Confidence Measures and Hypothesis Selection Strategies for Speaker Segmentation

C. Vaquero*, A. Ortega, J. Villalba, E. Lleida

University of Zaragoza
Maria de Luna 1, 50018 Zaragoza, Spain
{cvaquero,ortega,villalba,lleida}@unizar.es

Abstract. This paper addresses the problem of speaker segmentation in two speaker conversations, proposing a set of confidence measures to assess the quality of a given speaker segmentation. In addition we study how these measures can be used to apply hypothesis selection strategies in order to improve the performance of a 2-speaker segmentation system and how they are related to speaker verification performance. Our approach for speaker segmentation is based on the eigenvoice paradigm. We present a novel PCA based initialization in the speaker factor space along with a modification of the speaker turn duration distribution. Three confidence measures are analyzed on the output of the proposed segmentation system for the NIST Speaker Recognition Evaluation 2008 summed condition, showing that they constitute a good measure to estimate the segmentation accuracy and can be used for applying back-off strategies.

Keywords: Speaker segmentation, confidence measures, hypothesis selection strategies

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 has explore 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 to obtain high accuracy in a two speaker segmentation task on telephone conversations.

Consequently, the speaker identification community has focused on improving the performance in the two speaker segmentation task on telephone conversations, a task related to speaker verification. In [3] several approaches using JFA

* This work has been supported in part by the program FPU from MEC of the Spanish Government

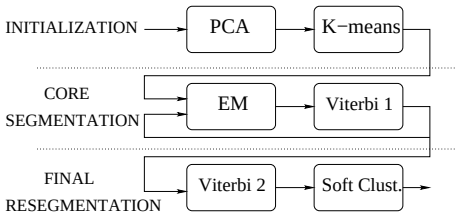

Fig. 1. *Block diagram of the proposed segmentation system*

and Variational Bayes are compared to a traditional Bayesian Information Criterion (BIC) based Agglomerative Hierchical Clustering (AHC) system [4]. In that study, most approaches show higher accuracy than the AHC system.

In this work we address the problem of speaker segmentation in two speaker conversations and how a set of confidence measures that assess the quality of a given speaker segmentation can be used to select the best segmentation among various hypotheses improving speaker segmentation performance. We use an eigenvoice approach for two speaker segmentation similar to the one presented in [2], introducing some modifications to achieve improved performance.

In Section 2 we introduce the segmentation system, and we describe three confidence measures to estimate the segmentation performance in Section 3. In Section 4 we explain a hypothesis selection strategy that uses the confidence measures to improve segmentation performance, while in Section 5, we evaluate the segmentation system, the confidence measures and the hypothesis selection strategy. Finally, in Section 6 we summarize the conclusions of this study.

2 Segmentation System

The proposed approach for speaker segmentation is detailed in [5]. We model every speaker by a Gaussian Mixture Model (GMM) adapted from an Universal Background Model (UBM) using an eigenvoice approach, so every speaker can be represented by a small set of speaker factors. We compute a set of 20 speaker factors for every frame over a 100 frame window, and we estimate a 2-Gaussian GMM on the speaker factors. Each one of these Gaussians will be assigned to a single speaker. A block diagram of the segmentation system is shown in Fig. 1.

2.1 Initialization

A good initialization is important to ensure that every Gaussian in the GMM corresponds to a single speaker. In [5], a new initialization is proposed: since speaker factors are priorly distributed following normal standard distribution 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. 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.

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: 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 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 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 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 [3] as soft-clustering.

3 Confidence Measures

In the following section we present a set of confidence measures that aims at determining the performance of the segmentation system explained in the previous section for a given audio recording. This set is of measures is analyzed in [5]

3.1 Bayesian Information Criterion

In order to use BIC as a confidence measure, given two sequences of acoustic feature vectors obtained by the segmentation system, we compute the BIC for two hypothesis: Each sequence belongs to a different speaker or both sequences belong to the same speaker. The confidence measure is the difference between BIC values. To avoid adjusting BIC penalty parameters, we force the models for both hypothesis to have the same complexity.

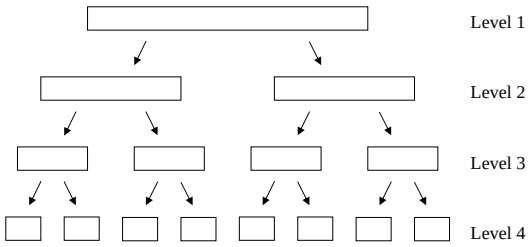

Fig. 2. *Slice partition diagram*

3.2 Kullback-Leibler Divergence in the Speaker Factor Space

Another way to measure the accuracy of a given segmentation is to compute the symmetric Kullback-Leibler (KL) divergence between the Gaussian speaker models obtained in the speaker factor space. In this approach we use the hypothetic segmentation labels to obtain two sequences of speaker factors, and Gaussian models are trained for each sequence. We can expect higher KL divergences between both Gaussian models when the segmentation is correct.

