Joint Speaker Segmentation, Localization and Identification for Streaming Audio

Joerg Schmalenstroeer, Reinhold Haeb-Umbach

Department of Communications Engineering,
University of Paderborn, Germany
schmalen@nt.uni-paderborn.de, haeb@nt.uni-paderborn.de

Abstract

In this paper we investigate the problem of identifying and localizing speakers with distant microphone arrays, thus extending the classical speaker diarization task to answer the question "who spoke when *and where*". We consider a streaming audio scenario, where the diarization output is to be generated in realtime with as low latency as possible. Rather than carrying out the individual segmentation and classification tasks (speech detection, change detection, gender/speaker classification) sequentially, we propose a simultaneous segmentation and classification by applying a Viterbi decoder. It uses a transition matrix estimated online from position information and speaker change hypotheses, instead of fixed transition probabilites. This avoids early hard decisions and is shown to outperform the sequential approach.

Index Terms: speaker diarization, acoustic scene analysis, Viterbi decoder

1. Introduction

Speaker diarization is the task of annotating an input audio channel with information that attributes temporal regions to specific speakers [1]. There are three primary domains which have been used for speaker diarization research: broadcast news audio, recorded meetings, and telephone conversations. Usually, a batch processing scenario is considered, i.e. the complete recording of the audio data is available at the beginning of the diarization.

In this paper we are also concerned with speaker diarization; however under three distinct differences. First, we consider an online scenario, where an audio stream has to be segmented and classified on the fly with as little latency as possible. Only a maximum latency of a few seconds is tolerable. For this reason, neither iterative nor multi-stage batch procedures can be used in order to avoid long system delays. The second major difference is that, besides the classical speaker diarization information, answering "who spoke when", we are also interested in the position of the speaker in the room. To this end we assume that microphone array data are available, from which speaker position information will be derived. While on the one hand the use of wall-mounted distant microphones leads to poor signal quality (low signal-to-noise ratio, reverberation), the use of position information can greatly improve the speaker change detection accuracy, if different speakers are assumed to be in different spatial locations [2]. The third difference is that, unlike in a traditional speaker diarization setup, the number of speakers is known in advance and models to carry out speaker identification are available. This assumption is reasonable for application in an intelligent home environment, which is considered here, since the system is meant to identify family members.

Localizing and identifying speakers on the fly can be used in intelligent environments to automatically select the most appropriate input/output device based on position information, and to adapt interfaces to user profiles and preferences, once the user/speaker is identified, as is investigated within the Amigo project (Ambient intelligence for the networked home environment, IST 004182, [3]). Another application is advanced video conferencing with automatic camera steering, microphone selection and metadata generation.

In our approach for online speaker diarization we carry out segmentation and classification in parallel rather than sequentially in order to avoid early unfavorable hard decisions, which cannot be corrected lateron due to the on the fly processing considered here. We employ a Hidden Markov Model (HMM), where the states represent speakers. Thus speaker information from the speaker identification module is taken as state observation probabilities, while speaker change information, obtained from BIC and the microphone array is used to obtain state transition probabilities. Diarization information is output at a fixed maximum latency using partial traceback at the Viterbi decoder. While the integrated approach in [4] uses fixed state transition probabilities and focuses on iteratively growing the HMM in the presence of an initially unknown number of speakers, we employ time-variant transition probabilites, which are obtained from speaker change information, and we assume that the number of speakers is known.

In the next section we outline the individual knowledge sources used in the diarization process. Section 3 describes the Hidden Markov Model on which the Viterbi algorithm, given in section 4, operates. Experimental results are presented in section 5, and we finish with some conclusions drawn in section 6.

2. Knowledge sources

Figure 1 gives an overview of the overall system architecture and the knowledge sources used, which are speaker change information from the Bayesian Information Criterion (BIC), speaker position derived from the microphone array, voice activity information (VAD), and the Gaussian mixture model (GMM) for speaker identification. Each knowledge source is modeled probabilistically, as described in the following.

