ADAPTIVE BEAMFORMING IN SPEAKER DIARIZATION

Tobias Grosser, Aoys Elsafadi, David Hübner, Ronald Böck, Andreas Wendemuth
OvGU, Otto von Guericke University Magdeburg
Institute for Electronics, Signal Processing and Communications
Universitätsplatz 2, 39106 Magdeburg

In this paper we outline the developement of a system for speaker diarization within group meetings to prepare those for automatic speech recognition. Since the acceptance of headsets for group meeting recordings is very limited a more comfortable approach is to use beamforming and to steer an adaptive beamformer towards the person who is speaking. In our setup we used 5 room microphones plus 4 headsets providing reference signals for performance evaluations. Recordings of virtual group meetings with limited vocabulary were done. Choosing from a vocabulary of only 1700 words, multiple sentences were constructed which were then spontaneously used within the group meeting recordings. Our goal is to capture the dynamics of group meetings but with limited vocabulary to make robust automatic speech recognition feasible.
We plan to combine two methods to prepare for speaker diarization, blind source separation and adaptive beamforming in combination with direction of arrival estimation for each speaker. In speaker diarization the goal is to find the segments of time in which each meeting attendee is speaking. Time segments in which two or more speakers are speaking are especially challenging. We plan to use blind source separation in conjunction with an adaptive beamformer to steer towards all speakers simultaneously and achieve one audio recording per speaker.

1 INTRODUCTION

In the last decades tremendous improvements in acoustic modeling of speech for Automatic Speech Recognition (ASR) were made. Nonetheless, the interaction between humans and computer systems via an ASR interface still regulary leads to unsatisfied users. The present paper presents preliminary results of multiple Beamforming algorithms, a Blind Source Separation (BSS) algorithm and an Automatic Speech Recognition engine based on the Hidden Markov Toolkit (HTK). It is a feasibility study to build up an Automatic Speech Recognition system for

group meetings with limited vocabulary. The vocabulary of the recorded group meetings is restricted to a set of 1700 words which were used to construct sentences in four categories namely nature (animals and plants), food and cooking, train traveling and bits and pieces. Participants of the one hour lasting meetings were asked to use only these sentences, subordinate clauses taken from those or a small number of predefined linking words. Our intention is to provide insights into the dynamics of group meetings which might lead to more robust and adaptive ASR systems.
In Fig. 1 a survey of the proposed speaker diarization system is shown in which BSS, Direction of Arrival Estimation and Beamforming are expected to show mutual enhancement and complementary aspects. The extracted audio signals from Blind Source Separation and Beamforming are merged in the subsequent fusion-stage the purpose of which is to compensate weaknesses of the BSS algorithm by the robustness of the Beamforming and vice versa. As the last stage before ASR speaker model based speaker diarization will be performed.

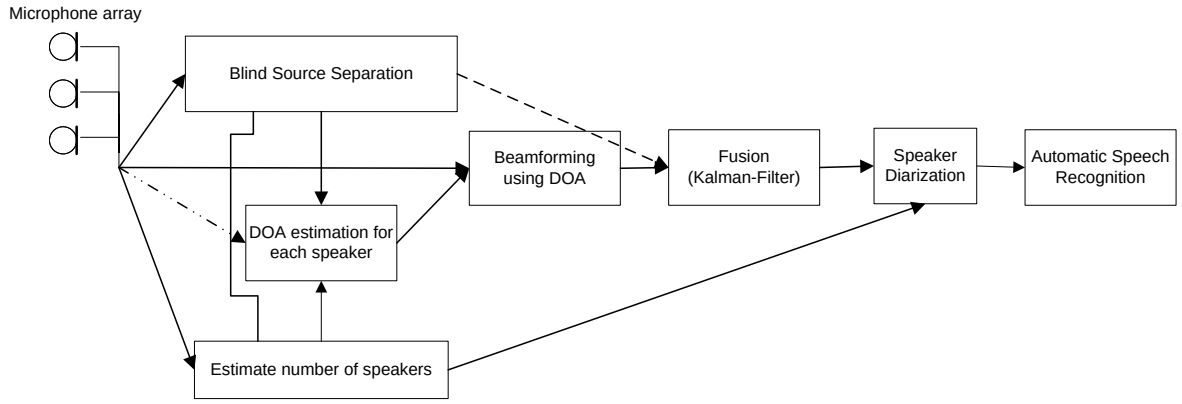

Figure 1: Survey of the proposed speaker diarization system

2 Beamforming

2.1 Simulating the array

In order to be able to intuitively rate the separation performance of a linear equidistant microphone array on two sources we first simulated our setup using 6 microphones and the geometry given in Fig. 4. The emitting signal is a single sine wave with constant frequency. One can see that the directivity for the single sine wave is good. Reception of signals is restricted to a circular arc of maximal 30°. Due to the broad spectrum of speech and resulting changing interference patterns of the microphone array over frequency, the directivity of the beamformer in practice is expected to decrease significantly.

