FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Intra-session Variability Compensation for Speaker Segmentation

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 addresses the problem of speaker segmentation in two speaker telephone conversations, proposing a segmentation approach based on factor analysis and a novel method for intra-session variability compensation to improve segmentation performance. The segmentation system is evaluated on the NIST Speaker Recognition Evaluation 2008 summed channel test condition, showing that intra-session variability compensation allows to obtain around a 20% relative improvement in terms of speaker segmentation error.

Index Terms: Speaker segmentation, speaker and session variability, intra-session variability

1. Introduction

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 [1]. Nowadays most state of the art speaker verification systems are based on this approach. Since then, researchers have explored its application to different areas, specially to study new speaker diarization methods. One of the most interesting of these methods is the one presented in [2], a novel approach for streaming speaker diarization, which shows several differences with traditional diarization systems. This method makes use of a simple Factor Analysis (FA) model composed of only eigenvoices [3] to obtain high accuracy in a two speaker segmentation task on telephone conversations. However, performance decreases significantly when the number of speakers is unknown.

Consequently, the speaker identification community has focused on improving the performance in the two speaker segmentation task on telephone conversations, a task quite related to speaker verification. In [4] several approaches using JFA and Variational Bayes are proposed, obtaining better performance than the traditional Bayesian Information Criterion (BIC) based Agglomerative Hierarchical Clustering (AHC) [5]. However all approaches presented in [4] only model inter-speaker variability to perform speaker segmentation. In [10] the same approaches are analyzed and inter-session variability compensation is added, showing that it decreases performance, since inter-session variability may contain useful information to separate different speakers in a single session, specially if they are talking over different channels.

In this work we address the problem of speaker segmentation in two speaker conversations. We propose two methods to compensate the variability presented by a single speaker during a session (intra-session variability) and an eigenvoice based approach for two speaker segmentation similar to the one presented in [2], obtaining competitive performance compared to state-of-the-art 2-speaker segmentation systems [4], and showing further improvement when the mentioned variability is compensated.

This paper is organized as follows: In Section 2 we present the proposed segmentation system, and we describe different types of variability that may affect a diarization system in Section 3. In Section 4, we introduce two approaches to compensate intra-session variability and in Section 5 we evaluate the speaker segmentation system and the proposed intra-session variability compensation approaches. Finally, in Section 6 we summarize the conclusions of this study.

2. Segmentation System

In the proposed speaker segmentation system, described in [6], 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 [3], according to:

$$M_s = M_{UBM} + Vy. \quad (1)$$

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, and as feature vectors we use 12 Mel Frequency Cepstral Coefficients (MFCC) including C0, computed every 10 ms over a 25 ms window. The dimension of the speaker subspace is 20, compared to the dimension of the supervector space that is $256 \times 12 = 3072$. This way every point estimate for a given speaker is defined by a set of 20 speaker factors.

To perform speaker segmentation given a sequence of feature vectors, as in [2], we estimate the speaker factors for every frame over a 100 frame window, with an overlap of 990 ms, an we estimate a 2-Gaussian GMM to model the stream of speaker factors obtained, after removing silence frames according to a Voice Activity Detector (VAD). Each one of these Gaussians

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

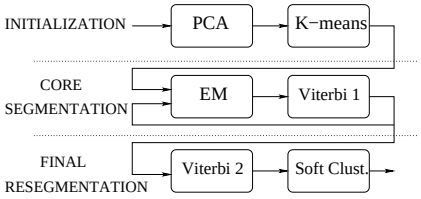

Figure 1: *Block diagram of the proposed segmentation system*

will be assigned to a single speaker. In contrast to [2], we estimate the GMM using all available data in the recording, rather than processing 1 minute slices and applying a clustering technique. The later allows stream processing with 1 minute latency but the former yields better results. A block diagram of the proposed segmentation system is shown in Fig. 1.

2.1. 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 [1]: 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 (100 frames, which is small compared to the number of frames that speaker recognition systems usually manage, around 10000), using MAP estimation, we can expect the posterior distribution of speaker factors for a single speaker to keep some properties of the prior. Assuming that the posterior variance is close to I, we can perform PCA to obtain the direction of maximum variability in the speaker factor space. Such direction should be the best one to separate speakers, since both are supposed to have a variance close to I and a different mean.

