International Journal of Engineering Research ISSN:2319-6890)(online),2347-5013(print)
Volume No.3 Issue No: Special 2, pp: 125-128 22 March 2014

The Approach of Speaker Diarization by Gaussian Mixture Model (GMM)

K. Rajendra Prasad
Asso.Prof, Dept. of IT
RGMCET, Nandyal

D. Jareena Begum
M.Tech in SE
RGMCET, Nandyal

E. Lingappa
M.Tech in SE
RGMCET, Nandyal

Abstract

Speaker identification is an important activity in the process of speaker diarization. We need to model the speaker by Gaussian mixture model (GMM) for speaker identification purpose. Large GMM is called as a Universal Background Model (UBM) which is adapted into each speaker model for speaker identification purpose. This paper focuses on speech clustering for speaker diarization. The speaker diarization includes the steps speech segmentation and the process of speech clustering. In speech segmentation, the features are extracted for each speech segment which is converted into Mel-Frequency-Cepstral- Coefficients (MFCC). Each speech segment is modeled by UBM adaptation. The relevant speech segments are grouped as speech clusters. This paper describes the speech segmentation, UBM adaptation, and speech clustering technique.

Keywords

GMM, UBM, MFCC, Speaker Diarization

1. Introduction

This paper is to concerns speaker diarization, which includes the problem of annotating an unlabeled audio(*segmentation*) and then finding the different segments of speech belonging to the speaker (*clustering*)[1].

We have seen as the much of the review on the diarization problem [1]–[3]. For its relative simplicity, the Bayesian Information Criterion (BIC) has becomes as a backbone and it gives an inspiration for the development of a number of initial approaches, mainly, includes bottom-up hierarchical clustering [4], [5]. Bottom-up approaches are used in general, where a number of clusters or models are be trained and successively merged until only one remains for each speaker [6],[7]. We also investigated a more integrated, top-down method that success is based on an evaluative Hidden Markov Model (HMM), where detected speakers can help the influence for the detection of other speakers [8],[9].

The HDPs have become well-known in field of Bayesian nonparametric statistics, and the use of Markov Chain Monte Carlo (MCMC) sampling methods have enabled the practical application of these methods to a variety of problems [9] , including diarization. Variational inference is another useful technique for approximate inference that was first applied to the diarization problem [5] and further extended in Diarization of telephone conversations [10]. These methods, alongside the successful application of factor analysis as a front-end for extracting speaker specific features [10],[11] , serve as a basis for much of the work discussed in this paper.

We have seen as the much of the review to developed an approach to diarization based on the successes of factor analysis-based methods in speaker recognition [12],[13], as well as diarization [10],[11]. Inspired by the ability of the Total Variability subspace to extract speaker-specific features on short segments of speech [13],[14].

The success achieved [12], however, was limited to the task in which we knew there were exactly two speakers in the given conversation. To solve the diarization problem in general, we must address the setting in which the number of participating speakers is unknown *a priori*. First, we motivated the use of a spectral clustering algorithm as an alternative to the previous approach involving K-means clustering based on the cosine distance. More importantly, we adapted a heuristic from previous work applying spectral clustering to diarization and used it to determine the number of clusters. Second, we verified that there exists a symbiotic relationship between clustering and segmentation; that is, better initial segmentations yield better speaker clusters, and conversely, better speaker clusters aid in providing cleaner speaker segments.

We taken the number of attempts at using factor analysis based methods for speaker diarization. The inspirations for our current saga[10],[11], also independently led to the work in[17], which uses PCA and K-means for two-speaker diarization in a way similar to our methods in[12]. Factor analysis-based features are used in [18] alongside the Cross Likelihood Ratio as a criterion for hierarchical clustering, while[19] performs clustering using PLDA as inspired by its recent success in speaker verification.

2. Acoustic Features

We first assume that the incoming audio has been transformed into a sequence of acoustic feature vectors. Specifically, we use raw Mel-Frequency Cepstral Coefficients (MFCCs) extracted every 10 ms over a 25 ms window. These MFCCs are 20-dimensional vectors and are the basis for our subsequent modeling. In practice, a number of variants can be used; for example, many speaker recognition systems also include first and second derivatives into their feature vector, cepstral mean subtraction, as well as a Gaussianization feature warping step that can normalize for short-term channel effects. However, in order to follow the footsteps of previous work as closely as possible, we limit our consideration to just the use of raw cepstral features. The rest of this paper assumes that all audio has been transformed into a sequence of acoustic feature vectors.