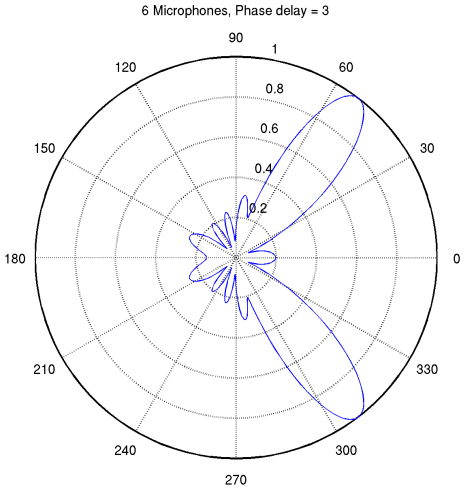

Figure 2: Simulated audio lab: Polar response of simulated equidistant line array consisting of 6 microphones. The beam is directed towards the first loudspeaker which is at 55°.

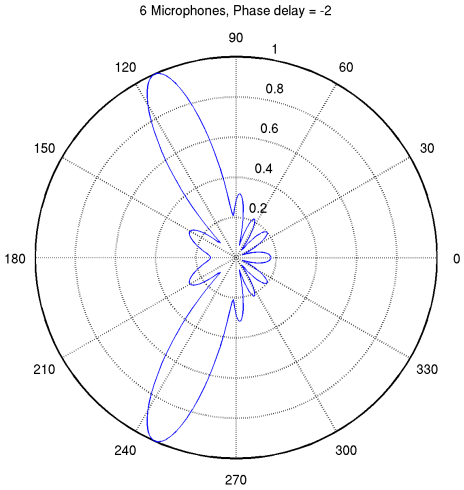

Figure 3: Simulated audio lab: Polar response of simulated equidistant line array consisting of 6 microphones. The beam is directed towards the second loudspeaker which is at 115°.

2.2 Beamforming-Algorithm evaluation

In order to assess the performance of different common beamformer methods we conducted recordings in our audio lab with cushion walls. Two speakers, one female and one male, were recorded separately via an headworn dynamic microphone. The first speaker counted from one to ten in his own rythm. When the second speaker was recorded she was listening to the voice

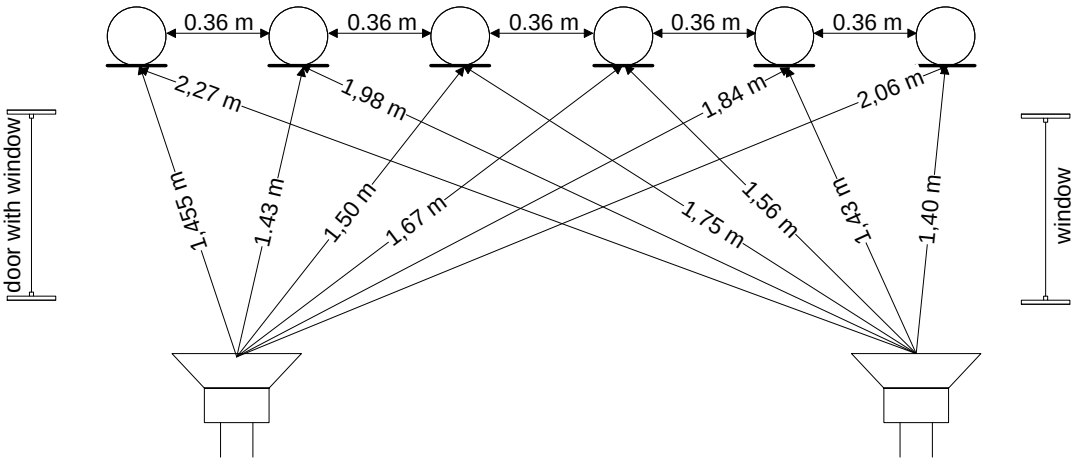

Figure 4: Geometry of the audio lab

of the first speaker over headphone and by that synchronizing with the first speakers voice. Both recordings were exported as one stereo audio file with one speaker only on the left and the other speaker only on the right side. We then used a microphone array with 6 elements to record the played back stereo recording emitted from two high quality loudspeakers within the audio cabin. The spacing between the microphones of the equidistant line array was $0.36m$.
This approach has two major advantages compared to simultaneous recording of both speakers. First the single recordings of the speakers can serve as reference signals. Moreover the loudspeakers have unlike speaking humans a constant position.
For performance assessment we used three criteria. The Source to Distortion Ratio (SDR) which is defined as

