum of linear predictive coding (LPCC) is discussed.

The cepstrum of linear predictive coding. For the audio signal $s(n)$, its' corresponding predictor $\tilde{s}(n)$ can be described by (3)

$$\tilde{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{3}$$

It implies that the estimated signal comes from the weighting of the previous signals and it is used to express the current state of the speech. Meanwhile, after derivation, it is easy to obtain the cepstrum of representation (3), it satisfy the form (4)

$$\begin{cases} \hat{h}(0)=0 \\ \hat{h}(1)=a_1 \\ \hat{h}(n)=a_k+\sum_{k=1}^{n-1}(1-k/n)a_k\hat{h}(n-k), \quad 1\le n\le p \\ \hat{h}(n)=\sum_{k=1}^{n-1}(1-k/n)a_k\hat{h}(n-k), \quad n>p \end{cases} \tag{4}$$

Where $\hat{h}(n)$ denotes the cepstrum of the linear predictive former (3). In our applications, it is used to express the feature vector for each sub-signal $c_i(t)$. Meanwhile, for the feature matrix (2), each δ_i comes from the calculation (4).

Weighting features. When all δ_i are obtained, we need to achieve one expressing feature which have a exact value. In order to achieve this objective, the weighting operation is employed. Instead of the utilization of GM in [9], directly weighting operation is suggested. Firstly the value of each weight is calculated by normalization. Under the reason that the higher value of cepstrum, the higher predictive error contained. Therefore, the higher cepstrum should be assigned smaller value of weight. Then the final feature value is weighed by (5)

$$\zeta=[\delta_1,\delta_2,\ldots,\delta_s]*([w_s,w_{s-1},\ldots,w_1]^H) \tag{5}$$

In the recognition process, this value ζ is used to compare with one threshold to judge weather one speech is the one can be identified. In order to enhance the accuracy of identification, the similar distance measure in [8] is used, which comes from the minimum description length theory. To simplify our representation, this form is skipped.

Experiments and Results

To testify the proposed algorithm, it is used to recognize the special audio signal. In our experiments, the audio signal is correlated to the dialect of Yunnan Province. Three research audio signals are different dialects which comes from Dehong, Kunming and ZhaoTong. They have the same contents, “How are you, Welcome to my home”. Three signals are given in Fig. 2, Fig. 3 and Fig. 4 respectively.

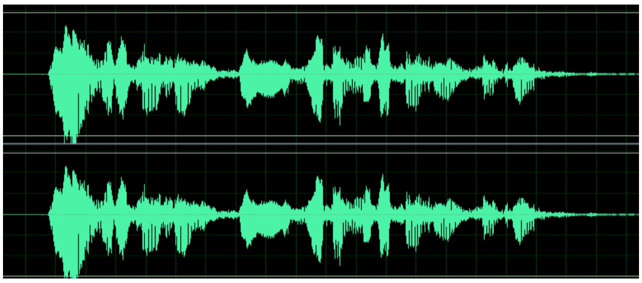

Fig. 2 The audio signal from Dehong dialect

The next is the dialect from Yunnan Kunming and ZhaoTong

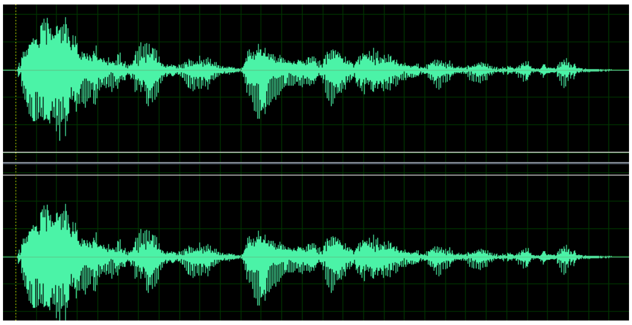

Fig. 3 The audio signal from Kunming dialect

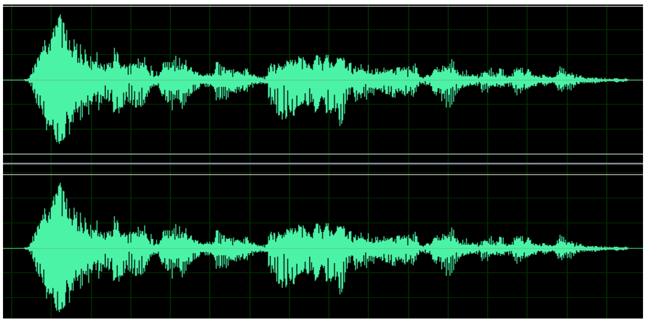

Fig. 4 The audio signal from ZhaoTong dialect

Then the proposed algorithm is employed to recognize these dialect speeches. The results are given in Table 1. Correspondingly, for comparison, the results from the algorithm [9] are also listed in table 1. In this experiment, 30 and 50 persons from different three areas give the speech which contain the contents above. The rate of accurate recognition is used as the results and the criterion.

Table 1 The comparison of rate of accurate recognition

dialect	Dehong		Kunming		ZhaoTong	
method	30 persons	50	30	50	30	50
proposed	98.9%	86.4%	99.4%	87.4%	98.7%	86.9%
Algorithm in [9]	99.1%	85.7%	99.3%	86.8%	98.8%	86.3%

It is obviously that our algorithm can achieve the similar results with the algorithm in [9] on the 30 persons case, but can achieve better recognition rate than the results by algorithm [9]. It implies that our algorithm can provide more stable efficiency fro special speaker recognition.

Conclusions

For the speaker recognition of dialect, the mixture algorithm is proposed in this paper. The audio signal is separated by using EMD and the cepstrum of linear predictive coding is suggested as the representation of the feature vector. After separating, the speech feature is described as an feature matrix and the weighting operation is used to achieve one value to help the recognition process. The experiment results indicate that the proposed algorithm is better than the similar algorithm in literature. The design objective of our algorithm is achieved.

References

[1] M.Smbur. Selection of Acoustic Features For Speaker Identification. IEEE Int.Conf. ASSP. Vol.23:176-182,1975.

[2] J.D.Markel,B.J.Oshika,A.H.Gray. long-term feature averaging for speaker recognition.IEEE Trans on Acoustics Speech and Signal Processing,1977,25(4):330-337.

[3] Douglas A. Reynolds Thomas F. Quateieri and Robert B. Dunn. Speaker verifiction using adapted gaussion mixture models.Digital Signal Processing,Academic Press,2000.

[4] S.Furui. Comparison of speaker recognition method using statistical features and dynamic feathers. IEEE Trans on ASSP,1981,19(3):324-350.

[5] G.Mclachlan and K.Basford. Mixture Models Inference and Applications to Clustering. Marcel Dekker,1998.

[6] D.A.Reynolds. Robust Text-Independent Speaker Identification Using Gaussion Mixture Models. IEEE Trans. Speech and Processing,1995,3(1):72-83.

[7] Manish Gehlot, Yogit Kumar, Harshita Meena, Varun Bajaj, Anil Kumar, EMD Based Features for Discrimination of Focal and Non-focal EEG Signals, Advances in Intelligent Systems and Computing, Vol.340, pp 85-93,2015.

[8] Min Chen, Jianhua Chen, Affinity propagation for the Context quantization, Advanced Materials Research, 2013, 791:1533-1536.

[9] Ying QiXing, Speaker Recognition Based on Speech Database Recorded with PDA, Thesis for Yunnan University, 2011.