International Journal of Advanced Research and Development
ISSN: 2455-4030, Impact Factor: RJIF 5.24
www.advancedjournal.com
Volume 1; Issue 6; June 2016; Page No. 87-90

Hierarchical clustering based gaussian mixture model for text independent speaker recognition

Pardeep Sangwan

Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi, India

Abstract

Speaker recognition is the process of automatically recognizing the person on the basis of the information contained in speech signal of the person. The extraction of this information, called feature extraction and then, feature matching process are implemented right after the pre-processing of the signal. Mel-Frequency Cepstral Coefficients (MFCCs) are taken as features for modeling by the Gaussian Mixture Model (GMM) during the identification process. The GMM algorithm [1, 2] is one of the clustering analyses for the text-independent speaker recognition. One of the main shortcomings of the traditional fuzzy clustering algorithm is the inability of determining the correct number of clusters except by trial and error. In this paper, an algorithm to determine optimal number of clusters automatically is developed by adopting the idea of hierarchical clustering and GMM. Numerical experiments demonstrate that the proposed algorithm achieves better performance than the traditional GMM where the number of clusters is fixed.

Keywords: GMM, GFM, Hierarchical Clustering, MFCC

1. Introduction

The speech signal is a slowly time varying signal called *quasi-stationary*. When examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. Therefore, *short-time spectral analysis* is the most common way to characterize the speech signal. A wide range of possibilities such as Linear Prediction Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC) [3] etc. exists for parametrically representing a speech signal for the task of speaker recognition. MFCC, the most popular feature, is used in this paper.

MFCC is based on the known variation of the human ear's critical bandwidths. This technique makes use of two types of filters, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. Rather than the speech signals, one can employ MFCCs that are less susceptible to the variations [4]. In speaker recognition, the GMMs are used to model the distribution of spectral feature vectors of speakers. The model parameters like *mean vectors, covariance matrices* and *mixture weights* are trained in an unsupervised classification using the Expectation Maximization (EM) algorithm [5]. This algorithm provides an iterative Maximum Likelihood estimation technique. Experiments have shown that GMMs are effective models capable of achieving high identification accuracy for short utterance lengths from the unconstrained conversational speech.

Gan *et al*. [6] proved that under certain conditions the Gaussian Mixture Model (GMM) becomes the Generalized Fuzzy Model (GFM) proposed by Azeem *et al*. [7]. But GFM requires the knowledge of prior clusters. For which one would need a clustering method. To avoid separate clustering, a Hierarchical clustering based modification of GMM is proposed in this paper. This modification provides a remedy for one of the main shortcomings of the traditional fuzzy clustering algorithm, where the number of clusters for obtaining the optimal arrangement is not calculated automatically as it needs the user intervention. By adopting the idea of hierarchical clustering and Gaussian Mixture Modeling, we propose an algorithm that can determine the optimum number of clusters automatically and efficiently. It is demonstrated experimentally that the proposed algorithm achieves better performance than GMM.

2. Gaussian Mixture Model

Let X={$\mathbf{x}_1$, $\mathbf{x}_2$, …., $\mathbf{x}_T$} be a d-dimensional feature vector of Mel Frequency Cepstral Coefficients (MFCC). Since the distribution of these vectors is unknown, it is modeled by a Gaussian Mixture Density, which is the weighted sum of *c* component densities, given by

$$p(\boldsymbol{x}_t/\lambda) = \sum_{i=1}^{c} p_i N(\boldsymbol{x}_t, \boldsymbol{\mu}_i, V_i) \qquad (1)$$

Where p_i, *i*= 1,…..,c, are the mixture weights, $N(\boldsymbol{x}_t, \boldsymbol{\mu}_i, V_i)$, *i*=1, …, c, are the d-variant Gaussian component densities with mean vector $\boldsymbol{\mu}_i$ and covariance matrix V_i

$$N(\boldsymbol{x}_t, \boldsymbol{\mu}_i, V_i) = \frac{exp\left\{-\frac{1}{2}(x_t-\mu_i)V_i^{-1}(x_t-\mu_i)'\right\}}{(2\pi)^{\frac{d}{2}}\,|V_i|^{\frac{1}{2}}} \qquad (2)$$

