MULTI-CHANNEL SPEAKER LOCALIZATION AND SEPARATION USING A MODEL-BASED GSC AND AN INERTIAL MEASUREMENT UNIT

Mehdi Zohourian, Alan Archer-Boyd, and Rainer Martin

Institute of Communication Acoustics
Ruhr-Universität Bochum, Germany
{mehdi.zohourian | alan.archer-boyd | rainer.martin}@rub.de

ABSTRACT

In this paper we propose a novel multi-channel algorithm to separate simultaneous speakers in an environment where the microphone array is subject to movement. When the microphones are mounted to a person's head, for instance, the movements can lead to ambiguities with respect to the sources and to distortions in the processed signal. The proposed system estimates the direction-of-arrival of the speaker's signals relative to the array and updates these estimates using an inertial measurement unit (IMU). A GMM-based localization model is used to compute the posterior probabilities of source activity in each time-frequency bin and its parameters are re-estimated during array movements. Then, a model-based generalized side-lobe canceler (GSC) whose components are continuously updated, is employed for the separation of sources. For various speeds of microphone array rotation, it is demonstrated that the IMU-based system delivers improved speech quality when compared to the baseline technique without IMU.

Index Terms— Multi-channel speech enhancement, beamforming, source separation

1. INTRODUCTION

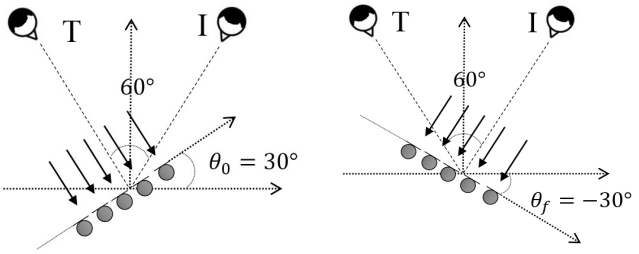
Adaptive beamforming is a multi-channel processing technique that is often used for the separation of acoustic sources and is closely related to blind techniques [1], [2], [3], [4]. When the position of the target sources is unknown the beamformer may be combined with an estimation of direction-of-arrival (DOA) for one or more relevant sources. This technique has been considered for various applications like hands-free mobile phone systems, man-machine interfaces, and assistive devices. However, the performance of these algorithms can be drastically reduced in highly dynamic environments where the sources and/or the array moves. While source tracking with fixed microphone arrays has received significant attention, the movement of the array, for example when worn by a moving listener, has received less consideration.

Adaptive beamformers use the spatial information provided by a microphone array in order to reduce interferencing sources and ambient noise. The generalized side-lobe canceler (GSC) [5] is the most prominent example of an adaptive beamforming algorithm. It performs best when the interferers originate from point sources and when there is only little reverberation. This approach has been extended by several authors to make it more robust, e.g. [6], and to cope with more general acoustic scenarios. For example, the TF-GSC [7] extends the GSC for arbitrary acoustic transfer functions. In [8] multi-channel eigenspace beamforming is used based on designing different linear constraints for extraction of a desired source. [9] contributes a model-based beamformer for source separation based on a localization algorithm, followed by a corresponding interference reduction scheme which is implemented in a GSC structure. This approach performs well when the position of microphones are fixed. However, as the microphone array moves in a multi-source scenario, the beamformer cannot easily track a specific target signal because of the ambiguities arising from array movements. There are also various techniques based on independent component analysis (ICA) that are used to improve the robustness of GSC beamformer e.g. [10], [11]. However, these blind methods need to track a large number of parameters to separate the target source from multiple concurrent sources and are susceptible to the fast array movements.

