VIVOLAB-UZ Speaker Diarization System for the Albayzin 2010 Evaluation Campaign

Carlos Vaquero, Alfonso Ortega, Eduardo Lleida

Communications Technology Group (GTC)
Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
{cvaquero,ortega,lleida}@unizar.es

Abstract

This paper describes the speaker diarization systems proposed by the VIVOLAB-UZ group for the Albayzin 2010 speaker diarization evaluation. Our approaches combine recent improvements in the field of speaker segmentation in two speaker telephone conversations, using eigenvoice modeling, with the traditional Agglomerative Hierarchical Clustering approach. We are presenting two submissions. Our first system uses a simple eigenvoice factor analysis model to extract a stream of speaker factors for every recording that enable better speaker separability. The speaker factor stream is used for speaker segmentation. Then, both the clusters obtained are agglomerated using Bayesian Information Criterion as distance metric, obtaining the speaker labels. Our second submission is exactly the same system, but it uses Viterbi resegmentation to refine speaker change points as a final step.

Index Terms: Speaker diarization, Factor Analysis, intrasession variability, Agglomerative Hierarchical Clustering, Bayesian Information Criterion

1. Introduction

The main breakthroughs in the field of speaker diarization have been introduced this decade, in part due to the NIST Rich Transcriptions (RT) evaluations. From 2002, NIST has coordinated several Rich Transcription Evaluations aiming at extracting information from audio recordings such as speaker turns or speech transcriptions. All these evaluations involved a Speaker Diarization task, that has become the framework for research and development of the state of the art speaker diarization technologies. In the beginning, the environment to evaluate speaker diarization approaches in the RT framework were telephone conversations and broadcast news. From 2005 the evaluation has focused on meetings. The current Albayzin evaluation does not differ much from those RT evaluations on broadcast news.

Most of the best performing systems presented in the RT evaluations are based on Agglomerative Hierarchical Clustering, that is, after a first segmentation, that gives a set of clusters, the system performs a bottom-up clustering until a stopping criterion is met [1]. Usually, Viterbi resegmentations are performed every time two clusters are merged, and several criteria are used for cluster merging and as stopping criterion. One of the most widespread is the Bayesian Information Criterion (BIC), that has shown to perform well for both cluster merging and stopping criterion.

On the other hand, recently, there has been a great advance in the field of speaker identification, in part motivated by the NIST Speaker Recognition Evaluations (SRE). One of the main breakthroughs of the last years has been the formulation of the Joint Factor Analysis (JFA) for speaker verification [2]. This has motivated the application of this new technique to different areas, mainly to the task of speaker segmentation in two speaker conversations. Some approaches for two speaker segmentation that make use of JFA are presented in [3], [4], [5].

VIVOLAB-UZ is submitting two systems, both based on a combination of a JFA based speaker segmentation system and a BIC based AHC system. The only difference is that the first system obtains speaker labels directly from the BIC AHC step, while the second uses these labels to perform a final Viterbi resegmentation.

2. System Description

Our speaker diarization systems fuse a JFA based speaker segmentation system and a BIC based AHC system. Currently our speaker segmentation system works with a given number of speakers (it was designed for 2-speaker conversations), so firstly, after running a speech activity detector (SAD), we split every recording into 5 minute slices and every slice is processed with the speaker segmentation system separately. We force the speaker segmentation system to find 10 speakers in every slice. Once we have 10 clusters for every 5-minute slice, we perform a BIC AHC algorithm over the whole recording to merge those clusters belonging to the same speakers until a stopping criterion is met. This way we obtain th output for the first system. Our second system will use directly this output to perform a Viterbi resegmentation. The different steps of the VIVOLAB-UZ speaker diarization system are described in the following subsections.

2.1. Features

The features used for all steps in the diarization system are 18 MFCC including c0, computed every 10 ms over a 25 ms window. No normalization is used on the features.

2.2. Speech Activity Detector

The first step in most speaker diarization systems is to separate the speech segments from those segments that does not contain speech. In our case, non-speech segments may contain music, noise or silence. To obtain the speech segments we train 64 gaussian GMMs for two classes using the development data. One GMM for speech and the other one for non-speech. We run a Viterbi segmentation, modeling every class with 10 tied-states [6] that share the same GMM as the observation distribution.

This work has been partially funded by the national project TIN2008-06856-C05-04.

