Online Speaker Change Detection by Combining BIC with Microphone Array Beamforming

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 consider the problem of detecting speaker changes in audio signals recorded by distant microphones. It is shown that the possibility to exploit the spatial separation of speakers more than makes up the degradation in detection accuracy due to the increased source-to-sensor distance compared to close-talking microphones. Speaker direction information is derived from the filter coefficients of an adaptive Filter-and-Sum Beamformer and is combined with BIC analysis. The experimental results reveal significant improvements compared to BIC-only change detection, be it with the distant or close-talking microphone.

Index Terms: speaker diarization, position estimation, beamforming, BIC.

1. Introduction

In speaker diarization or acoustic scene analysis information about ”who spoke when” is to be gleaned from recorded speech signals. While classical applications of this technology are in the field of automatic annotation of prerecorded audio or multimedia data [1], new emerging applications are found in the realm of telephone or video conferencing or for intelligent user services. In the latter case, information about the speaker is gathered in order to automatically establish user profiles and preferences and adapt an interface to an individual user [2], [3]. While the same technologies, such as speaker change detection and speaker recognition are used, there are however, remarkable differences compared to the aforementioned classical applications:

- The data are not prerecorded, but have to be processed online. Decisions on speaker change and identity should be taken with minimum latency.
- In order to enhance the usability of human-to-machine or human-to-human communication, distant microphones are preferred over close talking microphones.

Most of the literature on speaker change detection does not address these two issues. In a typical setup speaker change detection based on the Bayesian Information Criterion (BIC) delivers hypothetical change points which are then reconsidered using a clustering approach [4]. Such a two-stage batch procedure, which assumes that the whole database is available before processing starts, is not applicable in an online streaming scenario, where latency (and also computational effort, to some extent) is crucial. To reduce delay other methods have to be found to asses the hypothetical change points proposed by BIC. In our case we utilized Direction-of-Arrival (DoA) information obtained from an adaptive microphone array beamformer. Using this information to filter the hypothetical change points large performance improvements could be obtained.

In the next section we briefly revisit speaker change detection by BIC. Section 3 gives a short review of how speaker direction information is derived from the otherwise blind adaptive microphone array beamformer proposed in [5]. After describing the experimental setup and the database in Section 4, Section 5 contains the experimental results and discusses the combined BIC-beamformer design.

2. Bayesian information criterion

The basic idea for identifying possible change points is to reformulate the task as a problem of model selection [1], [6]. Given the set of feature vectors $\boldsymbol{X}(n) = \{\boldsymbol{x}(n), \dots, \boldsymbol{x}(n+M-1)\}$, where M denotes the window size, two models for $\boldsymbol{X}(n)$ are proposed. The null hypothesis H_0 states that all feature vectors are independent and identically distributed (i.i.d.) samples drawn from the same Gaussian $\mathcal{N}(\boldsymbol{X}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, while in the alternative hypothesis H_1 the first b vectors $\boldsymbol{x}(n), \dots, \boldsymbol{x}(n+b-1)$ are assumed to be drawn from $\mathcal{N}(\boldsymbol{X}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and the remaining from the Gaussian $\mathcal{N}(\boldsymbol{X}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Here we used a sliding window approach, where $b = \frac{M}{2}$ is fixed and the window always contains the M most recent feature vectors [7].

Since the parameters of the Gaussians are not known they have to be estimated from the data themselves. This gives the following log-likelihood of the data $\boldsymbol{X}(n)$ under the hypotheses H_0 and H_1 respectively [1]:

$$\log p(\boldsymbol{X}(n)|H_0) = -\frac{M}{2}\log|\widehat{\boldsymbol{\Sigma}_0}| - \frac{MD}{2}(\log(2\pi)+1) \quad (1)$$

$$\log p(\boldsymbol{X}(n)|H_1) = -\frac{M}{4}\log(|\widehat{\boldsymbol{\Sigma}}_1||\widehat{\boldsymbol{\Sigma}}_2|) - \frac{MD}{2}(\log(2\pi)+1) \quad (2)$$

where $\widehat{\boldsymbol{\Sigma}}_0$, $\widehat{\boldsymbol{\Sigma}}_1$, $\widehat{\boldsymbol{\Sigma}}_2$ are estimates of the respective covariance matrices. Since we are interested in low latency and thus small window sizes, all covariance matrices are assumed to be diagonal.

