Speaker Diarization Using Gaussian Mixture Turns and Segment Matching

Arlindo Veiga[1,2], *Carla Lopes*[1,2], *Fernando Perdigão*[1,2]

[1] Department of Electrical and Computer Engineering, University of Coimbra
[2] Instituto de Telecomunicações – Pólo de Coimbra
{aveiga, calopes, fp}@co.it.pt

Abstract

Speaker diarization aims to detect "who spoke when" in large audio segments. It is an important task in processing of broadcast news audio, making easier the audio segments selection and indexing task. In this paper an unsupervised speaker diarization scheme is proposed using a Gaussian Mixture Model as a Universal Background Model, Bayesian Information Criterion and fingerprint detection. A decoder that outputs a mixture sequence is used with high mixture transition penalization. Homogeneous segments tend to produce sequences with only one mixture allowing speaker turns to be detected using mixture transitions. Results for the Catalan broadcast news 3/24 TV channel are reported.

Index Terms: speaker diarization, audio fingerprint, GMM, BIC

1. Introduction

Gaussian Mixture Model (GMM) has the ability to model arbitrary densities distributions. They have shown to have excellent performance in speech modelling tasks. Many speaker recognition systems use a large GMM called Universal Background Model (UBM) [1] to represent all speaker distribution features and to adapt it to speaker's models. The Expectation Maximization (EM) is a well-established algorithm that is used to estimate the GMM parameters.

Since the speaker diarization contest relies on the assumption that do not exist prior knowledge about the speakers, the use of speaker models is not allowed. However UBM can still be used for unsupervised speaker segmentation and classification considering each mixture as a cluster. A decoder that outputs a mixture sequence was used with a high mixture transition penalization. Homogeneous segments should result in a single mixture sequence in most of the time. The ALIZE/Mistral project [2] proposes a system for speaker diarization that use a similar decoder for speaker clustering. Furthermore, the so called Bayesian Information Criterion (BIC) [3] is used in most proposed approach for unsupervised speaker clustering because it requires only one training parameter estimation.

Before speaker clustering, and in order to deal with non-speech audio segments, a segmentation system is used to detect silence, and music. The fingerprint system [4] is used to ensure that similar segments have same label. Acoustic or audio fingerprinting refers to a condensed representation of an audio signal that can be used to identify an audio sample or quickly locate similar items in audio streams. A binary representation of spectral patterns computed by the convolution of spectrogram with a mask is used. With this technique we can quickly and easily discover repeated segments with high confidence. To deal with classification inconsistencies in repeated segments, we use majority vote to determine which label should be used. In speaker diarization we use also this technique to set the same speaker to similar audio segments.

Finally, the BIC algorithm is used to cluster homogeneous segments. This step performs a substantial increase in the systems' performance.

2. System description

An audio segmentation system is used to segment audio files into five classes: clean speech, music, speech with music in background, speech with noise in background and others non-speech events (silence and noises). Non-speech segments are discarded and all speech segments are labelled with different speaker identification, creating one cluster per speech segment. To merge similar clusters, a UBM with 256 components is trained with the entire audio session. Then, a model based clustering method using this model is applied by means of a mixture sequence decoding. This will be explained in more detail in section 4.
. The result of the decoding is a sequence of segments with speaker labels. The fingerprinting system is then used to search for label inconsistencies in similar segments. Each segment is used as a "jingle" to verify if there is another equal segment in the audio session. In that case the speaker labels are made equal.

A final verification is applied to the labelled segments based on BIC. A pair-wise BIC difference (ΔBIC) between each cluster is computed and the decision to merge clusters is based on this value.

Figure 1 describes the speaker diarization system.

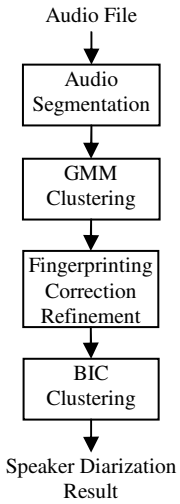

Figure 1: *Speaker diarization block diagram.*

The audio segmentation system used here is the same presented on audio segmentation contest.

3. Features

Audio features are computed every 100ms within a 200ms Hamming window, resulting on a frame-rate of 10 frames per second. The dimension of the feature vector is 48, corresponding to 24 static parameters and their first order derivatives. The static parameters consists of 16 Mel frequency cepstral coefficients with more 8 parameters shown in Table1.

Table 1. *Acoustic Feature Set used in combination with standard MFCC Features.*

Number	Feature description
1	Frame Energy in dB
2	Zero Crossing Rate
3	Spectral Centroid
4	Spectral roll-off - 90%
5	Max normalized correlation coefficient in each frame
6	frequency of the max normalized correlation coefficient
7	Harmonicity measure (proportional to the harmonic duration)
8	Spectral Flux

Audio segmentation and UBM training uses all the 48 features of the acoustic vector while BIC clustering uses only MFCC features (16). In most papers 12 MFCC have proved to be very efficient for speaker turns task. In the [5] the authors report that the use of first order derivative coefficients deteriorates the system performance. The BIC value depends of dimensions of feature vector.