2.3. Segmentation System

In the proposed speaker segmentation system, described in [5], we use a factor analysis approach to model the desired sources of variability. As a starting point we try to capture the variability present among different speakers. For this purpose, we model every speaker by a Gaussian Mixture Model (GMM) adapted from a Universal Background Model (UBM) using an eigenvoice approach [7], according to:

$$M_s = M_{UBM} + Vy. \tag{1}$$

Where M_s is the speaker GMM supervector, obtained concatenating all Gaussian means, M_{UBM} is the UBM supervector, V is the low rank eigenvoice matrix, and y is the set of speaker factors, which follows a standard normal distribution $N(y|0, I)$ a priori. This way every speaker is represented by a GMM supervector in a high dimension space, and in such space we allow the speakers to lay in the low dimension subspace generated by the column vectors of V, which point to the directions of maximum variability among speakers. We refer to this variability as inter-speaker variability and to the low rank subspace as the speaker subspace.

In our approach we use a 256 Gaussian UBM. The dimension of the speaker subspace is 20, compared to the dimension of the supervector space that is $256 \times 18 = 4608$. This way every point estimate for a given speaker is defined by a set of 20 speaker factors.

Once we have the speaker factors we apply Within Class Covariance Normalization (WCCN) to compensate intra-session variability and to ensure that the variance of the speaker factors is close to I for every speaker as in [8]. WCCN is a normalization method that allows to obtain a linear transformation for a given set of features belonging to different classes so that the within class covariance matrix S_w defined in Eq. 2 is equal to the identity matrix I. This technique assumes that all classes have the same covariance matrix.

$$S_w = \frac{1}{S-1}\sum_{s=1}^{S}\frac{1}{N^s-1}\sum_{n=1}^{N^s}(y_n^s - \mu^s)(y_n^s - \mu^s)^T \tag{2}$$

$$\mu^s = \frac{1}{N^s}\sum_{s=1}^{N^s} y_n^s \tag{3}$$

To obtain the linear transformation we first obtain S_w as shown in Eq. 2 and then we apply Cholesky decomposition, so the transformed speaker factors y' will follow this expression:

$$y' = Ry \tag{4}$$

$$S_w^{-1} = R'R \tag{5}$$

where R is the upper triangular matrix obtained by Cholesky decomposition.

To perform speaker segmentation given a sequence of feature vectors, we estimate the speaker factors for every frame over a 100 frame window, with an overlap of 990 ms, we transform the speaker factors using WCCN, and we estimate a 10-Gaussian GMM to model the stream of speaker factors obtained, after removing non-speech frames according to the SAD. Each one of these Gaussians will be assigned to a single speaker. We perform this process over 5-minute slices, obtaining 10 new speakers for every slice.

2.3.1. *Initialization*

We have detected that a good initialization is quite important to ensure that every Gaussian in the GMM corresponds to a single speaker. In our approach, we use prior knowledge about speaker factors proposed in [2]: A priori, speaker factors are assumed to be distributed according to the standard normal distribution $N(y|0, I)$. Since we obtain speaker factors from a small data sample, using MAP estimation, we can expect the posterior distribution of speaker factors for a single speaker to keep some properties of the prior. In addition, since we perform WCCN on the speaker factors, we will be closer to fulfill this assumption. Assuming that the posterior variance is close to I, we can perform PCA to obtain those directions of maximum variability in the speaker factor space. Then we will use 9 ($N_{spks} - 1$) directions to obtain, using K-means, a first clustering.

This strategy gives ten clusters that can be seen as a first speaker segmentation, and then K-means clustering is performed over the 20 speaker factors to reassign frames to the ten clusters and a single Gaussian is trained on each of them.

2.3.2. *Core Segmentation*

The 10 Gaussians previously trained serve as initial GMM of the whole recording. Then a two stage iterative process is applied until convergence: first several Expectation-Maximization (EM) iterations are used and then, every Gaussian is assigned to a single speaker and a Viterbi segmentation is performed. According to this new frame assignment, 10 Gaussian models are trained and the iterative process restarts again. Convergence is reached when the segmentation of the current iteration is identical to that obtained in the previous one.

To avoid fast speaker changes, in the Viterbi segmentation, we modify the speaker turn duration distribution using a sequence of tied-states [6] for every speaker model. This way, we avoid the state duration to follow a geometric distribution that cannot accurately model real speaker turn durations. Each speaker model is composed of 10 states that share the same observation distribution, a single Gaussian in this case. Tied-states are not considered for the silence, but a single state without an observation distribution is used, since the algorithm is forced to go through the silence state according to the SAD labels. We have observed that this way of modeling speaker turn duration yields better results than modifying the transition probability.