$$SDR := 10\log_{10}\frac{||S_{target}||^2}{||e_{interf}+e_{noise}+e_{artif}||^2} \tag{1}$$

sets the average power of the target signal in relation to all kind of distortion signals. On the other hand the commonly used Source to Interference Ratio (SIR)

$$SDR := 10\log_{10}\frac{||S_{target}||^2}{||e_{interf}||^2} \tag{2}$$

compares the target signal with the interference signal and by that neglects spectral power coming from either noise or artifacts of the algorithm. We found that the SDR correlates much better with the perception of seperation results than the SIR when listening to the audio files. As another criterion which reflects the acceptance of separation results we used the correlation coefficient introduced by Bravais and Pearson:

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \quad \text{with} \quad -1 \leq \rho \leq 1. \tag{3}$$

A positive correlation is indicated by positive values of ρ, e.g. if X increases Y increases as well. When ρ is negative a negative correlation exists, e.g. if X increases Y decreases. A value

around zero indicates that there is no linear correlation between the variables.
For performance assessment and in particular decomposition of the results of the beamforming algorithms we used matlab scripts from the Signal Separation Evaluation Campaign [2]. The function

$$[s_{target}, e_{interf}, e_{artif}] = bss\_decomp\_gain(se, i, S) \tag{4}$$

assumes that a given estimate $\hat{s}(t)$ of a source $s_i(t)$ can be decomposed into a sum

$$\hat{s}(t) = s_{target}(t) + e_{interf}(t) + e_{artif}(t) \tag{5}$$

where $s_{target}(t)$ corresponds to the target source $s_i(t)$, $e_{interf}(t)$ accounts for the interferences of the unwanted sources, and $e_{artif}(t)$ comprises all unwanted artifacts induced by the separation algorithm.
To measure the effectiveness of the different beamformer methods for the signal separation task we used the evaluation criteria defined above not only for the output signal of the beamformer but also for the individual microphone signals. We call a beamformer to be *effective*, only if the SDR and the correlation coefficient ρ of its output is higher than the maximum value of those criteria of all individual microphones, where the criteria are calculated on a time shifted version of the microphone signals to compensate for sonic delay. The time shift is calculated by computing the cross-correlation between the reference signal and the recorded signal of the individual microphone. The time shift leading to the maximum cross-correlation is used to shift the microphones' signal before the evaluation criteria are calculated.
In Table 1 the presence of each speaker within the single microphones can be concluded from the values in the table. The loudspeaker of speaker 1 is closest to mic 5 while mic 1 is the closest microphone to the loudspeaker of speaker 2. This is only partly reflected in the values of SDR, ρ and SIR, probably due to reflections of two windows in the audio cabin (Fig. 4).
In Table 2 we show results of 5 different beamformers. The first column ($Y_{DSBgeom}$) contains performance criteria of a simple Delay and Sum beamformer where time delays are calculated based on measurements of microphone and loudspeaker positions in R^3. The time delays are used to bring the target signal components into phase which leads to a constructive summation. The interferer signal components are likely to add, at least partly, destructively. The results show that the Delay and Sum beamformer performs effectively according to our definition for speaker 1. $Y_{DSBgeom}$ has $SDR = -7.76\,dB$ which is greater than $SDR = -12.25\,dB$ for the best single microphone *mic 3*. Equivalent its correlation coefficient is $\rho = 0.38$ which is greater than $\rho = 0.24$ for *mic 3*. On the other hand $Y_{DSBgeom}$ does not perform effectively for speaker 2 for which the correlation coefficient $\rho = 0.60$ is slidely less than the one of the best microphone *mic 2* which is $\rho = 0.61$.
The most effective beamforming algorithm in terms of our evaluation criteria is $y_{RobustGSC}$ ([1]) for which the $SDR = -7.25\,dB$ of speaker 2 is significantly higher than for the best microphone (mic 3: SDR=-12.25 dB) and the same holds for the correlation coefficient $\rho = 0.40$ compared to the best microphone (mic 3: $\rho = 0.24$).
The beamformer marked with $Y_{DSBxcorr}$ is a Delay and Sum Beamformer which uses cross-correlation to establish an optimal shift between each of the room microphones and the reference signal drawn from the headworn microphone of the person to be amplified. The time-shifted room microphones are then added to form the beamformers output.

