DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4354

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

GAUSSIAN MIXTURE MODEL (GMM) BASED K-MEANS METHOD FOR SPEECH CLUSTERING

K Rajendra Prasad
Dept. of CSE
Institute of Aeronautical Engineering
Dundigal, Hyderabad-500043, India

Abstract: Speech recognition and speaker identification are two important problems in speech clustering. K-means is an efficient clustering method, however it is required to modify this method in speech clustering to the purpose of modeling of speaker voices. In this paper, the pre-processing is performed using HTK and MATLAB Tools. Gaussian Mixture Modeling is used for modeling of speaker voice. In GMM modeling, several parameters are derived for the purpose of defining the shape or structure of a speaker voice. Feature extraction is a process that extracts data from the voice signal that is unique for each speaker. GMM-based k-means method is derived for efficient clustering results and respective results are discussed in experimental section for demonstrating efficiency of proposed method.

Keywords: Speech Recognition, speaker identification, HTK, GMM Modeling, MFCC

I. INTRODUCTION

The speech recognition is an active research area that can be widely used in the application of human machine interaction [2]. Automatic speech recognition is a vital technology that may process and transform speech signals into a sequence of bag of words and it can be used linguistic units of an algorithm, which can be described as a speech program. Some systems are speech understandable systems that are capable of defining whetherspeechunits follows meaning units or not, that framed from vocabularies of thousands of words in speech operational systems. There are two key steps are used in speech processing for recognition of speaker identity, these are (a) speech content and (b) The speaker identity [1]. Aim of speech recognizers [4] is to extract and analyze thelexical speech units and verify whether the speech segment is spoken by particular speaker or not using modeling technique. It is augmented with several speech related applications.

.In a speech clustering, the extraction of speech features is the initial part of the work, and regarding to this we use the HTK [5] tool kit; it extracts the speech data directly into MFCC [7] raw form. The MFCC form is one of the better representations for defining of voice characteristics. Before the speech clustering, we need to give the precise model of speaker voice. Thus, we use a statistical modeling, namely, GMM for the same purpose. GMM have number of mixture components and mainly the parameters of GMM are estimated. The important parameters of GMM are mean and variance. By these parameters, we obtain the shape of speaker voice. UBM is another fashion of GMM, but it is very large GMM, because it was trained by the huge amount of speech data of various channels. The main task in speech clustering is to determine that which utterance is spoken by a particular speaker. This challenging task is effectively done by any clustering algorithm with the help of GMM and UBM modeling concepts.

Every spoken utterance is then estimated statistically by GMM and these statistical values are adapted with UBM parameters by EM algorithm. For every two utterance statistical parameters are mutually compared and finally dissimilarity features are extracted. Especially the statistical parameter of mean of super vectors of utterances is compared either in Euclidean or with Cosine space for obtaining of dissimilarity features. These features are used in the MS algorithm for producing of clustering results. The cosine based MS clustering gives more robust results than Euclidean based MS clustering. But we would improvise the present technique of cosine based MS clustering algorithm by multiple viewpoints in our proposed work. This proposed work is said to be MVS-MS clustering. The expansion of MVS is called as multi-view points based similarity measure. The cosine metric depends on a single viewpoint, MVS is also chosen based metric, but it calculates the similarity features based on multi view points.

II. SPEECH CLUSTERING PROCESS

According to the steps of speech clustering and it can be described as follows: conversion of the speech data from a wav file to MFCC raw form, modeling the speech data by GMM and UBM concepts (generation of super mean-vectors), factor analysis of the data, MS clustering process.

2.1 Conversion of Speech Data to MFCC

The speech feature are very important in speech processing problems. The features are extracted initially from co-efficient of speech segments. These are broadly classified into two types, they are Mel frequency Cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) coefficients. Time-domain representation of speech segment is derived from either MFCC or PLP. In a time discrete representation, we digitized as Mel frequency Cepstral coefficients [8] (MFCCs) which are probably the most commonly practiced in speech spectrum of automatic address recognition (ASR) systems.