We denote here a set of all parameters contained in the probability model by $\lambda = \{p_i, \mu_i, V_i\}$ *i*= 1, …, c. While training the GMM, these parameters are estimated in some sense such that they best match the distribution of the training vectors. The maximum likelihood (ML) estimation method is adopted for learning the parameters. For a sequence of training vectors *X*, the likelihood of the GMM is

$$p(X/\lambda) = \prod_{t=1}^{T} p(x_t/\lambda) \qquad (3)$$

The ML estimation finds a new parameter model $\bar{\lambda}$ such that $p(X/\bar{\lambda}) \geq p(X/\lambda)$. Maximizing $p(X/\lambda)$ is not easy; hence an auxiliary function ***J*** is used

$$J(\lambda,\bar{\lambda}) = \sum_{i=1}^{T} p(i/x_t,\lambda)\log[\bar{p}_i N(x_t,\bar{\mu}_i,\bar{V}_i)] \qquad (4)$$

Where $p(i/x_t,\lambda)$ is the a posteriori probability for acoustic class *i*, *i* = *1,…,c* and it satisfies the relation:

$$p(i/x_t,\lambda) = \frac{p_i N(x_t,\mu_i,V_i)}{\sum_{j=1}^{c} p_j N(x_t,\mu_j,V_j)} \qquad (5)$$

Maximizing the ***J*** function is accomplished using the EM algorithm. Setting the derivatives of the *Q* function with respect to $\bar{\lambda}$ to zero, estimation formulas are derived as

$$\bar{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i/x_t,\lambda) \qquad (6)$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} p(i/x_t,\lambda)x_t}{\sum_{t=1}^{T} p(i/x_t,\lambda)} \qquad (7)$$

$$\bar{V}_i = \frac{\sum_{t=1}^{T} p(i/x_t,\lambda)(x_t-\bar{\mu}_i)(x_t-\bar{\mu}_i)}{\sum_{t=1}^{T} p(i/x_t,\lambda)} \qquad (8)$$

The model parameters comprising *mixture weights*, *mean vectors* and *covariance matrices* are trained in an unsupervised classification using EM algorithm [5].

3. Hierarchical Clustering based GMM

This algorithm combines the GMM algorithm with the framework of hierarchical clustering. To elaborate this, first consider the whole data set as one cluster and then invoke GMM to divide the data set into two clusters. This process is continued and at each stage certain number of clusters is chosen by some criteria. This process is repeated until the number of clusters attains the cluster range as per Ward's method [8] to split the clusters.

The Ward's method to split the cluster invokes GMM algorithm explained as under.

Set C_i (*i* = 1,2,…,*k*), where C_i is a *i*th cluster; *W*i (*i* = 1,2,…,*k*) is the cluster variance of the *i*th cluster. Split the cluster C_i into C_i^1, C_i^2, the corresponding variances being V_i^1, V_i^2. Then the decrement of cluster variance is defined as follows:

$$\delta(C_i) = W_i - W_i^1 - W_i^2 \qquad (9)$$

where $W_i = \sum_{x \in C}\|x - \bar{\mu}\|^2$. Here $\delta(C_i)$ is the error sum square of cluster variance that indicates the compactness of clusters.

For any two clusters C_i, C_j if $\delta(C_i) > \delta(C_j)$, the smaller cluster variance results from splitting C_i. This is reverse to the Ward's method, where two clusters are merged leading to the minimum cluster variance.

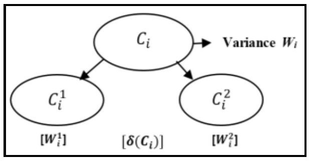


Fig 1: Error in cluster variance after splitting

The basic criteria to measure the clustering effect include the divergence and the compactness which imply that the inter-cluster distance must be as large as possible, and intra-cluster distance must be as small as possible. In [8] a new index was proposed based on cluster variance.

Let *T(K)* be the total variance at K^{th} stage of clustering, then

$$S(K) = \frac{T(K-1)-T(K)}{T(K)-T(K+1)} = \frac{Max[\delta(K)]}{Max[\delta(K+1)]} \qquad (10)$$

Where $Max[\delta(K)]$ and $Max[\delta(K+1)]$ are the values of the maximum error variance at K^{th} and $(K+1)^{th}$ stage of clustering respectively.