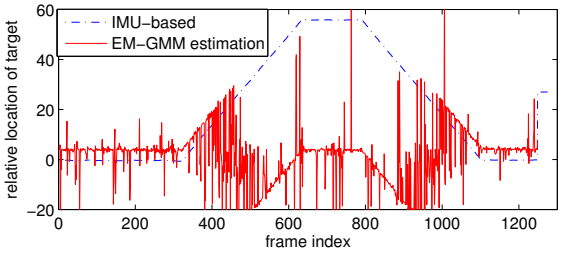
Several approaches have been also proposed for localizing and tracking multiple sources [12], [13]. Most of these approaches are based on Kalman or particle filtering (PF) applied to localization algorithms. These methods assume a specific model for source movement [12] or array movement [14] and are tailored to specific applications or consider just one source. Often, the localization and separation of multiple sources are treated separately and the generation of audio signals for the localized sources is not considered. To illustrate the ambiguites which arise when multiple sources are present and the array rotates, we present in Fig. 1 a simulation result for a scenario with two speakers T (target) and I (interferer) and a linear array of five microphones. Initially, the target speaker T is in broadside direction of the array as shown in Fig. 1 (a). After a rotation of the array by 60 degrees the interfering speaker I is in broadside direction, Fig. 1 (b). Then, the array is rotated back to its initial position. As we like to lock on to the target speaker, the direction-of-arrival relative to the array axis requires an update. Fig. 1(c) shows the performance of a conventional adaptive localization model (our baseline system) [9] for estimation of azimuth location of the target speaker relative to the array. We depict the estimated relative angle $\hat{\theta}(t) - \theta_0$ as well as the true relative direction $\theta(t) - \theta_0$ at each time frame. Here, $-\theta_0$ denotes the initial azimuth. Obviously, when the array starts moving the model cannot track the target position accurately and will shift to the DOA of the interference.

In this paper we extend the baseline model-based adaptive source separation technique [9] to a dynamic scenario and aim to extract one or more target speakers when the microphones rotates. As in [9] we use a Gaussian mixture model (GMM) based localization method that controls the beamformer during microphone movement. This model is adapted based on information obtained

This work has received funding from the People Programme (Marie Curie Actions) of the European Unions Seventh Framework Programme FP7/2007-2013/ under REA grant agreement PITN-GA-2012-317521.



(a) Initial position (b) Final position



(c) Estimated angles of speaker T w.r.t. initial position.

Fig. 1. Performance of the baseline system [9] for target speaker tracking in a two source scenario. The dashed line indicates the relative target-array angle as measured by the IMU. Note, that between frame indices 400 - 1000 the baseline sytem confuses the target direction with the direction of the interferer.

from the localization algorithm and a 9-axis inertial measurement unit (IMU). The IMU comprises a 3-axis accelerometer, 3-axis gyroscope and 3-axis magnetometer that provides the relative position of the moving object at each time step with respect to its initial position before movement. The adaptive localization model is then used for the adaptation of the fixed beamformer, blocking matrix and the adaptive noise canceler of a GSC beamformer.

The remainder of the paper is organized as follows: In Sec. 2 we describe the multi-channel signal model used in this paper. Sec. 3 will discuss the proposed system and will explain the GMM re-estimation step. Experimental results and conclusion will be described in Sec. 4 and Sec. 5, respectively. We like to point out that the evaluations reported here consider array rotations only.

2. MULTI-CHANNEL SIGNAL MODEL

In general, we consider an array of M microphones receiving signals from S speakers. Using the convolution operator $*$, the received signal at each microphone m is written as

$$x_m(n) = \sum_{i=1}^{S} s_i(n) * h_{im}(n) + \nu_m(n) \tag{1}$$

where $s_i(n)$ represents the source signal, $h_{im}(n)$ represents the room impulse response from source i to microphone m, $\nu_m(n)$ is the noise at microphone m, and n is the sampling index. In order to analyze signals in STFT domain, we take a K-point discrete Fourier transform (DFT) over overlapping windowed signal segments (frames). Using matrix notation and neglecting cyclic effects we obtain

$$\mathbf{X}(k,b) = \mathbf{H}^H(k,b)\mathbf{S}(k,b) + \mathbf{V}(k,b) \tag{2}$$

where $\mathbf{H}^H$ denotes the hermitian transpose of matrix $\mathbf{H}$

$$\mathbf{H}(k,b) = [\mathbf{h}_1(k,b), \mathbf{h}_2(k,b), .., \mathbf{h}_S(k,b)]^T \tag{3}$$
$$\mathbf{h}_i(k,b) = [h_{i1}(k,b), h_{i2}(k,b), .., h_{iM}(k,b)]^T$$

and the signal vectors are given by

$$\mathbf{X}(k,b) = [X_1(k,b), X_2(k,b), .., X_M(k,b)]^T \tag{4}$$
$$\mathbf{S}(k,b) = [S_1(k,b), S_2(k,b), .., S_S(k,b)]^T$$
$$\mathbf{V}(k,b) = [V_1(k,b), V_2(k,b), .., V_S(k,b)]^T .$$

In these equations, (k,b) indicate frequency index and frame index respectively. The received signals contain a mixture of target and competing speakers and are analyzed through the proposed algorithm that implements a GSC beamformer at its core.