Speaker change information from location information. A Filter-and-Sum Beamformer (FSB) using a blind adaptation method [5] delivers an estimate of the Direction-of-Arrival (DoA) as a byproduct of speech enhancement. In the case of multiple, distributed arrays, even the speaker position can be estimated in Cartesian coordinates. Every 10ms a position (or

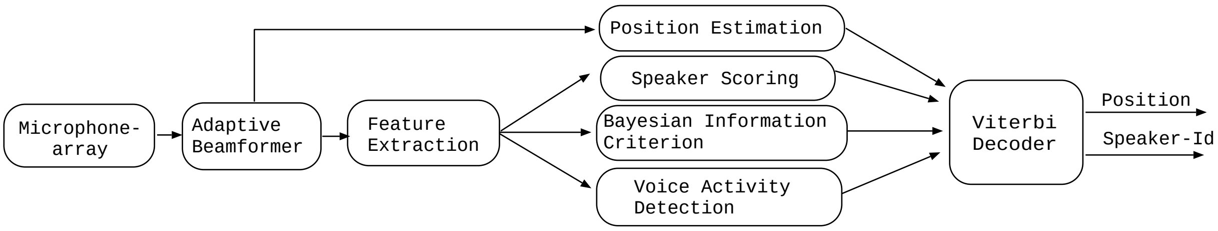

Figure 1: *Overall system architecture and knowledge sources.*

direction) estimate is computed. The feature x^{pos} to be used in the diarization task is the variance of the position estimates within a time window of width 0.5s. From training data we estimate the parameters of the Gaussian $p(x^{pos}|c)$, where the binary variable c indicates the presence or absence of a speaker change in that interval.

Speaker change information from BIC. We compute feature vectors from the audio signal at the output of the FSB using the ETSI AFE advanced feature extraction front end [6]. Next, BIC values are obtained from the feature vectors within a sliding window of 0.6s width [7, 8]. The feature x^{bic} to be subsequently used is the variance of the BIC values within a window of 0.5s. $p(x^{bic}|c)$ is again modeled as a Gaussian with parameters estimated on the training data.

Speaker id information from GMM. Speaker identification is based on a Gaussian Mixture Model. The GMM $p(x^{sid}|s)$ for each speaker s is obtained from a universal background model (64 Gaussians, diagonal covariance matrix) by using Bayesian adaptation on speaker-specific training data (approx. 3min / speaker). The feature vector x^{sid} consists of the 12 MFCCs $c_1, \ldots, c_{12}$ from the ETSI AFE, their first and second order derivatives, and a voicedness feature [9] and its first and second order derivatives.

Voice activity information from VAD. We employ an energy-based voice activity detector or alternatively the ETSI XAFE voice activity detector. It delivers a soft output speech indicator $P(\text{speech}|x^{vad})$, which can assume any value between 0 (absence of speech with high confidence) and 1 (presence of speech with high confidence).

We also experimented with a gender classifier to further enhance the speaker identification. However, the performance gain was marginal and therefore we decided to leave it out in the following.

3. Hidden Markov Model

Figure 2 displays the ergodic HMM used for speaker diarization. It shows a scenario with three speakers. The number of hidden states is equal to the number of speakers (which is assumed to be known a priori), plus one state for "silence". Note that we do not carry out a speech activity detection and frame dropping upfront, but propagate the soft VAD output to the Viterbi decoder, again to avoid early hard decisions.

The observation probability of each state is given by the GMM. While the GMM was trained on speech frames only, the input x^{sid} to be classified may also contain non-speech frames, as frames classified as non-speech are not eliminated prior to the HMM decoder. As a consequence the GMM likelihood must be multiplied by the probability of the frame being a speech frame, resulting in the following observation probability for a state $s = j$ representing a speaker:

$$b_j(x(k)) = p(x^{sid}(k)|s = j)P(\text{speech}|x^{vad});\ j : \text{spk} \quad (1)$$

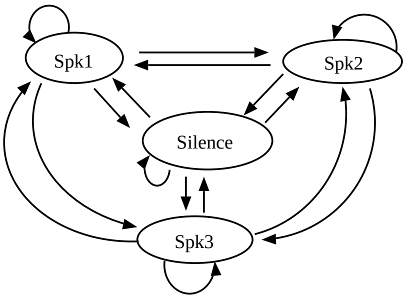

Figure 2: *Hidden Markov Model for speaker diarization*

If s denotes the "silence" state, the observation probability is modified to

$$b_j(x(k)) = p(x^{sid}(k)|s = j)(1 - P(\text{speech}|x^{vad}));\ j : \text{sil} \quad (2)$$

where an average GMM-score is taken as $p(x^{sid}|s = \text{sil})$.