Table 1: Audio lab: Signal to Distortion Ratio, Signal to Interference Ratio, and Pearson's correlation coefficient for recordings done in an audio lab with cushion walls. Speaker 2 is significantly louder than speaker 1.

		mic 1	mic 2	mic 3	mic 4	mic 5	mic 6
speaker 1	SDR	-18.10 dB	-15.35 dB	-12.25 dB	-15.63 dB	-12.39 dB	-12.70 dB
	SIR	-6.92 dB	-3.63 dB	-5.12 dB	-1.22 dB	6.59 dB	25.80 dB
	ρ	0.12	0.17	0.24	0.16	0.23	0.23
speaker 2	SDR	-5.20 dB	-2.37 dB	-4.56 dB	-6.96 dB	-8.02 dB	-10.60 dB
	SIR	22.47 dB	22.09 dB	27.96 dB	16,16 dB	18.44 dB	16.15 dB
	ρ	0.48	0.61	0.52	0.41	0.37	0.28

Table 2: Audio lab: SDR, SIR, and Pearson's correlation coefficient ρ for recordings done in our audio lab with cushion walls. $Y_{DSBgeom}$ is a simple Delay and Sum Beamformer using the measurements from Fig. 4 . y_{GSC} is a Generalised Sidelobe Canceller. $y_{RobustGSC}$ is a modified version of the GSC proposed by Hoshuyama et al. [3]. $Y_{DSBxcorr}$ is a Delay and Sum Beamformer based on cross-correlation with the reference signal.

		$Y_{DSBgeom}$	y_{GSC}	$y_{RobustGSC}$	$Y_{DSBxcorr}$
speaker 1	SDR	-7.76 dB	-7.56 dB	-7.25 dB	-7.60 dB
	SIR	44.48 dB	13.15 dB	13.95 dB	10.96 dB
	ρ	0.38	0.39	0.40	0.39
speaker 2	SDR	-2.47 dB	-1.41 dB	-0.87 dB	-0.65 dB
	SIR	25.86 dB	26.28 dB	25.59 dB	27.95 dB
	ρ	0.60	0.64	0.67	0.68

3 BSS

The main task of Blind Source Separation BSS is to extract acoustic sources of mixed signals by canceling the interference of the other sources involved in the mixing signal. An identical copy of the original source at the output stage of the BSS is not expected.

In most real acoustic applications signals are received at microphone side in different arrival time. That means they suffer from delay according to multi propagation paths and from reflections at the room walls. According to this fact, the received signals in a real acoustic application are in most cases convolutive signals of the propagated signals from the speakers to the microphones. However, many Blind Source Separation based Algorithms are normally used to separate these convolutive signals. The Cocktail-party is a common example for convolutive signals, in which simultaneously talking speakers in a room are involved and recovering a certain speech is required. On the other hand, if all the signals arrive at the microphone's side at the same time the signals are called Instantaneous-Mixed-Signals.

Moving from time domain into frequency domain to deal with convolutive signals can be achieved by using the Fast Fourier Transform (FFT). Depending on the number of available sources and microphones we can distinguish between the following scenarios: 1. Extracting/Separating signals from a mixed signal by using just a single microphone. In this case a possible way to separate the mixed signals (unknown number of sources) is to use the Head-Related-transfer function HRTF altogether with prior knowledge of the source statistics. Another related method for separating mixed signals of unknown number of sound sources by using a single microphone depends on Auditory Scene Analysis and Psycho-Acoustics. 2.TITO System: A TITO System (Two-input-two-output) is a special case of the MIMO (Multi Input Multi Output) system in which the mixing signals of two sources are recorded at two microphones. In under-determined Systems the number of speakers is larger than the number of the available microphones. Under-determined mixed signals are very difficult to separate even under perfect knowledge of the mixing system.

3.1 Practical Experiment

In this part we experimented with the FastICA-based BSS algorithm to compare the results with Beamforming. Our FastICA-Based BSS tool requires that the number of extracted signals equals the number of used microphones. For that reason we used only 2 microphones.