3.3 Core Segmentation System Convergence

In Section 2 we saw that the core segmentation runs until convergence. A way to estimate the quality of the output of the core segmentation system is to study how long did it take to converge. We can expect the system to converge fast when it can easily find the correct segmentation and to converge slow otherwise. This measure is probably less correlated with the previous measures described.

4 Hypothesis Generation and Selection

Assuming that the proposed set confidence measures enables us to obtain an idea of how good is a segmentation hypothesis, we can generate several segmentation hypotheses and use the confidence measures to select the best among them.

4.1 Hypotheses Generation

In order to generate different segmentation hypotheses for a given conversation we split the conversation into 2 slices and then every slice into two new slices iteratively until we obtain eight non-overlapping slice, as shown in Fig. 2.

Every time we split a given slice, we look for the silence interval that is closest to the middle of the slice and we split the slice at a point as close to the middle as possible. This way, we obtain a tree composed by 4 levels, each level having 2^{l-1} where l is the level number as pointed in Fig. 2.

Then we perform the segmentation algorithm on every slice independently, obtaining a two speaker segmentation hypothesis for every slice. After this, we

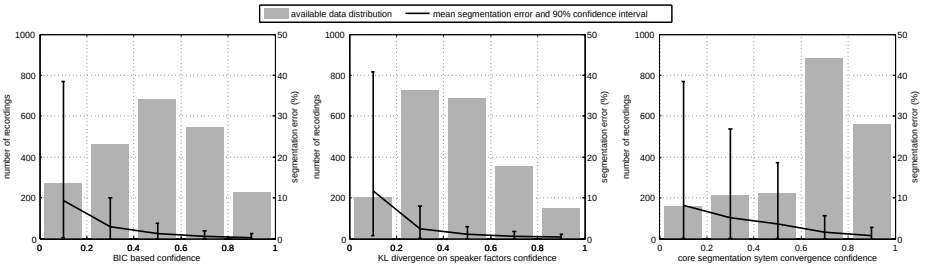

Fig. 3. *Distribution of the available data, mean segmentation error and 90% confidence interval for different confidence measures.*

compute a confidence measure for every slice, and for every level we keep those slices having higher confidence measure, discarding the remaining slices. This way we obtain a set composed by a variable number of slices for every level. Every set is supposed to contain the best segmented slices for the given level. Then we perform AHC using BIC and full covariance single gaussian models in the MFCC space to agglomerate the 2 speakers obtained in one slice with those obtained in the other slices belonging to the same level, until we have two speakers. In this step we force the agglomerative procedure to always merge speakers belonging to different slices.

Once we have the segments from the best slices of every level agglomerated in two speakers or clusters, we train a 32 gaussian GMM for every speaker and perform the resegmentation steps described in Section 2, obtaining one segmentation hypothesis for every level, a total of 4 segmentation hypotheses. Finally we compute a confidence measure for every segmentation hypothesis and select that having the highest confidence measure.

5 Performance Evaluation

5.1 Experimental Setup

To evaluate the proposed segmentation system and the confidence measures, we use the 2213 five minute telephone conversations from the NIST SRE 2008 summed channel condition. Performance is measured in terms of segmentation error rate. We will use as baseline the segmentation system described in Section 2, which obtains a 2.2% segmentation error rate in the evaluation dataset.

5.2 Confidence Measures

To analyze the proposed confidence measures, first we normalize them to be in the range $[0, 1]$ and then we divide the dataset into 5 subsets according to a uniform division of the confidence measure range. Fig. 3 represents the distribution of the recordings and the mean segmentation error with the 90% confidence

interval (CI) over the previously proposed confidence measure ranges, for each confidence measure. We can observe that all confidence measures follow the expected behavior: as they increase, the mean segmentation error decreases and so does the 90% CI. This way, we can assure that given a segmentation output with a high value in its confidence measure there is a high probability of having a good segmentation. However, we can not assure that given a low confidence measure the segmentation is wrong, since the CI is large is that case, simply we can not be sure if it is right or not This behavior does not allow to predict the segmentation error given the confidence measures in all cases, but it is enough to consider them as confidence measures of segmentation quality. Comparing them, we can see that BIC and KL divergence behave in a similar way and they both could be used to detect very well segmented recordings, while the CI of the convergence measure does not decrease as fast when the confidence measure increases, and most recordings concentrate on higher values of this measure.

To show the capability of the proposed confidence measures, we use a single confidence measure, obtained as a linear combination of the three proposed measures. For this analysis we use the same weight for all measures (1/3), but in general the weights can be estimated to optimize a cost function or simply to empathize a confidence measure more reliable than other. Again, we divide the dataset into 5 subsets according to a uniform division of the confidence measure range. We analyze the mean speaker segmentation error rates and their typical deviations for the single confidence measure obtained. In addition, to study the relation of these confidence measures to speaker verification performance, we analyze such performance in terms of Equal Error Rate (EER) and Minimum Detection Cost Function (minDCF) for the summed channel condition of the NIST SRE 2008. Results are displayed in Table 1.