The transition probabilities are formed by the speaker change information. Let $s(k-1) = j$ and $s(k) = i$. We define the binary random variable $c(k)$ to be one if there is a speaker change from time $k-1$ to k (i.e. $i \neq j$) and zero else. The transition probability between "speaker" states (i.e. excluding the "silence" state) is then given by $p(c(k)|x^{pos}(k), x^{bic}(k))$. Assuming that x^{pos} and x^{bic} are statistically independent, we find

$$p(c(k)|x^{pos}(k), x^{bic}(k)) = \frac{p(x^{pos}(k), x^{bic}(k)|c(k))P(c(k))}{p(x^{pos}(k), x^{bic}(k))}$$
$$= \frac{p(x^{pos}(k)|c(k))P(c(k))}{p(x^{pos}(k))} \frac{p(x^{bic}(k)|c(k))P(c(k))}{p(x^{bic}(k))} \frac{1}{P(c(k))}. \quad (3)$$

If we assume that $P(c(k))$ is constant for all $c(k)$, the term can be further simplified and we obtain the transition score

$$a_{ij}(k) = \frac{p(x^{pos}(k)|c(k))}{\sum_{c'} p(x^{pos}(k)|c')} \cdot \frac{p(x^{bic}(k)|c(k))}{\sum_{c'} p(x^{bic}(k)|c')} \quad (4)$$

The "silence" state needs again special treatment. A transition to the silence state is assigned the probability $p(c|x^{bic})$, and a loop in the "silence state" is assigned $(1 - p(c|x^{bic}))$. The reason for the absence of the position dependent feature is that a transition to silence obviously does not correspond to a Direction-of-Arrival or position change of the input signal. Note that the transition score need not be a probability. Further note that the transition scores are time variant.

4. Viterbi decoder

By unfolding the state transition diagram of Figure 2 over time a trellis diagram is obtained, on which the Viterbi algorithm operates to deliver the single best state sequence through that trellis, given the (joint) observations $x(1), \ldots, x(K)$:

$$\hat{s}_1^K = \underset{s_1^K}{\operatorname{argmax}} \sum_{k=1}^{K} \left(\log b_j(x(k)) + \kappa \log a_{ij}(k)\right) . \tag{5}$$

It is known from the literature that minimum length constraints and heuristic smoothing rules may be applied to suppress excessive state switching [1]. They are based on the assumption, that every speaker speaks at least for a small time period, producing more than just a few observation vectors. Here we avoid unwanted state oscillations by using a heuristic weighting factor κ, which controls the importance of the transition score relative to the observation score.

To guarantee a latency smaller than a given maximum value, partial traceback is implemented and initiated every second from the state with the currently best score. The traceback outputs the unique part of the state history and forces an output at least after a maximum time delay of $\tau_{\max}$. However, an unanimous state sequence is typically found already after a shorter delay.

Sofar the speaker position information information has only been used to find speaker changes. However, the output $\hat{s}_1^K$ of the Viterbi-based diarization can now be combined with the speaker position information, to obtain the augmented diarization information "who spoke when and where". Once speakers have been assigned to positions, the position of individual speakers can be tracked over time, e.g. by filtering the position estimates of each speaker using a state space model of the movements.

5. Experimental results

The problem under consideration in this paper is different from the classical speaker diarization task, as was desribed earlier. The databases used in the DARPA EARS Rich Transcription effort cannot be used since they do not contain microphone array data, from which position information could be extracted. The databases compiled within the CHIL project include microphone array data. However, the multi-channel recordings mainly consist of lecture recordings, where the same speaker is active for more than 90% of the time, which is inappropriate for meaningful speaker change detection experiments [10].

We therefore decided to use a self compiled database for our experiments. The database contains about 1.5 hours of spoken texts from a total of 5 male and 5 female speakers, recorded by a linearly arranged microphone array (6 microphones; intermicrophone distance of 5cm) at 2.8m distance from the speakers. More details on the database can be found in [2]. For the experiments reported below the database was divided in three subsets according to the average length of contiguous speaker segments: fast (below 2s), medium ($3-4$s) and slow (more than 4s). The performance is measured by the speaker diarization error rate (DER), which is the ratio of frames the detected state label (including silence) does not coincide with the ground-truth state label versus the total number of frames (no tolerance) [1].