MFCC values are derived from the Fourier Transform of an audio stream. The Mel frequency always determines the

perceptual weighting i.e. it gives more clarity about the perceptual sounds such as different speech data from various speakers.

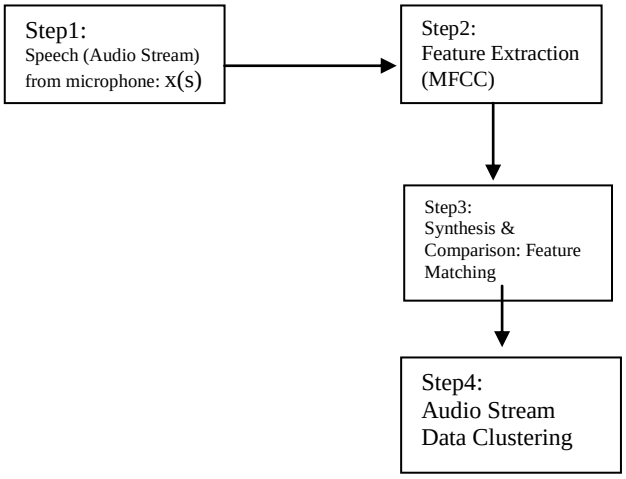

Figure 1: Outline of Speech Data Processing

Spectrum nothing but is the log followed by inverse Fourier Transform or it is the spectrum of a spectrum. A spectrum gives the knowledge about how the frequencies are changed for a audio signal. It is especially used in audio recognition application and also in speech clustering. The procedure for finding of MFCC values are as follows:

Step1: Apply the Fast Fourier Transform for speech frame (speech segment window)

Step2: Find Mel-Frequency scaling filter groups

Step3: Take the logarithm for Mel-Filter

Step4: Further, apply discrete cosine transform for obtaining of MFCC

The relationship between Mel-Frequency and the actual frequency can be denoted by the equation (1).

$$\text{Mel}(f) = 2595 \log (1 + f / 700) \quad (1)$$

MFCC values are denotes the optimum representation of a speaker and it retrieves the speech coefficients (or the coefficients of audio stream data). After extracting the MFCC's we will find Gaussian mixture model (GMM) of various trends (or speech by different speakers). Gaussian distribution formula will help in determining the approximated shape of clusters (by Gaussian distributions).

2.1 MFCC pre-processing steps

The data of speech can be read out from the respective wav file and we follow the some MFCC pre-processing steps. These are described as follows:

a) Pre-Emphasis Step

The high-pass filter is applied on wave file data for emphasizing of the higher frequencies and these are to be compensates for the human speech which tends be high frequencies. The first order with high-pass filter is commonly used along with a typical co-efficient value of 0.97

b) Framing Step

e divide the full-length audio segment according to the time-domain for the utterance can be considered as a division fixed duration segments called as *frames*. Generally we follow the frame duration values will be from 20 ms to 30 ms (usually 25 ms) and this is a frame for generating for every 10 ms (thus consecutive 25 ms frames generated every 10 ms will overlap by 15 ms).

c) Windowing

In feature extraction we multiply each frame by a window function. The main purpose of window function is that it gives the smooth effect of results. All features are generally a spectral in the form of Fourier Transform. Without using a specified window will have the undesirable spectral artifacts and the technique of Hamming window becomes the most popular. Therefore, we prefer the Hamming window technique in windowing.

In data stream clustering, we use distance metrics [20], [22], and [23] after projecting the speech stream data into a high-dimensional space to extracting of dissimilarity features between different utterances.

3. The Total Variability Approach

To enhance the classical method of modeling speakers using Gaussian Mixture Models (GMMs)[29], recently developed methods apply factor analysis to supervectors a vector consisting of stacked mean vectors from a GMM in order to better represent speaker variabilities and compensate for channel inconsistencies[13]. One such approach is Total Variability, which decomposes a speaker and session dependent supervector M as

$$M = m + T w \quad (1)$$

Where 'm' is still the speaker- and session-independent supervector taken from the Universal Background Model (UBM), which is a large GMM trained to represent the speaker-independent distribution of acoustic features. 'T' is a rectangular matrix of low rank that defines the new total variability space and is a low-dimensional random vector with a normally distributed prior N(0,1). The remaining variabilities not captured by 'T' are accounted for in a diagonal covariance matrix, the vector 'w' can be referred to as a "total factor vector" or an i-vector, short for "Intermediate Vectors" for their intermediate representation between an acoustic feature vector and a supervector.