MFCC values are directly derived from the Fourier Transform of speech segment. The MFCC always find the perceptual values i.e. It may with more clarity of perceptual sounds for the different speech data from several speakers. Spectrum is required for defining of log of speech frequency and it is followed by inverse Fourier Transform for obtaining of meta spectrum or detailed schema of spectrum. A spectrum gives the knowledge about how the frequencies are shifted for an audio signal. It is particularly applied in audio recognition application and also in speech clustering. The operation for finding of MFCC [8] values are as follows [13]:

Step1:Use the Fast Fourier Transform of speech frame for the speech segment
Step2: Find the coefficients in the form of Mel-Frequency scaling filter groups
Step3: Apply logarithmic of Mel-Frequency groups
Step4: Use the discrete cosine transform for obtaining of Mel-Frequency-Cepstral-Coefficients (MFCC)

Eqn. (1) shows describes the relationship between Mel-Frequency of speech segment and the actual frequency

Mel (f) = 2595 log (1 + f / 700) [13] (1)

MFCC values are denoted the optimal representation of a speaker and it represents the speech segment or speech utterance in the of speech coefficients. It is required to model the shape of speaker voice using his speech segments. In this regard, the Gaussian mixture model is used for defining the shape of speaker voice and it derives the GMM mean-super vector for analyzing of speech clustering concept. Further steps shows that how to apply pre-processing techniques for speech data for removing of noise. These are described as follows

2.2 Mel-Frequency-Cepstral-Coefficients pre-processing steps

Initially the speech data is carried out in the form of .wav files. These .wav files may be converted into MFCC form using pre-emphasis step

a) **Pre-Emphasis Step**;For obtaining of required frequencies, high-pass filter is applied to targeted.wav files, which are achieving to emphasizing the higher frequencies of respective .wav files. The first order with high-pass filter is usually used along with a typical co-efficient value of 0.97

b) ***Framing Step;***The full-length speechutterance is segmentedusing the time-domain frequency levels of thespeechutterance and it can be explored as a time division based fixed segment that normally used in human bodies.In this domain, generally it is noted that the frame duration values will be from 20 ms to 30 ms (usually 25 ms) in many speech related experiments, sometimes we also measure for every 10 ms (thus consecutive 25 ms frames generated every 10 ms will overlap by 15 ms).

c) ***Windowing;***Inwindowing, feature extracted values are multiplied with each frame of a window function. The objective of window function is that it gives smooth results in related speech processing results. Most of the features of either speech utterance or speech segment generally use the spectral concept in the Fourier Transformation technique. It is also show that we get undesirable artifacts even we use the spectral concept in the case of without applying a window.In the respective windowing concepts, we prefer the Hamming window for better results.

In speech stream clustering, distance metrics [2], [3], and [7] are used for measuring the similarity features between any two speech segments.

2.2 Modeling the Speech Data

The modeling of speaker voice is a vital step in the speech recognition. The characteristic and shape of voice is answered with the modeling step. Gaussian Mixture Modeling (GMM) is a robust method for speaker modeling [2]. Initially the voice communication data are gathered and further it is transformed as MFCC form by Hidden Markov Tool (HTK-Tool Kit) [5], and the MFCC data is modeled by GMM parameters [12].

The grandness of the EM algorithm is to accommodate a set of Gaussians to speech data is reported, with issues in interpreting the oral communication using this example. The Gaussian distribution formula will assist in determining the approximated shape of clusters (by Gaussian distributions).

2.3. Gaussian Mixture Model (GMM)[13]

The speech features (or respective MFCC features) are initially modeled by a GMM with 'm' components. However, it is adapted with universal background model (UBM). The UBM may finds the average shape model of human voice. Thus, GMM with UBM is used for producing of reliable GMM means super vectors. This mechanism is greatly succeed in the speech related experimental results and it called as UBM-GMM [9]. The speech features are modeled by GMM as a weighted sum of them-component densities which are in the following equation.

$$p\left({}^{x}\!/_{\lambda}\right) = \sum_{i=1}^{m} wN({}^{X}\!/_{\mu_i}, \Sigma_i) \qquad (2)$$

Where x refers the d-dimensional vector, w_i can be prior probability with condition of$\sum_{i=1}^{m} w_i = 1$. Now, the GMM is defined by its parameters$\lambda = \{w_i, \mu_i, \sum_i\}_{i=1}^{m}$, and the estimation of continuous probability density function (PDF) reduces to finding the proper values of λ.. The following diagram illustrates the GMM [12] modeling process of speaker data. Following figure 1. Shows the illustrative steps of Gaussian mixture modeling of speakers data and it is most commonly used methodology for modeling of speaker data in the real-world speech applications.

.

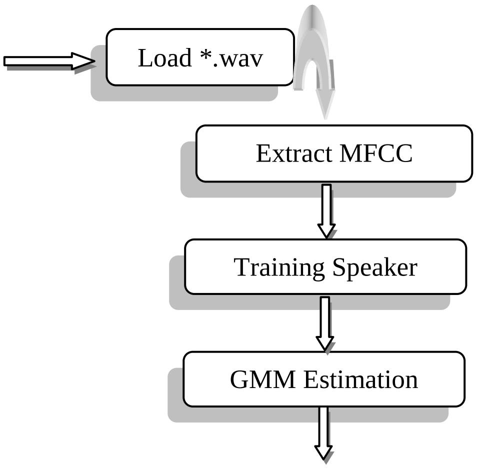


Figure.1 GMM modeling process of speaker data

The UBM is commonly known as large GMM [11] because of it is trained by huge number of speech segments and speech utterances of various speakers. In the experimental section, it is tried with different number of mixture components, like 64-GMM, 128-GMM, 256-GMM, 512-GMM, 1024-GMM etc. In UBM, the speech segments are taken from wide number of sources via a different channels by different speakers.

