ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation

Javier Franco-Pedroso, Ignacio Lopez-Moreno, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
{javier.franco, ignacio.lopez, doroteo.torre, joaquin.gonzalez}@uam.es

Abstract

This paper describes the ATVS-UAM systems submitted to the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation. The ATVS-UAM audio segmentation system is based on a 5-GMM-MMI-state HMM model. Testing utterances are aligned with the model by means of the Viterbi algorithm. Spurious changes in the state sequence were removed by mode-filtering step. Finally, too sort segments were removed. The ATVS-UAM speaker diarization system is a novelty approach based on the cosine distance clustering of the Total Variability speech factors -the so-called iVectors- performed in two steps, followed by a Viterbi decodification of the probabilities based on the distances between the candidate speaker centroids and the iVectors stream.

Index Terms: audio segmentation, speaker diarization, viterbi, factor analysis, maximum mutual information.

1. Introduction

In the recent years the speaker and language recognition community dedicates special attention to the *real conditions* challenge. This challenge involves audio recordings preceding from different sources in addition a single speaker, such as noise, channel effects, speech or music. Speaker turns in a conversation also causes significant degradation in performance for poor segmentations. Such challenge motivates the ATVS-UAM participation in Albayzin 2010. Recently, Factor Analysis (FA) methods have shown excellent results facing some of these problems such as the compensation of the channel and speaker variability. Moreover, FA is currently the state-of-the-art technology for speaker and language recognition, with promising results in other fields such as speech recognition. A successfully FA scheme for speaker diarization was firstly proposed by Castalado [1] in 2008 and later extended in [2]. Castaldo uses low dimensional speaker vectors that are obtained over highly overlapped windows of one-second length. Thus FA generalizes as a secondary parameterization of the input speech stream. This new short-term speaker-factors space shows excellent results when classical speaker diarization techniques are applied on it. In [3] Najim and Kenny enhances the classical FA scheme by: a) Modeling together speaker and channel variability, in what is called total variability. Additional improvements can be achieved with a discriminative training of the target classes such as Linear Discriminant Analysis (LDA) [4] and b) Estimating the posterior probabilities of a speaker participating in the conversation as the cosine distance between the averaged iVectors over the training and testing utterances [4].

Other concerns that have been addressed during the design of the ATVS-UAM Audio Segmentation System were the use of features that includes information of the time dependency structure of the speech, such as Shifted Delta Cepstral coefficients (SDC) [5] and the usage of Maximum Mutual Information (MMI) [6] to improve the discrimination rate while maximizing the mutual information between acoustic classes. In multi-class problems such as Language Recognition or even Speech Recognition, GMM-MMI and HMM-MMI models have shown notable discrimination improvements, also motivating their usage for this submission.

The rest of the paper is organized as follows. Section 2 describes feature extraction for each system. Then, we describe system details for each evaluation task, audio segmentation (Section 3) and speaker diarization (Section 4). Finally, conclusions are presented in Section 5.

2. Feature extraction

2.1. Audio segmentation

Audio Segmentation parameterization consists in 7 MFCC with CMN-Rasta-Warping concatenated to their 7-1-3-7 Shifted Delta Coefficients (SDC).

SDC features have been widely used in Language Recognition due to the fact that they capture the time dependency structure of the language better than the speed or acceleration coefficients (also known as delta and delta-delta). Similarly, SDC features are expected to distinguish the time dependency of the speech over the music or noise.

2.2. Speaker diarization

The front-end parameterization for speaker diarization is illustrated in the Figure 1. It follows a classical Speaker Recognition recipe: 19 MFCC coefficients concatenated to their deltas and followed by Cepstral Mean Normalization (CMN), RASTA filtering and feature warping.

All the training data labelled as ‘speech’, ‘speech with noise in background’ and ‘speech with music in background’ is used to train a 1024-mixtures UBM model. Given this UBM, sufficient stats are extracted for every labeled segment. The total variability subspace is then modeled following the FA recipe. The next step is to compute a LDA matrix that discriminates among speakers. Such matrix is trained with the speaker labels provided and compensated statistics, called iVectors.

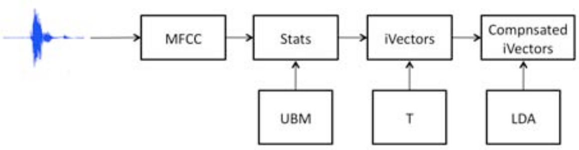

Figure 1: Schematic diagram of the feature extraction scheme for speaker diarization.

As in [1] our back-end parameterization computes iVectors every 20ms over a one second length window. Resulting iVectors are projected over the space defined by the LDA matrix.

3. Audio segmentation system

The ATVS-UAM-UAM audio segmentation system is illustrated in the Figure 2. It is based on the Viterbi alignment of the audio stream using a five-state HMM. One for each target acoustic class: 'speech', 'speech with noise in background', 'speech with music in background' 'music' and 'others'.