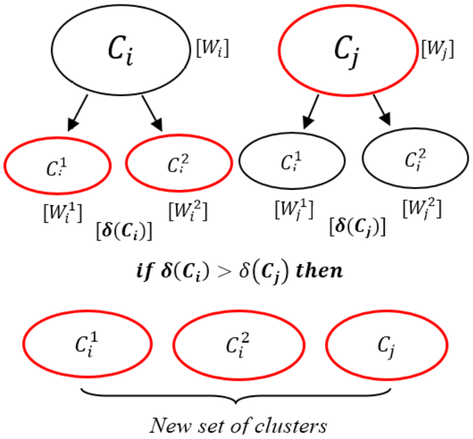


Fig 2: Methodology of cluster splitting

Equation (10) expresses the variation of cluster variance in the two adjacent splitting processes. The maximum index shows that there is less error variance in the next splitting, that is to say that the optimum number of clusters is the one with the maximum in (10).

Algorithm

Inputs: **X**: Feature vector set
Cmax: Maximum number of clusters

Output: **C**opt: optimum number of clusters

1. Invoke GMM(X, 1); to create a model considering whole data as one cluster; set K=1 and Repeat.
2. Invoke GMM(C_i, 2) on each cluster *i* = *1,…,K*; split it into C_i^1, C_i^2;
3. Compute $\delta(C_i)$ for each cluster using (9). If i_o is the index that leads to the maximum value of $\delta(C_i)$, then replace C_{i_o} with $C_{i_o}^1$, $C_{i_o}^2$;
4. K= K +1; until K > Cmax
5. For each C_i, compute $S(K)$ using (10);
6. Set **C**opt = K that leads to the maximum value of $S(K)$;
7. Return **C**opt ;

4. Validation Results

We have used two databases. Figure 3 depicts the male and female speech waveforms of the same text spoken.

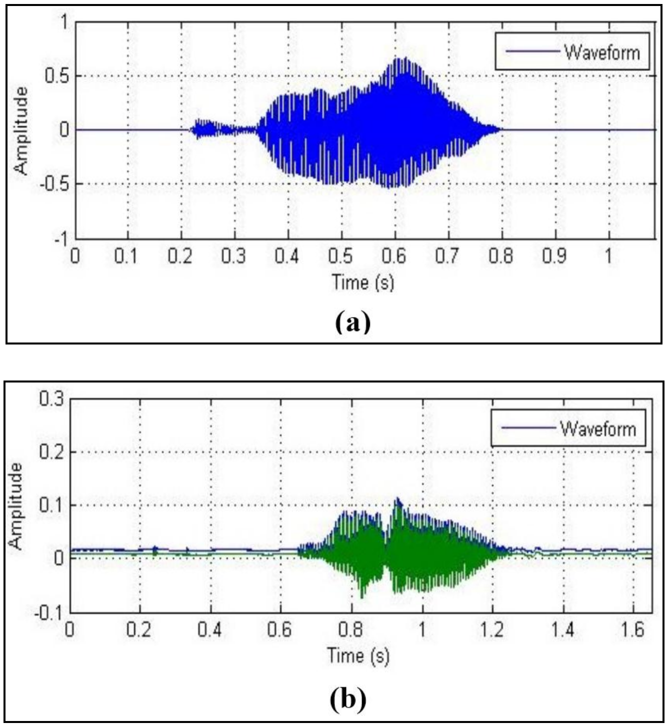
Database 1

Contains 100 speech samples at the rate of 5 speech samples per speaker. These are the outcome of each speaker spoken on different text of length 28 sec to 30 sec, sampled at 16

KHz. In which 3 samples are given for training and 2 for testing. Each speech sample is divided into frames of length 30 ms. Each frame is windowed using Hann window. These windowed frames are transformed into frequency domain using FFT. Then logarithmic of this spectrum is wrapped with Mel filter banks containing 20 triangular filters yields Mel Frequency Cepstral Coefficients (MFCC) called the feature vectors of dimensionality 20. Table 1 shows sample feature vectors of a speech sample. These feature vectors are modeled using GMM and proposed algorithm with unsupervised learning and Table 3 shows the comparison of both for Database 1.