Confidence	0.0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
Recordings	133	413	743	699	225
μ Seg Err (%)	14.1	4.0	1.6	0.6	0.2
σ Seg Err (%)	14.0	6.6	3.9	2.4	0.4
Trials	2198	6811	12201	11089	3430
EER (%)	12.65	10.53	9.37	7.79	6.07
minDCFx10	0.549	0.563	0.477	0.431	0.307

Table 1. *Performance of the speaker segmentation for 5 uniform confidence measure ranges.*

We can see that as the confidence measure increases, the segmentation error decreases, and so does the typical deviation of the segmentation error. Moreover, we can see that the proposed confidence measures can be used to segregate those test segments that will give better performance in a speaker verification system.

To analyze this effect, we split our dataset into two subsets. Splitting is done setting a threshold in the confidence measures. We found this threshold to be 0.55 in order to keep same number of recordings in both subsets. Fig. 4 shows

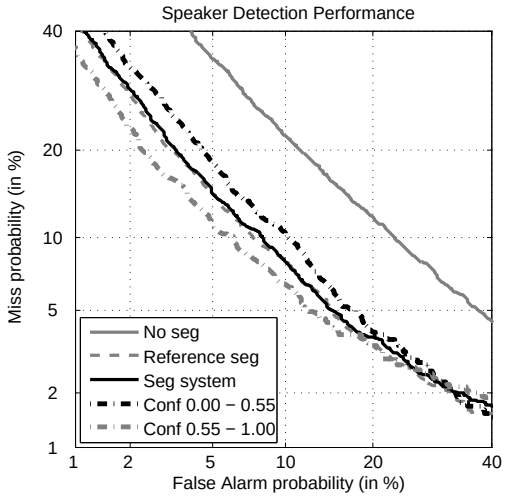

Fig. 4. *DET curves for non-segmentation, reference segmentation and the proposed segmentation, for all data and 2 confidence measure ranges*

Detection Trade Off (DET) curves obtained for the NIST SRE 2008 short2-summed condition, when not using any segmentation, using the segmentation system without hypothesis selection, and considering the ground truth segmentation. For comparison, the DET curves obtained using the segmentation system for the mentioned subsets are represented. We can see that in terms of speaker verification, the segmentation system enables the verification system to perform as well as if ground truth segmentation was provided. However, the most interesting thing is that using the confidence measures, we can segregate half of the dataset that enables the verification system to obtain much better performance that the remaining half of the dataset.

5.3 Hypothesis Selection Evaluation

In order to evaluate the proposed method for hypothesis generation and selection, we will obtain a new confidence measure. Since the number of iterations that takes core segmentation system to converge cannot be used to obtain a confidence measure for the whole conversation in those levels where more than one slice is available (we perform the core segmentation algorithm on each slice, not on the whole conversation), we will use a linear combination of BIC and KL divergence. This time we train the weights for every confidence measure using logistic linear regression, trying to be able to segregate correctly those slices having less than 5% segmentation error. For this purpose we use the FoCal toolkit and the data of the NIST SRE 2000 Switchboard segmentation task as development data.

We perform the method described in Section 4 to obtain several segmentation hypothesis and we select the hypothesis having higher confidence measure. This way we obtain the results presented in Table 2.

Segmentation system	Seg error (%)	σ(%)
Level 1	2.2	6.1
Level 2	2.2	5.9
Level 3	2.4	6.0
Level 4	2.6	6.4
Hypothesis selection	**1.9**	**5.3**
Best possible selection	1.3	3.5

Table 2. *Performance of the segmentation system and standard deviation step by step.*

As we can see, the confidence measures can be used for select among a set of different segmentation hypothesis, improving the global performance of the system. The segmentation error using hypothesis selection based in the confidence measures is lower than the segmentation error at every level. However, the hypothesis selection could be better. If we could obtain a confidence measure which had a monotonic behavior with respect the segmentation error rate, we could always select the best segmentation, obtaining a 1.3% segmentation error.

6 Conclusions

In this study, we have introduced three confidence measures in order to determine the accuracy of the segmentation system and a hypothesis generation and selection strategy to take advantage of the confidence measures. Given a dataset, the proposed confidence measures make possible the segregation of those recordings that will give good segmentation from those that will give less performance in average. This way, we have increased the performance of the proposed segmentation syste, reducing the segmentation error rate from 2.2% to 1.9% on the summed dataset from the NIST SRE 2008. This system is competitive compared to other segmentation systems recently proposed in [3].

References

[1] P. Kenny et al, "A Study of Inter-Speaker Variability in Speaker Verification", IEEE Transactions on Audio, Speech and Language Processing, 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] Reynolds, D. et al "A Study of New Approaches to Speaker Diarization", in Proc Interspeech, 1047–1050, Brighton, UK, 2009

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

[5] Vaquero, C. et al "Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification", to appear in 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.