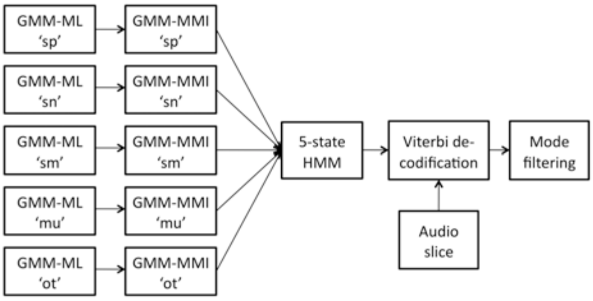

Figure 2: Schematic diagram of the ATVS-UAM audio segmentation system

Each HMM state consists in a 1024 mixtures GMM, previously trained by means of 5 iterations of the Maximum-Likelihood criterion, and enhanced later by means of 18 iterations of the Maximum Mutual Information criterion. This latter step were carried out using the HMM Toolkit STK software from BUT Speech@FIT (Brno University of Technology, Faculty of Information Technology) [7]. All development data provided for the evaluation were used to train these GMMs and no additional data were used.

The SDC features stream is previously divided into 60 seconds length audio slices that are independently processed. Initial 2 seconds of each slice are overlapped with the previous one.

Viterbi alignment is performed using the HMM Toolbox for Matlab by Kevin Murphy [8].

After the Viterbi decodification, a mode-filtering step over a 700 ms sliding window is used to avoid spurious changes between states. Finally, for each class, very short segments were removed –those ones with length smaller than around 3 seconds.

Table 1 summarizes ATVS-UAM audio segmentation system testing timing.

Table 1: *Breakdown timing for ATVS-UAM audio segmentation system.*

Testing (per 4 hours session file)	
Feature extraction	14 minutes
Viterbi decodification + mode-filtering	20 hours

4. Speaker diarization system

ATVS-UAM speaker diarization system (Figure 3) is based on the previous works [2] and [3].

The MFCC features stream is firstly divided into 90 seconds length audio slices –contiguous windows are 33% overlaped-. Compensated iVectors in each slice are clusterized based on their cosine distance. The number of clusters is controlled by maximum allowed distance between the vectors to the centroid of the cluster. In our implementation we used as centroid the averaged vector in each cluster and it represents a candidate speakers model. Candidate speaker models are accumulated over all the slices in the test session, together with the frequency of appearance of their cluster. Since speakers are expected to appear in more than one slice, a secondary clustering is used to merge the first iteration centroids, obtaining then an enhanced set of candidate speakers. A prior probability is assigned to each of the candidate speakers based on its relative frequency of appearance in the entire session.

In a second pass over the slices we compute the probability of each candidate speaker with the stream of iVectors. Such probability is estimated using the cosine distance and normalized with the prior probability of each candidate speaker. The final diarization labels are obtained with a Viterbi decodification of these scores.

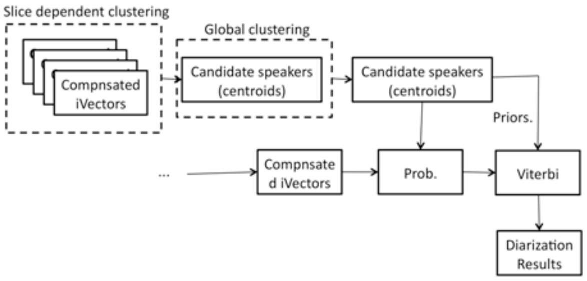

Figure 3.Schematic diagram of the ATVS-UAM speaker diarization system.

Table 2 summarizes ATVS-UAM speaker diarization system testing timing.

Table 2: *Breakdown timing for ATVS-UAM speaker diarization system.*

Testing (per 4 hours session file)	
Feature extraction	40 minutes
iVectors computation	32 hours
iVectors clustering + Viterbi decodification	15 minutes

5. Conclusions

This paper summarizes the ATVS-UAM participation in Albayzin 2010 Evaluations. ATVS-UAM submits results for two of the four proposed evaluations: Audio Segementation and Speaker Diarization. In the latest case we present a novelty approach based on FA to model the total variability subspace. The so-computed iVectors are clustered based on an estimation of the likelihood using cosine distance. Thus, centroids to each cluster can be considered candidate speakers. Likelihoods for each candidate speakers are computed in a second pass over the iVector stream. The final sequence of decisions is computed using the Viterbi algorithm. The ATVS-UAM Audio Segmentation system submitted is based on a five states HMM, each of them trained independently with a 1024 gaussians GMM using MMI. The final sequence of decisions is obtained as an enhanced Viterbi decodification.

6. References

[1] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, “Stream-based Speaker Segmentation Using Speaker Factors and Eigenvoices,” in Proc. ICASSP, Las Vegas, Nevada, Mar. 2008, pp. 4133 – 4136.

[2] P. Kenny, D. Reynolds, F. Castaldo “Diarization on Telephone Conversation using Factor Analysis”. IEEE Journal on Selected Topics In Signal Processing. 2010.

[3] Najim Dehak, Patrick Kenny, Rda Dehak, Pierre Ouellet, and Pierre Dumouchel, “Front end Factor Analysis for Speaker Verification,” IEEE Transactions on Audio, Speech and Language Processing, 2010.

[4] Najim Dehak, Reda Dehak, James Glass, Douglas Reynolds, and Patrick Kenny, “Cosine similarity scoring without score normalization techniques,” in Odyssey, 2010

[5] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller, Jr., “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features.” *Proc. ICSLP 2002*, Sept. 2002, pp. 89-92.

[6] P. Matejka, L. Burget, P. Sckwarz, J Cernocky. “Brno University of Technology System for NIST 2005 Language Recognition Evaluation”, in Proceedings of Odyssey 2006. Puerto Rico.

[7] http://speech.fit.vutbr.cz/

[8] http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.htm