To obtain an impression whether the individual components of our system are state-of-the-art, we tested our speaker identification software on recordings of the CHIL database, for which classification results of other laboratories were available and obtained results comparable to those. The sliding-window BIC algorithm used here is similar to [8], which was there shown to deliver good results. Finally, the adaptive Filter-and-Sum Beamformer, from which the position information is obtained, has been shown to deliver good performance in [11].

5.1. Ground-truth speaker segmentation

When the ground-truth speaker change points are known, the diarization performance is limited by the performance of the speaker identification software. This performance, denoted as "ground-truth change points" in Table 1, serves as a benchmark for our system.

5.2. Sequential diarization

A first and very simplistic approach to diarization is to carry out speaker recognition on a sliding window of fixed width and assign the detected speaker label to the center frame of the window. Figure 3 displays the DER as a function of the window width. It comes to no surprise that the DER increases on the databases with the shorter speaker segments.

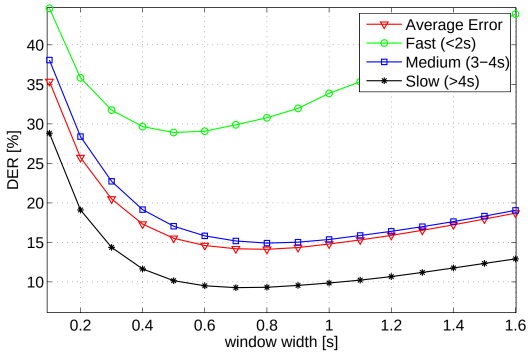

Figure 3: *DER performance using fixed window size*

A common approach to diarization is to first detect speaker changes with the Bayesian Information Criterion (BIC), and then apply speaker recognition on the found segments. Several variants of the BIC rule and related techniques have been proposed in the literature. We have chosen the metric decision rule, which is to our experience quite robust to varying acoustic environments and thus needs less calibration [2]. In Figure 4 DER is shown against the BIC threshold δ, which effects the average length of the detected segments. A high threshold favors longer segments. For the optimal value of δ, this approach gives slightly better results than the fixed window approach. However, good performance is obtained for a wider range of values for δ, while in the fixed window approach the DER depends strongly on the window width.

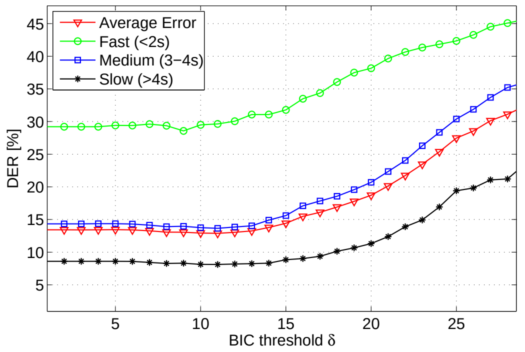

Figure 4: *System performance using BIC*

The first two results lines of Table 1 summarize these results: it contains the minimum DER of each curve of Figures 3 and 4. This DER performance is achieved if the optimal value of the parameters window width and the BIC threshold are used on each database.

Speech duration	$< 2s$	$3 - 4s$	$> 4s$	Avg.
Fixed window	29.00	15.14	9.10	14.21
BIC	28.76	13.91	7.94	12.98
Viterbi (Pos,BIC,$\kappa = 1$)	22.62	11.52	6.83	10.69
Viterbi (Fixed,$\kappa = 5$)	25.53	10.05	5.72	9.66
Viterbi (Pos,$\kappa = 7$)	21.66	9.32	5.69	8.95
Viterbi (BIC,$\kappa = 7$)	24.03	9.48	5.35	9.08
Viterbi (Pos,BIC,$\kappa = 7$)	22.80	6.80	4.27	7.05
Ground-truth change pts	11.09	4.05	2.46	4.00

Table 1: *Diarization error rates of different setups*

5.3. Joint segmentation and classification

Our approach proposed in this paper carries out segmentation and classification in one step using a Viterbi decoder. Table 1 contains the results for different variants: Combining position and BIC information as described by eq. (4) gives good results ("Viterbi (Pos,BIC,$\kappa = 1$)"). The DER can even be further reduced by weighting the contribution of the transition score relative to the observation score. For "Viterbi (Pos,$\kappa = 7$)" only the position feature is used to compute the transition scores, while for "Viterbi (BIC,$\kappa = 7$)" only the BIC information is used, in both experiments we used $\kappa = 7$.