Several decision rules have been proposed for BIC. In our experiments we found that the metric decision criterion for detecting change points [8] was quite robust. A possible speaker change point is detected, if the differences between the BIC value of a local maximum ($\mathit{BIC}(max)$) and the corresponding minima ($\mathit{BIC}(min_L), \mathit{BIC}(min_R)$) are both larger than a threshold δ, see Fig.1. In this criterion the terms which are independent of the input

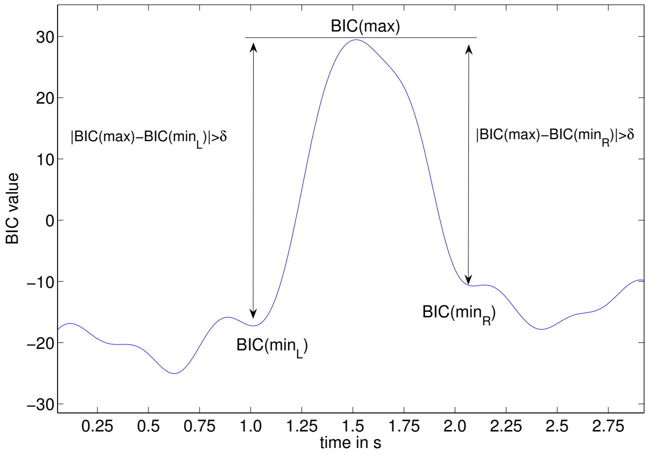

Figure 1: *Metric decision rule*

data, e.g. the model complexity, are unimportant and it is sufficient to consider

$$BIC(\boldsymbol{X}(n)) = -\frac{M}{2}\log|\widehat{\boldsymbol{\Sigma}}_0| + \frac{M}{4}\log(|\widehat{\boldsymbol{\Sigma}}_1||\widehat{\boldsymbol{\Sigma}}_2|). \quad (3)$$

3. DoA estimation by FSB beamforming

In [5] we presented a Filter-and-Sum Beamformer (FSB) whose coefficients are adapted such that they form the principal component of the power spectral density matrix of the microphone signals. While the adaption works blindly, i.e. does not require the estimation of the Direction-of-Arrival (DoA) of the desired signal, DoA information can be derived from the filter coefficients themselves [9].

The FSB output signal $y(n)$ is given by

$$y(n) = \sum_{m=1}^{M} x_m(n) * f_m(-n) \quad (4)$$

where $x_m(n)$ is the m-th microphone signal and $f_m(n)$ is the m-th filter impulse response. For the Direction-of-Arrival information the signal delay between two microphone channels must be estimated. This can be done by calculating the cross-correlation

$$\phi_{ij}(\lambda) = f_i(-\lambda) * f_j(\lambda) \quad (5)$$

between the i-th and the j-th filter impulse response. Here $\lambda = kT$ denotes the lag, which is an integer multiple of the sampling period T. As the FSB filters can model fractional delays a resolution below the sampling period can be obtained by interpolation. Let this interpolated cross-correlation be called $\tilde{\phi}_{ij}(\tau)$, with $\tau = lT'$ and $T' < T$. The delay between the signals at microphones i and j is then estimated by

$$\delta_{ij} = \underset{\tau}{argmax}|\tilde{\phi}_{ij}(\tau)| \quad (6)$$

and the Direction-of-Arrival can be calculated by

$$\alpha_{ij} = \arcsin\left(c \cdot \frac{\delta_{ij}}{s_{ij}}\right) \quad (7)$$

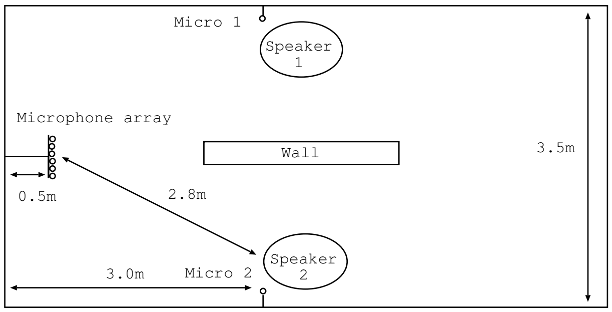

Figure 2: *Database recording setup*

where c is the speed of sound and s_{ij} is the distance between the i-th and the j-th microphone.