3. PROPOSED ALGORITHM

Fig. 2 shows a block diagram of the proposed fully adaptive algorithm used to extract the desired speaker. Principally, the algorithm consists of two main parts: First, a GSC beamformer with a beamformer $\mathbf{W}_f(k,b)$ looking into the target direction, a blocking matrix $\mathbf{B}(k,b)$, and a noise canceler $\mathbf{W}_V(k,b)$ that attempts to extract target speech at each time-frequency bin. Secondly, a GMM-based estimation model that includes an IMU, a localization algorithm delivering $\hat{\theta}(k,b)$ and the computation of posterior probabilities $P_{\theta_T|\hat{\theta}}(k,b)$. All of them are utilized to update the GSC parameters while the array turns. Each part of this approach will be discussed in more detail in the following sections.

3.1. Generalized side-lobe canceler (GSC)

The proposed GSC beamformer is composed of a delay-and-sum beamformer with steering vector $\mathbf{W}_f(k,b)$, the blocking matrix $\mathbf{B}(k,b)$, and an adaptive noise canceler with coefficient vector

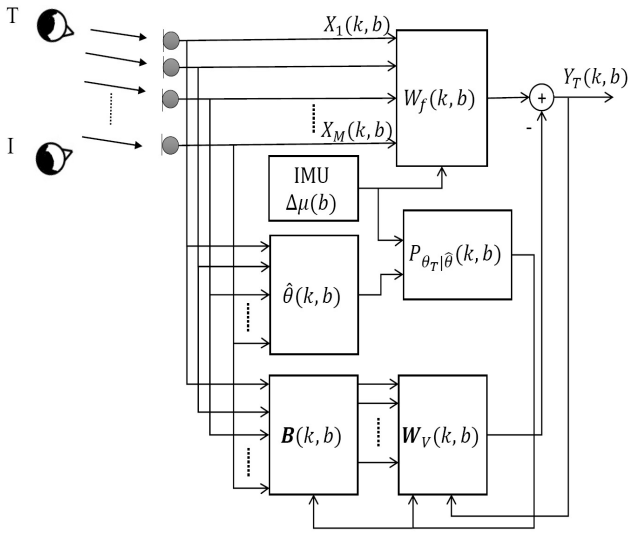


Fig. 2. Block diagram of the proposed method.

$\mathbf{W}_V(k,b)$. As in [9] a frame-wise posterior probability of the target speaker activity in each time-frequency bin is employed for the adaptation of the blocking matrix and the noise canceler. This scheme first estimates the target subspace

$$\mathbf{P}(k,b) = (1 - P_{\theta_T|\hat{\theta}}(k,b))\mathbf{P}(k,b-1) + P_{\theta_T|\hat{\theta}}(k,b)\frac{\mathbf{X}(k,b)\mathbf{X}^T(k,b)}{\left\|\mathbf{X}^2(k,b)\right\|} \quad (5)$$

and then computes the blocking matrix

$$\mathbf{B}(k,b) = D_{(M-1)M}\left(\mathbf{I}_{M\times M} - \mathbf{P}(k,b)\right) \quad (6)$$

where the operator $D_{ab}(\cdot)$ selects the first a rows and b columns of the matrix argument. $P_{\theta_T|\hat{\theta}}(k,b)$ is the probability of presence of speaker T in each time-frequency bin, $\mathbf{P}(k,b-1)$ is the target signal subspace, and $\mathbf{I}_{M\times M}$ is an identity matrix.

Then, we can write the output signal in the (k,b)th time-frequency bin as:

$$Y(k,b) = \mathbf{W}_f^H(k,b)\mathbf{X}(k,b) - \mathbf{W}_V^H(k,b)\mathbf{B}(k,b)\mathbf{X}(k,b) \quad (7)$$

where the adaptive noise canceler uses a normalized least mean squares (NLMS) algorithm

$$\mathbf{W}_V(k,b+1) = \mathbf{W}_V(k,b) + \alpha(k,b)\frac{Y^*(k,b)\mathbf{B}(k,b)\mathbf{X}(k,b)}{||\mathbf{B}(k,b)\mathbf{X}(k,b)||^2} \quad (8)$$

with adaptive step-size (α_f denotes a fixed stepsize factor)

$$\alpha(k,b) = \left(1 - P_{\theta_T|\hat{\theta}}(k,b)\right)\alpha_f \ .$$

Obviously, the noise canceler is adapted in those time-frequency bins which do not contain the target signal. In a multi-source scenario and when several source are to be extracted the above equations need to be executed for each source.