One way to interpret (1) is to see the columns of 'T' as a limited set of directions from which M can deviate from m, the latter of which is a starting point, or bias, taken from the UBM. Ultimately, for some utterance u_S, its associated i-vector w_S can be seen as a low-dimensional summary of the speaker's distribution of acoustic features with respect to the UBM.

To avoid getting bogged down in the mathematics, we omit the details regarding the training and estimation of 'T'

and 'W' via an Expectation Maximization (EM) algorithm. A thorough treatment can be found in [16]. For convenience throughout the rest of this paper, we use the term "i-vector extraction" to denote estimation of the posterior distribution of 'w' (mean and covariance). Moreover, the term "i-vector" refers specifically to the posterior mean of 'w', while "i-vector covariance" will refer to its posterior covariance. Lastly, the cosine similarity metric has been applied successfully in the Total Variability subspace to compare two speaker i-vectors [13]. Given any two i-vectors 'w1' and 'w2', the cosine similarity score is given as

$$\cos\_score(w_1, w_2) = \frac{(w_1)^t (w_2)}{\|w_1\| \cdot \|w_2\|}$$

4. Analysis of Speaker Clustering

The GMM-UBM training is performed by MAP adaptation technique [21] of the means, from this adapted model, finally we form a GMM supervector. A universal background model (UBM) is a speaker-independent model. It has represents a speaker independent distribution of the feature vectors: these vectors are used to form the model. It is trained with a large amount of speech data from a set of speakers, using the algorithm of EM.

K-means clustering using the cosine distance is capable of achieving good clustering results on conversations containing any number of speakers [12],[15],[16]. Unfortunately, K-means requires as input the number of clusters, we adapted the use of a heuristic to estimate the number of speakers in a conversation by using a spectral clustering method, which analyzes the eigen-structure of an affinity matrix. This technique gave reasonable performance; however, its success as a heuristic only served to further inspire the development of a more principled approach. The explorations of touched upon the use of Bayesian model selection as an analog for determining the number of speakers in a conversation. Bayesian methods

have the advantage of naturally preferring simpler models for explaining data. At least in theory, they are not subject to the over fitting problems which maximum likelihood method.

The clustering stage involves grouping the previously-extracted segment i-vectors together in such a way that one cluster contains all the segments spoken by a particular speaker. And unless given *a priori*, the number of speakers (clusters) 'K' must also be determined at this stage. Because it is known that we are strictly diarizing conversations (involving two or more participants), we require that $\hat{K} \geq 2$, where $\hat{K}$ is our estimate of 'K' .There exist many different ways to perform clustering.

Given a set of segments with associated cluster labels, we use the exact same re-segmentation algorithm discussed in both [10], [12] to refine our initial segmentation boundaries. At the acoustic feature level, this stage initializes a 32-mixture GMM for each of the 'K+1' clusters (Speakers $\{S_1, \ldots, S_K\}$ and non speech NS) defined by the previous clustering. Posterior probabilities for each cluster are then calculated given each feature vector z_t (i.e., $P(S_1|z_t), \ldots, P(S_K|z_t), P(\bar{N}S|z_t)$). Pooling these across the entire conversation provides a set of weighted Baum-Welch statistics from which we can re-estimate each respective speaker's GMM. To prevent this unsupervised procedure from going out of control, the non-speech GMM is never re-trained.

We can further refine the diarization output by extracting a single i-vector for each respective speaker using the (newly-defined) segmentation assignments. The i-vector corresponding to each segment (also newly extracted) is then re-assigned to the speaker whose i-vector is closer in cosine similarity. We iterate this procedure until convergence when the segment assignments no longer change. This can be seen as a variant of K-means clustering, where the "means" are computed according to the process of i-vector estimation.

TABLE 1A: Speaker Recognition in Speech Clustering (Using Distance formula)

Train/Test	T1	T2	T3
S1	0.7073	2.9367	5.1568
S2	2.8871	0.7476	6.9923
S3	2.4129	3.3374	0.9776

TABLE 1B: Speaker Recognition in Speech Clustering (Using kull-lieber divergence test)

Train/Test	T1	T2	T3
S1	0.5156	1.4436	3.3220
S2	0.7943	0.3932	2.5177
S3	1.4262	2.0819	0.7288

TABLE 1C: Speaker Recognition in Speech Clustering (Using Gaussian likelihood measure)

Train/Test	T1	T2	T3
S1	0.3910	0.5779	1.5182
S2	0.5298	0.2920	1.2698
S3	0.9319	0.9146	0.4052

TABLE 1D: Speaker Recognition in Speech Clustering (Using Dynamic Time Warping)

Train/Test	T1	T2	T3
S1	0.5366	1.5967	2.6127
S2	2.8425	0.6779	6.5741
S3	1.8984	2.6358	0.9506

Speech recognition experiments are carried out by proposed technique and results are presented in the tables (TABLE1A, TABLE1B, TABLE1C, TABLE1D) .The training speakers are defined in the rows. The testing data of same sequence of three speakers are taken column wise. The cells (S1,T1), (S2,T2) , (S3,T3) have lowest values in their rows compared to other values in the rows. Thus speaker identification is done perfectly by GMM modeling concept.