This strategy gives two clusters that can be seen as a first speaker segmentation, and then K-means clustering is performed to reassign frames to the two clusters and a single Gaussian is trained on each of them. Using this frame assignment directly as segmentation output gives reasonably good results, as we will see later, in Section 5.

2.2. Core Segmentation

The 2 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 (Viterbi 1 in Fig. 1). According to this new frame assignment, 2 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 [7] 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 VAD labels. We have observed that this way of modeling speaker turn duration yields better results than modifying the transition probability.

2.3. Viterbi Resegmentation and Soft Clustering

The output of the core segmentation system gives accurate speaker labels in most cases, but these labels can be refined by means of Viterbi resegmentations (Viterbi 2 in Fig. 1). In this case we model every speaker with a 32 component GMM according to the output of the core segmentation system using as features 12 MFCC including C0. Again we use 10 tied-states for speaker models and a single state for all silence frames.

After this resegmentation we retrain the GMM models and run a forward backward decoding to perform a soft reassignment of the frames to the two speakers. GMM models are retrained according to the soft reassignment and a final Viterbi resegmentation is performed. This approach was first presented in [4] as soft-clustering.

3. Speaker, Session and Intra-session Variability in Speaker Diarization

In the proposed approach for speaker segmentation we only take into account inter-speaker variability. However there are other sources of variability that may affect a segmentation or diarization system. In speaker recognition systems, one of the hardest problems is to deal with the variability present in a speaker recorded over different sessions. This is known as inter-session variability and includes variability due to the speaker, since his speech may vary along different recording sessions, as well as variability due to the recording environment. There are several techniques to model this variability. Some of the more recent and successful approaches have been Nuissance Attribute Projection (NAP) for SVM-GMM speaker recognition systems [8], Eigenchannel modeling, or JFA [9]. All this techniques assume that the speaker is modeled by a supervector (usually a GMM-sv) in a high dimension space and different sessions for a given space produce different estimations of the speaker supervector. The variability in these estimations or inter-session variability is assumed to lay in a low dimension sub-space, so all inter-session variability compensation techniques try to estimate the component of the speaker session in such space and remove it to obtain a session independent speaker supervector.

The question is if inter-session variability compensation is useful for speaker diarization. Speaker diarization systems aim at answering the question "Who spoke when?" in a unsupervised fashion. We can think that inter-session variability compensation do not help for speaker diarization, for two main reasons: First, diarization is performed over one session without prior knowledge of the speakers involved, so we will never get the same speaker over different sessions. Secondly, in many scenarios session variability models may enhance diarization performance since different speakers may use different communication channels. This is the case of telephone conversations or meetings in a room where the speakers remain static.

Finally, a single speaker can present variability during a single session when we process such session in small segments. We will refer to this variability as intra-session variability. Some examples of this variability includes emotions or excitement of the speaker as the conversation evolves, or the unbalanced phonetic load present in small segments as in the proposed system (1 second segments). Intra-session variability is not usually taken into account for speaker recognition, since state of the art systems usually integrate over all observations of a given speaker obtaining an average model, which may differ from session to session. In such case intra-session variability modeling and compensation will only be useful as far as it is re-

lated to inter-session variability. Actually, both intra and inter-session variability share many sources of variability, but some of them are more critical than others. For example, channel is a source of inter-session variability that in general will not introduce intra-session variability (but it could, e.g., if a speaker is recorded in a room with a far field microphone and he moves as he talks). On the other hand, unbalanced phonetic load will be more critical for intra-session variability modeling, specially as the segments to analyze in a given session become smaller.

However, intra-session variability is very important and should be taken into account in the task of speaker diarization, since in such task we analyze small and pure segments and try to agglomerate them to obtain pure clusters that should belong to a single speaker. In the following section we describe an approach for supervised intra-session variability compensation.

4. Intra-session variability compensation

Given a recording, the segmentation system proposed in section 2, produces a sequence of speaker factor vectors estimated every 10 ms over 1 sec. window. Assuming that a set of S recordings is available and each recording contains a single speaker, we can obtain a sequence $y^s = y^s_1, ..., y^s_{N^s}$ of N^s speaker factor vectors for every recording session s. The speaker factors obtained from a session belongs to the same class (same speaker), so we can study the inter-session and intra-session as between-class and within-class variability respectively. This approach is similar to the one presented in [11], but in that case it was used for speaker recognition and inter-session variability compensation.