Audio fingerprinting system use a binary representation of spectral patterns and is computed every 20ms within 240ms Hamming window resulting 50 frames per second. Binary representation of each frame is saved on 32 bits integer and searching similar segments is based on *xor* operations and bit counts.

4. GMM Clustering

HTK [6] tools were used to initialize and train each session's background model. The number of mixture used was 256 because the number of speakers per recording ranges from 30 up to 250. It is expected to have one dominant mixture per speaker.

A simple decoder was used in order to find the best mixture sequence, given the audio features. It corresponds to a Viterbi decoder with 256 states in parallel, each one using a single mixture Gaussian of the UBM and a high transition penalization between different states. Homogeneous segments uttered by the same speaker tend to have the same mixture Gaussian during most of the segment time. So, given two segments, it is possible to estimate if they are uttered by the same speaker analyzing the decoding results of a concatenation of these two segments. Figure 2, illustrate a pair of segments and the mixture sequence. The algorithm computes the mean mixture index in the first (fixed) segment and in all other forward segments. If the difference between them means is low, the second segment becomes with the label of the first one.

The resultant mixture sequence is highly dependent on the transition penalty used in the decoding process. Using just a simple mean criterion implies the use of a high mixture transition penalty (in this case a value of -100 was used as a logarithmic weight).

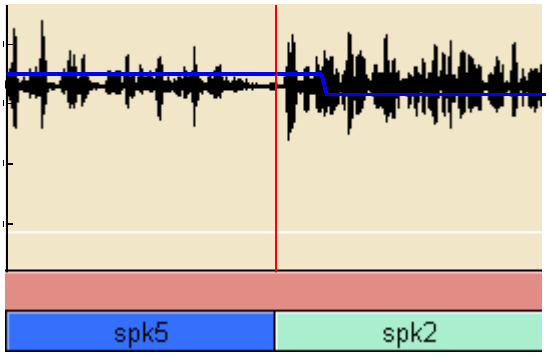

Figure 2: *Two concatenated segments and mixture sequence.*

It is possible to use different criteria to analyze the mixture sequence in each segment in order to decide if they are uttered by the same speaker. In future work a more robust criterion will be searched for.

The last step consists in finding similar segments using the fingerprinting technique. If two segments are found to be repeated, then the same label is set for both segments. This corrects inconsistencies on labelled segments.

5. BIC Clustering

To reduce the GMM clustering diarization error rate (DER), another clustering algorithm is used, based on BIC [3] computation. Using speaker turn detention approach, ΔBIC requires tuning only one parameter (λ). Considering two segments, X_1 and X_2, each one represented by a single Gaussian, $X_1 \sim N(x;\boldsymbol{\mu}_{X1},\boldsymbol{\Sigma}_{X1})$ and $X_2 \sim N(x;\boldsymbol{\mu}_{X2},\boldsymbol{\Sigma}_{X2})$, a concatenated segment X can still be represented by a single Gaussian $X \sim N(x;\boldsymbol{\mu}_X,\boldsymbol{\Sigma}_X)$ if X_1 and X_2 are similar segments (uttered by the same speaker). This results in a positive ΔBIC value, defined as [3,5]:

$$\Delta BIC = -R + \lambda P \tag{1}$$

where:

$$R = \tfrac{N_X}{2}\log\left(|\Sigma_X|\right) - \tfrac{N_{X1}}{2}\log\left(|\Sigma_{X1}|\right) - \tfrac{N_{X2}}{2}\log\left(|\Sigma_{X2}|\right), \tag{2}$$

$$P = \frac{1}{2}\left(p + \frac{1}{2}p(p+1)\right) \tag{3}$$

λ is an empiric parameter normally with a value between 0.5 and 2. In this work 0.6 was used after tuning with the some reference sessions.

6. Results

The DER on the test database of the contest is 55.84 %. The time to diarize the test sessions was around 4 hours, about half an hour per session. Most of time is spent in the Viterbi algorithm. The code was mainly implemented in Matlab (with optimizations in C++ mex) and the used machine was a Cray CX1.

7. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, no. 1, pp. 19-41, Jan. 2000.

[2] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a Free Toolkit for Speaker Recognition," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2005), pp. 737-740, 2005.

[3] S. S. Chen and P. S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 127-132, 1998.

[4] C. Neves, A. Veiga, L. Sá, and F. Perdigão, "Audio Fingerprinting System for Broadcast Streams," in Proceedings of the Conference on Telecommunications (ConfTele2009), pp. 481-484, 2009.

[5] P. Delacourt and C. J. Wellekens, "DISTBIC: A Speaker-based Segmentation for Audio Data Indexing," Speech Communication, vol. 32, no. 1, pp. 111-126, Sep. 2000.

[6] S. Young et al., The HTK Book (for HTK Version 3.4). Cambridge, U. K: Cambridge University Engineering Department, 2006.