The accuracy is further enhanced by calculating the mean Direction-of-Arrival $\overline{\alpha}$ by

$$\overline{\alpha} = \frac{2}{M^2 - M}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\alpha_{ij} \quad (8)$$

over all possible microphone pairs (i, j) of the linear array. Outliers in the Direction-of-Arrival information are rejected through a median filter.

The experiments showed a mean adaptation time for the beamformer of approximately 40 feature vectors (0.4 s) and the accuracy improved with the increasing number of microphones.

4. Experimental setup and database

Most of the existing databases on speaker change detection contain either single-channel recordings or recordings of multiple, distributed microphones. Recordings of microphone arrays with known, fixed microphone array geometry are hardly available. Note that the CHIL project is concerned with similar research issues as studied here and that a large database will become available in the course of the project [10].

We carried out recordings in a setup as depicted in Fig.2 to compile our own database. Two speakers sitting in a medium-sized room are alternatingly reading aloud passages of newspaper articles. A passage has a minimum duration of 2.5 s. Audio signals are captured by two near-field microphones (”Micro 1”, ”Micro 2”; dynamic microphone with 0.2 m distance towards speaker), one being close to the first, the other close to the second speaker. Further, a linear 6-element microphone array (”Microphone array”; prepolarised condenser microphones) with an interelement distance of 0.05 m is placed at a distance of approximately 2.8 m from the speakers. The array is mounted at a height of 1.2 m to be representative of being mounted on top of a display. The whole setup should feature a typical scenario where people communicate either with a system or a remote partner via distant microphones and displays.

Since handlabeling the data is an expensive and error-prone task we tried to choose a setup which allows for close-to-perfect automatic labeling of speaker changes. To this end we placed a removable wall between the speakers resulting in an attenuation of -13 dB of the speech of the second speaker compared to the signal of the speaker sitting next to the near-field microphone. An adaptive interference canceler was used to further attenuate the other speaker's signal down to -20 dB. Automatic labeling of change points can now be achieved by a simple energy criterion and by incorporating the aforementioned constraints on the minimum speaker duration. These labels are then used as ground truth change points for both the near-field and the far-field microphone signals, as all signals are captured simultaneously.

The database contains about 1.5 h of spoken texts from a total of 5 male and 5 female speakers. A mixed single-channel signal is derived by adding the two processed single-channel near-field microphone signals. Speaker change detection experiments are now carried out with this mixed signal and with the signal of the distant microphone array.

5. Experimental results

Speaker change detection exhibits two error classes: missed detections (MD) and false alarms (FA). For the evaluation we adopted the definition of missed detection rate (MDR) and false alarm rate (FAR) proposed in [11]:

$$MDR = \frac{100 \cdot \textit{number of MD}}{\textit{number of change points}} \% \qquad (9)$$

and

$$FAR = \frac{100 \cdot \textit{number of FA}}{\textit{number of change points} + \textit{number of FA}} \%. \qquad (10)$$

A change point at time k is counted as missed, if no change point is detected in the range of $[k-1s, \ldots, k+1s]$, i.e. a two-second window around the change point. Further, the false alarm count is incremented, if a change point is detected at time k, although no change points occurs in the range of $[k-1s, \ldots, k+1s]$.

In all experiments described here we used a feature vector consisting of 39 Mel Frequency Cepstral Coefficients (MFCC) and 36 Linear Predictive Cepstral Coefficients (LPCC). The ETSI advanced feature extraction front end [12] was used to compute the MFCC features from 16 kHz input data. LPCC were computed from the enhanced signal after the two-stage Wiener Filter. The combined feature vector yielded slightly better results than MFCC or LPCC alone (see Fig.3). The receiver operating characteristic (ROC) depicted in Fig.3, 4 and 6 were obtained by varying the metric decision threshold δ in the range of $[10, \ldots, 80]$.