4.1. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a well known technique for dimensionality reduction in pattern recognition that, given a set of features belonging to different classes and laying in the feature space, seeks the orthogonal basis for such space that enables better discrimination between different classes by maximizing between-class variance and minimizing within class variance. Linear discriminant analysis assumes that the observations belonging to each class are normally distributed and that within class covariance is kept across different classes. The speaker factor vectors satisfy the first assumption, while the second is expected to be satisfied since we do not expect the posterior covariance of y^s to be very different of the prior as we explained in section 2.

In our problem we estimate between-class covariance (S_b) and within class (S_w) covariance as:

$$S_b = \frac{1}{S-1}\sum_{s=1}^{S}(\mu^s - \mu)(\mu^s - \mu)^T \quad (2)$$

$$S_w = \frac{1}{S-1}\sum_{s=1}^{S}\frac{1}{N^s-1}\sum_{n=1}^{N^s}(y^s_n - \mu^s)(y^s_n - \mu^s)^T \quad (3)$$

$$\mu^s = \frac{1}{N^s}\sum_{s=1}^{N^s} y^s_n \quad (4)$$

$$\mu = \frac{1}{S}\sum_{n=1}^{N^s}\mu^s \quad (5)$$

The problem reduces to find the matrix v of eigenvectors that satisfies:

$$S_b v = \lambda S_w v, \quad (6)$$

and project the speaker factors onto v or onto a low rank matrix A obtained selecting those eigenvectors having higher eigenvalues, for dimensionality reduction.

4.2. Within Class Covariance Normalization

Within class covariance normalization (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. 3 is equal to the identity matrix I. Again this technique assumes that all classes have the same covariance matrix.

To obtain the linear transformation we first obtain S_w as shown in Eq. 3 and then we apply Cholesky decomposition, so the transformed speaker factors y' will follow this expression:

$$y' = Ry \quad (7)$$

$$S_w^{-1} = R'R \quad (8)$$

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

5. Performance Analysis

5.1. Experimental Setup

We study the performance of the proposed segmentation system and intra-speaker variability compensation in terms of segmentation error rate. As development data to train the UBM, V matrix, LDA and WCCN we use all telephone data available from 1conv and 8conv conditions from the NIST SRE evaluations 2004, 2005 and 2006. As evaluation data we use the summed channel test condition from the NIST SRE 2008. This condition comprises 2213 2-speaker telephone conversations of around five minutes length each. As ground truth for segmentation error rate computation we extract the segmentation labels from the ASR NIST transcriptions obtained separately on each telephone of the conversation.

5.2. Segmentation Performance: Baseline

As we explained in Section 2, the proposed segmentation system comprises several steps, including PCA initialization, K-means clustering, iterative EM and Viterbi segmentation in the speaker factor space, a Viterbi resegmentation using MFCC features and a last soft-clustering resegmentation. Table 1 shows the results obtained by the segmentation system after every step:

Segmentation system	Seg error (%)	σ(%)
PCA	20.2	14.3
+K-means	4.9	8.8
Core segmentation system	3.1	6.6
+Viterbi resegmentation	2.3	6.2
+Soft-clustering	2.2	6.1

Table 1: *Performance of the segmentation system and standard deviation step by step.*

Given these results we can extract several conclusions. First, speaker factors enable easy separability between speakers. Just with PCA and K-means clustering we get 4.9% segmentation error. Note that at that point, frames are assigned to one speaker or the other assuming statistical independence, no context or temporal information is used. Completing the core system gives great improvement and results are comparable to those obtained with the best systems presented in [4]. Moreover, after resegmentations results improve further.

5.3. Intra-speaker Variability Compensation

To study the performance of intra-speaker variability compensation we compare the segmentation error obtained before the resegmentation stages (after the core segmentation in Fig. 1) with and without using the intra-speaker variability compensation methods described in Section 4. For comparison when using LDA for dimensionality reduction we show results using 20 speaker factors (baseline system) and 50 speaker factors.