5. Conclusion

The sample results are presented for speaker identification. These results are helpful in speech diarization for deriving of speech clustering results. This paper explains the clear approach of speech clustering by GMM, and UBM concepts. GMM is a finest statistical model for representing the speaker in a high-dimensional space. By the PCA concept we denote each speaker utterance in a two-dimensional space. Hence, the Euclidean space is used for classification of speakers in an unsupervised manner which is the significant step of speaker diarization.

References

i. *S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 5, pp. 1557–1565, Sep. 2006.*

ii. *M. H. Moattar and M. M. Homayounpour, "A review on speaker diarization systems and approaches," Speech Commun., vol. 54, no. 10, pp. 1065–1103, 2012.*

iii. *X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 2, pp. 356–370, Feb. 2012.*

iv. *D. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations," in Proc. NIST Rich Transcript. Workshop, 2004.*

v. *F. Valente, "Variational Bayesian methods for audio indexing,"Ph.D. dissertation, Univ. De Nice-Sophia Antipolis—UFR Sciences, Nice, France, Sep. 2005.*

vi. *X. Anguera, C. Wooters, and J. M. Pardo, "Robust speaker diarization for meetings: ICSI RT06's evaluation system," in Proc. ICSLP, 2006.*

vii. *T. H. Nguyen, H. Sun, S. Zhao, S. Z. K. Khine, H. D. Tran, T. L. N. Ma, B. Ma, E. S. Chng, and H. Li, "The IIR-NTU speaker diarization systems for rt 2009," in Proc. RT'09, NIST Rich Transcript.Workshop, 2009.*

viii. *S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," Comput. Speech Lang., vol. 20, no. 2, pp. 303–330, Jul.2006.*

ix. *D. Blei andM. Jordan, "Variational inference for dirichlet process mixtures," Bayesian Anal., vol. 1, no. 1, pp. 121–144, 2006.*

x. *P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," IEEE J. Sel. Topics Signal Process. vol. 4, no. 6, pp. 1059–1070, Dec. 2010.*

xi. *F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Streambased speaker segmentation using speaker factors and eigenvoices," in Proc. ICASSP, 2008, pp. 4133–4136.*

xii. *S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting intra-conversation variability for speaker diarization," in Proc. Interspeech, 2011.*

xiii. *N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 4, pp. 788–798, Jul. 2010.*

xiv. *S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," in Proc. IEEE Odyssey, 2010.*

xv. *S. Shum, N. Dehak, and J. Glass, "On the use of spectral and iterative methods for speaker diarization," in Proc. Interspeech, 2012.*

xvi. *S. Shum, "Unsupervised methods for speaker diarization,"M.S. thesis, Mass. Inst. of Technol., Cambridge, MA, USA, Jun. 2011.*

xvii. *C. Vaquero, A. Ortega, and E. Lleida, "Intra-session variability compensation and a hypothesis generation and selection strategy for speaker segmentation," in Proc. ICASSP, 2011, pp. 4532–4535.*

xviii. *D. Wang, R. Vogt, S. Sridharan, and D. Dean, "Cross likelihood ratio based speaker clustering using eigenvoice models," in Proc. Interspeech, 2011.*

xix. *J. Prazak and J. Silovsky, "Speaker diarization using PLDA-based speaker clustering," in Proc. IDAACS, 2011.*

xx. *P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 4, pp. 1448–1460,May 2007.*

xxi. *M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Univ. College London, London, U.K., May 2003.*

xxii. *E. Khan, J. Bronson, and K. Murphy, Variational Bayesian EM for Gaussian Mixture Models. 2008 [Online]. Available: http://www.cs. ubc.ca/~murphyk/Software/VBEMGMM/index.h ml*

xxiii. *M. J. Johnson, PYHSMM: A Python Library for Bayesian Inference in (HDP-)H(S)MMS. 2010.*