2.4. Universal Background Model (UBM)

A universal background model (UBM) is average speaker model and it is developed for presenting a speaker independent distribution of the speech features in mean vectors format The mean super vectors are used for the speaker modeling..

2.5. The MAP adaptation Technique

The MAP adaptation is a key step in the initial designing step of universal background model - UBM [4], later, an iterative steps are applied for performing of effective modeling. Key steps are organized as follows: these are E (estimate) step and the M (Maximize) step. The posterior probability of training vectors for a Gaussian component is$p(i/X_t)$.

$$p(i/X_t) = \frac{w_{oi}N(X_t/\mu_{oi},\Sigma_{oi})}{\sum_{j=1}^{m} w_{oj}N(X_t/\mu_{oj},\Sigma_{oj})} \qquad (3)$$

The Eqn. (3) shows the posterior probability that can be used for reassigning of training vector $X_t$of speech segment to the Gaussian mixture component of the UBM model.

2.6 Mean Shift Clustering Process

The Mean Shift rule will be viewed as agglomeration rule or as the way of finding the nodes in a very non-parametric distribution. During this section we are going to gift the intuitive plan behind the Mean Shift mode-seeking method, similarly because the mathematical derivations of this rule. In addition, we tend to gift two variants of this rule which may be applied for agglomeration functions. Finally, the extension of the normal MS to the cosine-based MS is given.

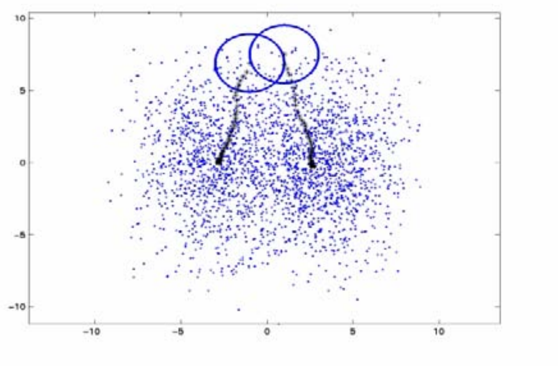


Fig 2: Mean Shift Mode Finding

III. PROPOSED SPEAKER RECOGNITION

The structure of proposed system consists of two modules

- Speaker Identification [6]
- Speech Recognition [2]

3.1 Speaker Identification

Feature extraction is a method that can be used for extraction of speech data from .wav file of speaker. All speech recordings of speaker data are stored in the form of audio file i.e in .wav file.TheMel Frequency Cepstral Coefficient (MFCC) is widely used co-efficient conversion technique that has to create the fingerprint of the either speech segments or speech utterances. The co-efficient values are depends on speaker voice that can handle the problem of known variation speech segments effectively. The MFCC values are characterizing the speaker voice using his speech data. The speaker identification is a primary problem of speech clustering and it can be performed by finding the dissimilarity matrix of speech segments and visually [14],[15],[16] shown how many speaker are really involved in the speech utterance. GMM is used for defining the GMM mean-supervectors of speech segments and these vectors are really obtains the wonderful results in speaker identification problem. Prim's logic is used for re-ordering the indices of dissimilarity matrix of speech segments. Image of re-ordered indices shows the effective speakers of speech utterance.

3.2 Speech Recognition System

Hidden Markov [10] is tool that can be used for processing the statistical features during the conversion of .wav to MFCC, however, initial setup is required for the development of statistical models of speech signals, which characterize the mean and variance properties. Hidden markov tool kit is used for defining HMM that creates the GMM model of speech segments of particular speaker data.
It is the key idea for implementing models of speech segments for every spoken word of speaker. The steps of speech recognition are summarized as follows:

- Markov models are derived for every pronounced spoken word.
- Compute the value of maximum likelihood of every spoken word of speaker during speech processing.