Database 2

Contains 50 speakers each of 10 speech samples spoken on different text of length 8 sec to 12 sec, sampled at 16 KHz. In which 8 are given for training and 2 for testing. 598 to 600 feature vectors are extracted using MFCC. Table 2 shows sample feature vectors of a speech sample. These feature vectors are modeled again using GMM and proposed algorithm with unsupervised learning and Table 4 shows the comparison for Database 2.



(a)

(b)

Fig 3: (a) Female speech, (b) Male speech waveform

Table 1: Sample Feature Vectors of a speech sample in Database1

x1	x2	x3	x4
-15.6007	-27.02	-38.592	-29.9592
10.39731	15.22464	17.39925	13.70034
3.893954	5.267123	4.884622	4.588396
4.795568	7.832418	8.600793	6.660986
2.767083	3.046189	4.119544	5.002978
3.89519	4.090614	5.382687	6.699311
2.528635	3.489308	3.555741	3.39125
4.457545	3.970372	3.152367	3.758308
3.2616	2.94929	4.767323	3.125902
3.553143	1.650853	2.684871	2.221281
2.582132	2.130911	3.888578	2.555095
2.495409	2.81024	3.232979	3.372506
2.584783	2.823473	2.568738	1.991695
2.206982	1.861845	2.176644	2.13453
1.545785	1.221605	1.791948	1.981811
1.038767	1.465766	1.740861	1.615948
1.286646	1.183937	0.936803	1.534526
1.354805	0.998143	1.157701	1.190844
0.771699	0.791281	0.52227	0.426698
0.305086	0.456838	0.564202	0.539112

Table 2: Sample Feature Vectors of a speech sample in Database2

x1	x2	x3	x4
-58.339	-59.3872	-58.5855	-57.8704
-4.7888	-4.48098	-4.67673	-4.29445
0.122317	-0.41278	-0.53994	0.140772
-0.44081	-0.98801	-1.21015	-0.72881
0.195514	-0.93674	-0.10391	0.167644
-0.62024	-0.71034	-0.02153	0.338255
-0.58534	-0.55659	-0.15957	0.25173
-0.69209	-0.67335	-0.53056	0.257806
0.037556	-0.33629	-0.73216	-0.11129
0.430269	-0.26964	-0.90938	0.361586
0.248077	0.146565	0.220096	-0.32232
-0.16225	-0.00923	0.274467	-0.33333
-0.21444	0.114748	-0.16072	0.080245
-0.17947	0.585853	0.301243	0.026033
-0.1058	0.390334	0.406734	-0.1866
0.229122	-0.04406	0.26952	-0.28692
0.355014	0.098887	0.242264	0.116573
0.075672	0.258626	-0.20641	0.008385
-0.22894	-0.31292	-0.08477	0.048503
-0.37524	-0.3588	0.27587	-0.00838

Table 3: Comparison of GMM and H C based GMM based on recognition % with Database 1

No. of Speakers	10	20
GMM	90%	95%
H C based GMM	90%	97.5%

Table 4: Comparison of GMM and H C based GMM based on recognition % with Database 2