Figure 5 shows the DER as a function of the weight κ, see eq. (5). Using the best value $\kappa = 7$, position and BIC information gives the best performance and clearly outperforms the sequential approach and also the best Viterbi decoder with fixed transition probabilities ("Viterbi (Fixed,$\kappa = 5$)"), see entry "Viterbi (Pos,BIC,$\kappa = 7$)" in Table 1.

The performance of the Viterbi decoder also depends on the allowed maximum delay $\tau_{\max}$, after which a traceback is forced. This partial traceback is needed to adhere to the real-time requirements of the system. In Figure 5 results for different delays $\tau_{\max}$ are given, showing that with a maximum delay of $3s$ good results are achievable and that the performance degradation is minor for delays $\tau_{\max} \geq 2s$.

6. Conclusions

In this paper we have presented an algorithm for parallel speaker segmentation and classification based on an HMM, where each state characterizes a speaker (or silence) and the transitions model the changes between the speakers. Appropriate observation and transition scores have been defined. The system is shown to outperform a two-stage approach, where first change detection is carried out and second speaker classification. The proposed method is particularly useful if iterative approaches which employ first tentative segmentation and clustering, which can later be refined by merging and resegmentation, are not affordable due to tough latency requirements.

If microphone array recordings are available, as is assumed here, evidence of speaker change can be further supported by speaker position or Direction-of-Arrival information derived from an adaptive beamformer. This further improves diarization performance.

7. Acknowledgement

This work has been supported by the European union project ”Amigo - Ambient intelligence for the networked home environment”. For more information visit ”www.amigo-project.org”

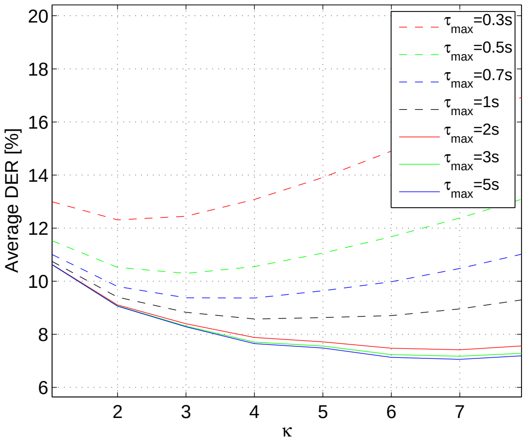

Figure 5: *Diarization performance using Viterbi(Pos,BIC,κ) with partial traceback. Average of whole database*

8. References

[1] S. Tranter, D. Reynolds, "An Overview of Automatic Speaker Diarization Systems", *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557-1565, Sept. 2006

[2] J. Schmalenstroeer, R. Haeb-Umbach, "Online Speaker Change Detection by Combining BIC with Microphone Array Beamforming", in *Proc. Interspeech 2006*, Pittsburgh, USA, Sept. 2006

[3] Amigo homepage: http://www.amigo-project.org

[4] S. Meignier et al., "Step-by-Step and Integrated Approaches in Broadcast News Speaker Diarization", *Comput. Speech Lang.*, no. 20, pp. 303-330, Sept. 2005.

[5] E. Warsitz, R. Haeb-Umbach, S. Peschke, "Adaptive Beamforming Combined with Particle Filtering for Acoustic Source Localization", in *Proc. ICSLP 2004*, Jeju, Korea, Oct. 2004

[6] ETSI ES 202 050 V1.1.3 , ”ETSI Standard Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms”, Nov. 2003

[7] M. Nishida, T. Kawahara, "Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing", *IEEE Trans. on Speech and Audio Processing*, Vol. 13, no. 4, July 2005

[8] S. Cheng, H. Wang, "A Sequential Metric-based Audio Segmentation Method via the Bayesian Information Criterion", in *Proc. Eurospeech*,

[9] A. Zolnay, R. Schlüter, H. Ney, "Extraction Methods of Voicing Feature for Robust Speech Recognition", in *Proc. EUROSPEECH*, Geneva, Sept. 2003

[10] CHIL - Computers In the Human Interaction Loop, ”http://chil.server.de”, 2006 Geneva, Sept. 2003

[11] E. Warsitz, R. Haeb-Umbach, ”Acoustic Filter-and-Sum Beamforming by Adaptive Principal Component Analysis”, ICASSP05, Philadelphia, USA, 2005