Used-Microphones	Extracted Signal (Speaker)	SDR in dB
3+4	1	-8.2220
3+4	2	-5.8812

4 Speaker diarization

Speaker diarization is the process of finding the segments of time within a meeting in which each meeting participant is talking. It is a a necessary procedure before automatic speech recognition (ASR) of group meetings can be achieved. ASR systems can be adapted on individual voices of a meeting only if it is known when the equivalent person is speaking. If all participants are recorded via an headworn microphone, speaker diarization can be achieved by applying a simple threshold to the signal energy of each channel. This, however, is seldom the case. Commonly used meeting recording setups contain a significant amount of cross talk between the channels meaning that there are many voices plus noise audible in each microphone signal. Speaker diarization in this case is harder to achieve. Hidden Markov Models ([4]) and generally methods from the field of *speaker identification* can be used to model the characteristics of the different speakers' voices which gives good results in a noise free environment and if only one speaker is speaking at a time. But what can be done if the recording equipment allows cross-talk and several speakers are audible simultaneously?
We have seen that methods from Beamforming and Blind Source Separation can be used to enhance the audio signal for individual voices to some extend. Unfortunately huge Beamforming arrays are usually not available during group meetings due to the exhaustive costs and the improvement in SDR for small arrays is insufficient to perform robust speaker diarization.

Moreover the results from Blind Source Separation methods on real convolutive recordings are very limited. Often performance is measured only for either instantaneous mixtures or artificially convoluted mixtures. For many simultaneously recorded voices BSS is not feasible yet.
In our upcoming research we will focus on finding parameters which can be used to characterize both the current status of a group meeting and the speaking style of individual speakers in terms of speech rythm and topic. Frequency and length of contributions of individual speakers could be such parameters and we will try to use these to make predictions on how probable the next contribution will be done by a particular speaker.

5 Automatic speech recognition

Our speech recognition engine is based on the Hidden Markov Toolkit (HTK) which is developed at the Cambridge University Engineering Department. It is a triphone system trained on the Kiel Corpus of read speech with a vocabulary of about $v = 1700$ words. The training material is transcribed phonetically with a set of $m = 39$ monophones. We used the HTK tools *HLEd* and *HHEd* to convert the monophone transcriptions into an equivalent set of word internal triphones. Each phoneme with a specific predecessor and successor phoneme builds one triphone model. Theoretically the number of possible triphones is $m^3 = 39^3 = 59319$. Since training material even for the biggest ASR speech corpora is not large enough to estimate robustly the parameters of so many models, HTK provides a convenient way of restricting the number of triphone models to those occuring in the training material which is 3109 in our case. If a specific triphone is included in the test corpus but was not in the training corpus it is constructed by tree based clustering during the recognition phase. Tree based clustering is a means of sharing model parameters like transition matrices from the Hidden Markov Models (HMM) and means and variances of the Gaussian Mixture Models accounting for the emission probabilities of the HMMs. Using our system we achieved a word accuracy of 90.46% using a bigram language model. The vocabulary size of the test set included 365 words.
The reported results were obtained on high quality non-spontaneous speech. Our future goals is to combine this ASR system with a robust speaker diarization method to get good performance on the simulated group meetings which were constrained to the vocabulary of the training material of our ASR system.

6 SUMMARY

In this paper we presented a skeleton of a speaker diarization system for group meeting recordings and results of an automatic speech recognition system based on a vocabulary of 1700 words which achieved a word accuracy of 90.46%. While we could not yet present a working diarization system we showed promising results for different beamforming algorithms.

7 Outlook

In the future we plan to use our simulated group meeting recordings to focus on interhuman interaction by analysing speech-markers for topic changes within the conversation. In particular speech-markers for misunderstandings, halting speech, and individual background which all

can lead to topic changes are of interest when dealing with speaker diarization. We will try to derive relevant parameters for speaker models which can be used to enhance speaker diarization. Moreover parameters to distinguish between different conversational types, like halting speech and spirited conversation will be investigated.

References

[1] J. Bourgeois and W. Minker. *Time-Domain Beamforming and Blind Source Separation*. Springer, 2009.

[2] C. Févotte, R. Gribonval, and E. Vincent. "Bss eval toolbox user guide." In *IRISA Technical Report 1706, Rennes, France, April 2005. http://www.irisa.fr/metiss/bss eval/*. 2005.

[3] O. Hoshuyama and A. Sugiyama. "Robust adaptive beamforming." In *Ed. Michael Brandstein and Darren Ward: Microphone Arrays : Signal Processing Techniques and Applications, New York 2001*.

[4] I. Shahin. "Speaker identification in the shouted environment using suprasegmental hidden markov models." *Signal Processing*, 88(11), 2700–2708, 2008.