No. of Speakers	10	20	30
GMM	80%	82.5%	85%
H C based GMM	80%	85%	88.33%

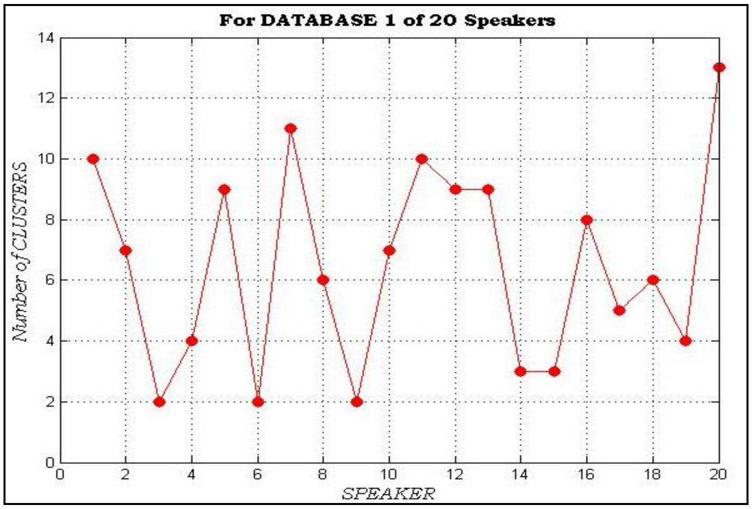


Fig 4: Speaker Vs No. of clusters for Database 1

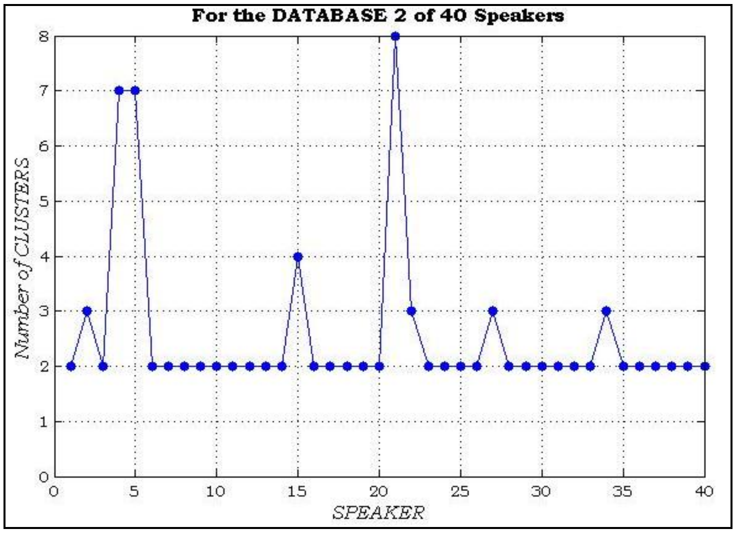


Fig 5: Speaker Vs No. of clusters for Database 2

5. Conclusion

This paper presents the results of recognition of a person based on the speech signals acquired out of different English text spoken though the language is not a barrier to the approach. The frame size of sampled speech signal is 256 samples per frame. The log-spectrum of these frames is wrapped with Mel scale which gives rise to Mel frequency cepstral coefficients serving as features. The authentication is performed using both GMM and Adaptive GMM on the features. Two databases have been used to validate the performance of both the methods. The results show that the performance differs from GMM to Hierarchical clustering based GMM. A recognition rate of 97.5% on database-1 and of 88.33% on database-2 are obtained with the new proposed algorithm whereas 95% on database1 and 85% on database2 with GMM on a controlled database. Further research can be done to reduce the computation time with the proposed algorithm as obtaining the optimal number of clusters is computationally expensive.

6. References

1. Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models, Digital Signal Processing. 2000; 10:19-41.
2. Reynolds DA, Heck LP. Speaker verification: from research to reality, ICASSP Tutorial, Salt Lake City, 2001.
3. Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani, Md. Saifur Rahman. Speaker Identification Using Mel Frequency Cepstral Coefficients, 3rd International Conference on Electrical & Computer Engineering ICECE, Dhaka, Bangladesh. 2004.
4. Lawrence Rabiner, Biing-Hwang Juang. Fundamental of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J, 1993.
5. Reynolds, Douglas Alan. A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification", PhD thesis, Georgia Inst. Of Technology. 1992-1993.
6. Ming-Tao Gan, Madasu Hanmandlu, Ai Hui Tan. From a Gaussian Mixture Model to Additive Fuzzy Systems, IEEE Trans. On Fuzzy Systems. 2005; 13(3):303-316.
7. Azeem MF, Hanmandlu M, Ahmad N. Structure identification of generalized adaptive neuro-fuzzy inference systems, IEEE Transactions on Fuzzy Systems. 2003; 11(5):666-681.
8. Zhehui Liang, Pingjian Zhang, Juanjuan Zhao. Optimization of the Number of Clusters in Fuzzy Clustering, Int. Conf. on Computer Design And Applications (ICCDA). 2010; 3:580-584.