Segmentation system	Seg error (%)	$\sigma(\%)$
Baseline (20 spk factors)	3.1	6.6
WCCN (20 spk factors)	**2.5**	**5.5**
50 spk factors	2.9	6.9
LDA 50 to 20	2.7	5.7
LDA 50 to 20 + WCCN	2.5	5.7
50 spk factors + WCCN	**2.1**	**5.6**

Table 2: *Performance of the core segmentation system with and without intra-speaker variability compensation.*

As we can see in Table 2, both LDA and WCCN approaches for intra-speaker variability compensation outperforms our baseline. Using WCCN directly on 20 speaker factors reduces the segmentation error from 3.1% to 2.5%, obtaining a 20% of relative improvement. Using LDA to obtain 20 dimension vectors from 50 speaker factors improves also the performance of the system compared to the baseline using both 20 and 50 speaker factors. In addition, the performance can be further improved applying WCCN after LDA.

However using LDA+WCCN on 50 speaker factors is not significantly better than using WCCN directly on 20 speaker factors. Moreover, the most critical step in the proposed system regarding computational cost is the speaker factor computation ($O(d^2)$, with d the dimension of the speaker factors), and once speaker factors are computed, the classification algorithm is fast compared to speaker factor computation ($O(d)$). Therefore, the computational cost of the system using LDA for dimensionality reduction is comparable to the cost of the system using 50 speaker factors and is much higher than the cost of the system using 20 speaker factors. For this reason, we show the results obtained with 50 speaker factors and WCCN for intra-speaker compensation. We obtain a 28% relative improvement when using WCCN on 50 speaker factors. It seems that a higher dimensionality enables WCCN to improve further.

Taking into account the computational cost, we can affirm that, even though LDA based intra-session variability compensation shows improvements, it is not useful for our system since using WCCN on low dimension speaker factor space performs as good as using LDA+WCCN on a higher dimension speaker factor, but this second approach is much more costly, and if we use WCCN directly on the higher dimension speaker factor space we obtain further improvement keeping the computational cost comparable to LDA+WCCN.

5.4. Results with the Full Segmentation System

In the previous subsections we have shown results for the core segmentation system, but the proposed segmentation system can increase its performance using Viterbi resegmentation after obtaining the core segmentation output.

Results in Table 3 show that while increasing the number of speaker factors is not effective after Viterbi and soft-clustering resegmentations, intra-session variability compensation using WCCN is still effective, obtaining a relative performance improvement of 18% for 20 speaker factors and 23% for 50 speaker factors. In addition it is shown that increasing the number of speaker factors may not be helpful if intra-session variability is not compensated, probably because some directions of the speaker space are related to intra-session variability.

Segmentation system	Seg error (%)	$\sigma(\%)$
Baseline + reseg	2.2	6.1
WCCN + reseg	**1.8**	**5.0**
50 spk factors + reseg	2.2	6.2
50 spk factors + WCCN + reseg	**1.7**	**5.2**

Table 3: *Performance of the segmentation system with WCCN for intra-speaker variability compensation.*

6. Conclusions

In this study, we have introduced two methods for intra-session variability compensation in the task of speaker segmentation and diarization, based on LDA and WCCN. In addition, we have proposed a two speaker segmentation system based on the one presented in [2], introducing a set of improvements, including a novel PCA initialization and a modification of the speaker turn duration distribution, that enables us to obtain a 2.2% segmentation error on the summed dataset from the NIST SRE 2008. We have shown that intra-session variability compensation can improve performance of a segmentation system, reducing the segmentation error rate to 1.8%.

7. References

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

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

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

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

[5] Reynolds, D. A. and Torres-Carrasquillo, P., "Approaches and applications of audio diarization", In Proc ICASSP, V:953–956, Philadelphia, PA, 2005.

[6] Vaquero, C. et al "Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification", to appear in Interspeech, Makuhari, Japan, 2010.

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

[8] Campbell, W. et al "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation", In Proc ICASSP, Toulouse, France, 2006.

[9] Kenny, P. et al, "Joint factor analysis versus eigenchannels in speaker recognition" IEEE Trans. Audio, Speech Proc. 15 (4), pp. 1435-1447, 2007

[10] Kenny, P. et al, "Diarization of Telephone Conversations using Factor Analysis" IEEE Journal of Selected Topics in Signal Processing, 2010.

[11] Dehak, N. et al, "Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysi", in Proc ICASSP, Taipei, Taiwan, 2009.