2.4. Clustering

Once we have a set of ten clusters for every 5-minute-slice, an AHC step is performed to obtain the final clusters that correspond to the actual speakers. For this purpose, BIC is considered both as clustering metric and as stopping criterion. Every cluster is modeled using a single full covariance gaussian using MFCC, and two hypotheses are considered for every pair of clusters: The null hypothesis, that is, assuming that both clusters belong to the same speaker, and the 2 speaker hypothesis, that is, assuming that every cluster belong to a different speaker. BIC is computed for both hypotheses, and the ΔBIC is computed as $\Delta BIC = BIC_{2spks} - BIC_{null}$. The pair of clusters having lower ΔBIC is merged. Clusters are not longer merged when $\Delta BIC > 0$. To penalize the 2 speaker hypothesis the λ parameter for model complexity penalization is set to 10.0.

This step will give the output of the first VIVOLAB-UZ submission.

2.5. Resegmentation

Only the second system performs a final Viterbi Resegmentation. For this purpose, we model every speaker with a 32 component GMM using MFCC, according to the output of the BIC AHC step. As in the core segmentatino system, we use 10 tied-states for speaker models and a single state for all silence frames.

3. Development data

As development data we have considered those recordings provided for this purpose in the evaluation (16 sessions from the Catalan Broadcast News database) and the English and Mexican Broadcast News Speech from the Hub4 database.

For training the GMMs used in the SAD the 12 first sessions from the provided development data were used. The remaining 4 sessions were used to check the SAD and to adjust the AHC parameters (λ).

For training the UBM, the Eigenvoice matrix V and the WCCN transformation the Hub4 database was considered. Performance will be degraded for using different languages and dialects to train the models to obtain the speaker factors, but we could not find any other labeled datasets in Catalan.

4. Computational cost

The proposed system runs in matlab and it is not optimized. The following table resumes the computational cost for every step of the proposed diarization system:

Operation	Computational cost, real time (rt)
Obtain speaker factors	$0.20 \times rt$
SAD	$0.02 \times rt$
Segmentation	$1.10 \times rt$
BIC AHC	$< 0.01 \times rt$
Viterbi resegmentation	$\approx 1.20 \times rt$

Table 1: *Computational cost of the speaker diarization system, step by step.*

5. Conclusions and Future Work

This is the first time that VIVOLAB-UZ group participates in a Speaker Diarization Evaluation. We have built a system combining two state-of-the-art technologies: JFA and eigenvoice modeling for speaker segmentation and BIC based AHC for speaker clustering. We believe in the potential of the eigenvoice modeling for the task of speaker segmentation, but due to the lack of training data, we do not expect this submission to obtain state-of-the-art performance.

As future work we plan to study the potential of the eigenvoice modeling using matched data to train the V matrix, and we plan to improve the speaker clustering using GMM to model every cluster instead of single gaussians, and using a bigger GMM to model the null hypothesis, canceling the complexity penalization term in the BIC computation. This should give better agglomeration and a more robust stopping criterion.

Finally we plan to introduce Bayesian approaches to determine the number of speakers. We are working in the development of a speaker diarization system that combines eigenvoice modeling for speaker segmentation and Variational Bayes for determining the number of speakers in the recording. We believe this approach can obtain competitive performance.

6. Acknowledgements

We would like to thank the organizers of Albayzin 2010 Speaker Diarization Evaluation for their effort in preparing the evaluation data, and also to all the organization of FALA 2010 for supporting the Albayzin Speaker Diarization Evaluation.

7. References

[1] Wooters, C., Huijbregts, M., “The ICSI RT07s Speaker Diarization System”, in Proc Machine Learning for Multimodal Interaction, LCNS, Springer, Heidelberg, 2007.

[2] P. Kenny et al, “A Study of Inter-Speaker Variability in Speaker Verification”, IEEE Trans. Audio, Speech Proc., 2008

[3] Castaldo, F. et al, “Stream Based Speaker Segmentation Using Speaker Factors and Eigenvoices”, in Proc ICASSP, 4133-4136, Las Vegas, NV, 2008.

[4] Reynolds, D. et al “A Study of New Approaches to Speaker Diarization”, in Proc Interspeech, 1047–1050, Brighton, UK, 2009

[5] Vaquero, C. et al “Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification”, in Proc Interspeech, Makuhari, Japan, 2010.

[6] Levinson, S.E., “Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition”, Computer Speech and Language, I:29–45, 1986.

[7] R. Kuhn et al. “Rapid Speaker Adaptation in Eigenvoice Space”, IEEE Trans. Speech Audio Proc. Vol 8, no. 6, 695–707, 2000.

[8] Vaquero, C., Ortega, A., Lleida, E. “Intra-session Variability Compensation for Speaker Segmentation”, to appear in FALA, Vigo, Spain, 2010.