The sliding window approach briefly described in section 2 yields one BIC-value per input frame. This sequence of BIC values is filtered by a fifth-order Chebycheff filter to smooth the BIC trajectory. The filter has a group delay of 56 feature vectors (0.56 s), adding to the delay already introduced by the blockwise processing. However the filtering was considered necessary in order to better identify relevant local maxima in the BIC stream.

5.1. Window size

In a first set of experiments we determined the optimal window size M using the mixed signal obtained from the near-field microphones. In Fig.4 it can be seen that a performance optimum is achieved for $M = 80$. If the window size is too small, the covariance terms in eq. (3) cannot be estimated reliably, and if the window size is too large the BIC trajectory is too smooth making it difficult to reliably identify change points. Conducting the same experiments with the signals of the distant microphones an increased optimum window size of $M = 100$ was observed. The reason is probably that due to the worse signal-to-noise ratio more smoothing is necessary.

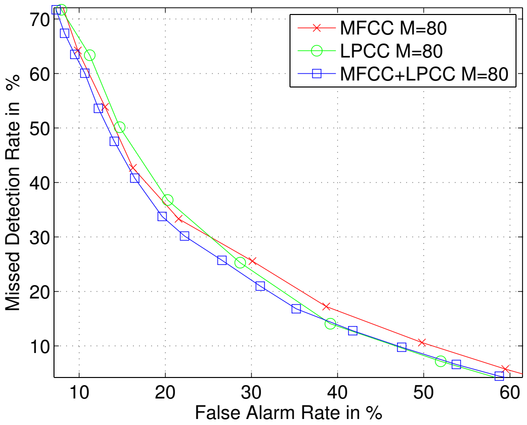

Figure 3: *ROC for different feature vectors*

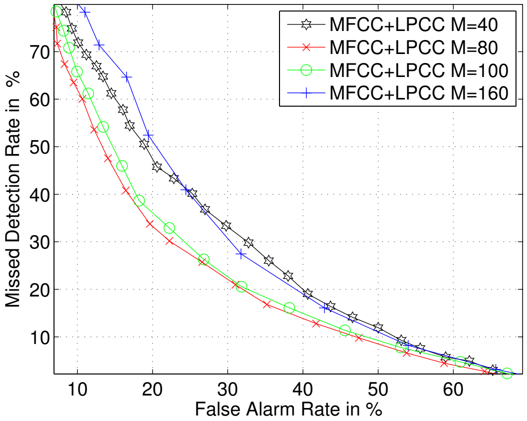

Figure 4: *ROC for different window sizes (near-field microphones)*

5.2. Combining BIC with direction information

Direction-of-Arrival (DoA) information obtained from the microphone array as outlined in Section 3 is used to improve the change detection accuracy achievable with distant microphone signals. The underlying assumption here is that the speakers are spatially apart and that they do not move fast while speaking. Then speaker changes indeed may be indicated by observed changes in Direction-of-Arrival, see Fig.5.

The direction information and the BIC evaluation were combined as follows: a change point was accepted if and only if both BIC indicates a speaker change and DoA evaluation indicates a direction change. If only one of the two hypothesizes a change it was considered a false alarm, be it because of poor BIC values or erroneous DoA estimates caused by low signal-to-noise ratio, reflections or speaker movements. Although not present in our database, speaker movements may well occur in practice and therefore an observed DoA change was not considered sufficient to decide on a speaker change.