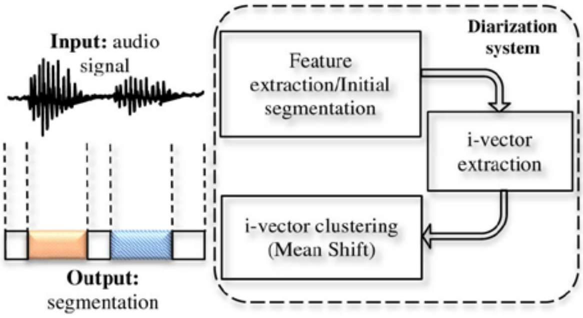
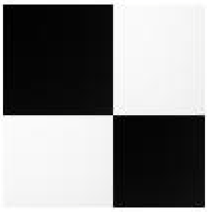


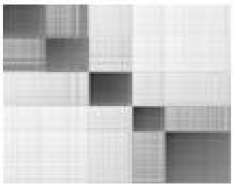
Fig.3 Speech diarization Process

- **Two speaker data will be appear like this:**



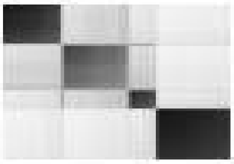
Mean value of two speakers data is 0.942136.

- **Three speakers data will be appear like this:**



Mean value of three speakers data is 0.837044.

- **Four speakers data will be appear like this:**



Mean value of four speakers data is 0.840727.

IV. CONCLUSION

This paper describes the methodology for assessing number of speakers from unlabeled speaker data and it uses the various dissimilarity measures for distinguishing speakers from the respective speech data. GMM is a statistical modeling method that can be define the shape of speech segments by mean parameter. Similarity measures are applied between the mean vectors of speech segments for deriving of similarity features, and then k-means is applied for discovering effective speech clustering results. It is proved that GMM based speech clustering assess the correct number of speakers as well as speech clustering results.

V. REFERENCES

[1]. Ronald M. Baecker, "Readings in human-computer interaction: toward the year 2000", 1995.

[2]. Melanie Pinola, "Speech Recognition Through the Decades: How We Ended Up With Siri", PCWorld.

[3]. Ganesh Tiwari, "Text Prompted Remote Speaker Authentication : Joint Speech and Speaker Recognition/Verification System".

[4]. Dr.RaviSankar, Tanmoy Islam, SrikanthMangayyagari, "Robust Speech/Speaker Recognition Systems".

[5]. BassamA.Q.Al-Qatab and Raja.N.Aninon, "Arabic Speech Recognition using Hidden Markov Model ToolKit (HTK)", IEEE Information Technology (ITSim), 2010,page 557-562.

[6]. AhsanulKabir, Sheikh Mohammad MasudulAhsan,".Vector Quantization in Text Dependent Automatic Speaker Recognition using Mel-Frequency Cepstrum Coefficient", 6th WSEASInternational Conference on circuits, systems, electronics, control & signal processing, Cairo,Egypt, dec 29-31, 2007,page 352-355

[7]. LindasalwaMuda, MumtajBegam and Elamvazuthi.,"Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and DTW Techniques ",Journal of Computing, Volume 2, Issue 3, March 2010

[8]. Mahdi Shaneh and AzizollahTaheri ,"Voice Command Recognition System based on MFCC and VQ Algorithms", World Academy of Science, Engineering and Technology Journal , 2009

[9]. RemziSerdarKurcan, "Isolated word recognition from in-ear microphone data using hidden markov models (hmm)", Master's Thesis, 2006.

[10]. Nikolai Shokhirev ,"Hidden Markov Models ", 2010.

[11]. L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in Speech Recognition", Proceedings of the IEEE Journal, Feb 1989, Vol 77, Issue: 2.

[12]. Suma Swamy, Manasa S, Mani Sharma, Nithya A.S, Roopa K.S and K.V Ramakrishnan, “An Improved Speech Recognition System”, LNICST Springer Journal, 2013.

[13]. Matthew Nicholas Stuttle, “ A Gaussian mixture model spectral representation for speech recognition”,2003

[14]. B Eswara Reddy, K Rajendra Prasad, “ Improving the performance of visualized clustering method”, International Journal of System Assurance Engineering and Management, Springer, 2015

[15]. K Rajendra Prasad, B Eswara Reddy, “ Context-Aware Graph-Based Visualized Clustering Approach”, Advanced Computing and Systems for Security, Springer, pp:247-260,2016

[16]. K Rajendra Prasad, “Assessment of clustering through data visualization methods”, 2015

Author

Dr. K. Rajendra Prasad Graduated in B.Tech(CSE) from Jawaharlal Nehru Technological University, Hyderabad in 1999. He received Masters Degree in M.Tech(CSE) from*Visvesvaraya Technological University***,** *Belgaum* ,in 2004.He received Ph.D in Computer Science & Engineering from JNTUA, Ananthapur, in 2015. Presently, he is working as Professor and Head of CSE Dept., Institute of Aeronautical Engineering, Hyderabad. He has more than 30 Publications in various International Journals and Conferences. He is a life member of CSI, and member of IEEE. His research interests are data mining &data warehousing, and databases.