Fig.6 compares the speaker change detection performance for different setups. The performance obtained from running BIC on

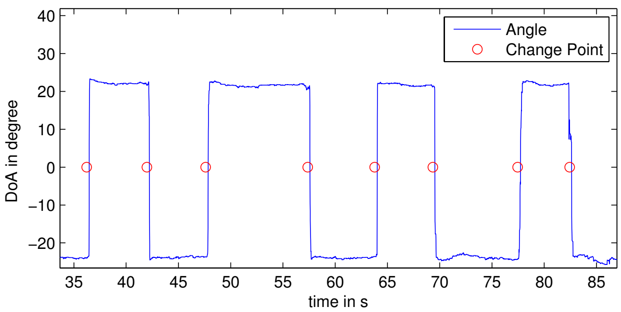

Figure 5: *Example for DoA estimation and speaker change detection*

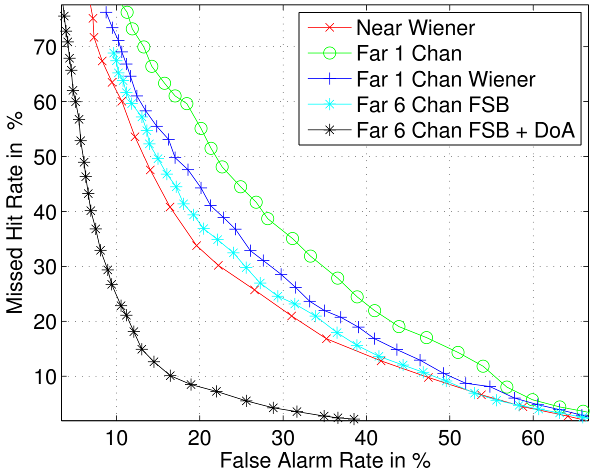

Figure 6: *ROC for different setups*

the mixed signal of the near-field microphones may serve as a baseline (”Near Wiener”). It can be seen that using the signal of a single microphone of the array (”Far 1 Chan Wiener”) greatly degrades accuracy. The worst case is a non-enhanced far-field microphone signal (”Far 1 Chan”). Interestingly, running BIC on the enhanced microphone array output signal, a performance is achieved which comes close to the performance obtained with near-field microphones. Drastic improvements even beyond the performance of near-field microphones are obtained when using Direction-of-Arrival information in this latter setup (”Far 6 Chan FSB + DoA”): an equal error rate of 13.8 % was obtained compared to 25.6 % with the near-field microphone.

6. Conclusions

In this paper we have investigated the impact of distant microphones on the performance of speaker change detection systems, based on the BIC criterion. Using the enhanced signal of a microphone array beamformer the loss compared to the performance achieved with close-talking microphones could be recovered almost completely. Combining BIC with Direction-of-Arrival information, which is a byproduct of the adaptive beamformer used, significant performance improvements could be obtained resulting in more reliable change point detection than was achievable with a close-talking microphone. Speaker changes are detected with an overall delay of about 1 second making it a valuable source of information for applications demanding real-time such as videoconferencing or intelligent user systems.

7. Acknowledgments

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”.

8. References

[1] C. Wu, C. Hsieh, “Multiple Change-Point Audio Segmentation and Classification Using an MDL-Based Gaussian Model”, IEEE Trans. on Audio, Speech and Language Proc., Vol 14, NO. 2, March 2006

[2] R. Haeb-Umbach, B. Kladis, J. Schmalenstroeer, “Speech Processing in the Networked Home Environment - A View on the Amigo Project”, Proc. Interspeech 2005, Lisbon

[3] Inspire Project Homepage, ”http://www.knowledge-speech.gr/inspire-project/”, 2003

[4] X. Zhu, C. Barras, S. Meignier, J. Gauvain, “Combining Speaker Identification and BIC for Speaker Diarization”, Proc. Interspeech 2005, Lisbon 2005

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

[6] 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

[7] S. Cheng, H. Wang, “METRIC-SEQDAC: A Hybrid Approach for Audio Segmentation”, Proc. Interspeech 2005, Lisbon

[8] S. Cheng, H. Wang, “A Sequential Metric-based Audio Segmentation Method via The Bayesian Information Criterion”, Proc. Eurospeech 2003

[9] E. Warsitz, R. Haeb-Umbach, S. Peschke, “Adaptive Beamforming Combined with Particle Filtering for Acoustic Source Localization”, Proc. ICSLP 2004, Jeju, Korea, 2004

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

[11] P. Delacourt, C. Welkens “DISTBIC: A Speaker-based segmentation for Audio Data Indexing”, Speech Communications, Vol. 32, pp111-126, 2000

[12] 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