3.2. Adaptation of the GMM-based localization model

Our localization method is based on the *steered-response power with phase transform* (SRP-PHAT) method [15] which scans the acoustic environment to find the direction of arrival of the most powerful source for each time-frequency bin. Then, for each signal frame b, we estimated a Gaussian Mixture Model (GMM) whose means represent the direction of arrival of the acoustic sources. The means, together with the weights and variances, are estimated using the *expectation-maximization* (EM) algorithm [16]. Finally, the posterior probability of target activity in the (k,b)th bin is found as

$$P_{\theta_T|\hat{\theta}}(k,b) = \frac{\pi_T\mathcal{N}(\hat{\theta}(k,b)|\mu_T,\sigma_T^2)}{\sum_{i=1}^{C}\pi_i\mathcal{N}(\hat{\theta}(k,b)|\mu_i,\sigma_i^2)} \quad (9)$$

where $\hat{\theta}(k,b)$ is the direction of arrival estimated by localization algorithm in each time-frequency bin, $\mathcal{N}(\hat{\theta}(k,b)|\mu_T,\sigma_T^2)$ is the normal distribution describing the direction of the target source, and μ_i, σ_i^2, π_i are GMM parameters indicating mean, variance, and weighting factor for all sources. The number of components C is selected to exceed the assumed number of acoustic sources in order model diffuse ambient noise which typically results in model components of large variance. As outlined above, the posterior probability is instrumental for the estimation of the blocking matrix and the noise canceler.

When the microphone array is fixed the EM algorithm finds these parameters with high accuracy; however, as soon as the array moves in the multi-source scenario the estimation becomes erroneous, as it was demonstrated in Fig. 1. For solving this problem, we propose here to use an IMU for the adaptation of the mean values. Furthermore, we re-estimate the weights and the variances of the GMM for each frame and smooth all parameters via a first order recursive system

$$\begin{aligned}\mu_i(b) &= \mu_i(b-1) + \Delta\mu_i(b)\\ \bar{\pi}_i(b) &= (1-\beta)\bar{\pi}_i(b-1) + \beta\pi_i(b)\\ \bar{\sigma_i^2}(b) &= (1-\beta)\bar{\sigma_i^2}(b-1) + \beta\sigma_i^2(b)\ .\end{aligned} \quad (10)$$

Here, b is the frame number and $\Delta\mu_i(b)$ is the direction of arrival update obtained through the IMU. These smoothed parameters are then used instead of the instantaneous values to compute the posterior probabilities in (9). While the array rotates it is obviously necessary to adapt the means of the GMM, i.e. the mean direction of arrival. However, it is also important to re-estimate and smooth the other GMM parameters at each frame. In order to investigate the benefit of re-estimation, the log-likelihood function using the GMM model is evaluated during the array movement. The log-likelihood function is expressed for a single frame b of data as follows [17, Section 9.2.2]:

$$\ln p(\hat{\theta}(k,b)|\{\pi_i,\mu_i,\sigma_i^2\}_{1\ldots C}) = \sum_{k=1}^{K}\ln\{\sum_{i=1}^{C}\pi_i\mathcal{N}(\hat{\theta}(k,b)|\mu_i,\sigma_i^2)\} \quad (11)$$

where $\hat{\theta}(k,b)$ is the estimated source position in frequency bin k and frame b using SRP-PHAT [15], K is the number of frequency bins at each frame.

Fig. 3 then depicts the log-likelihood for a longer signal with ten successive cycles of array rotation. In this figure, we consider three different methods: without using the IMU (method 1), with using the IMU and fixed GMM weight and variance parameters (method 2), and with IMU and with re-estimated and smoothed GMM parameters (method 3). According to Fig. 3 and when the IMU is used, the log-likelihood output has less outlying low values but when the GMM parameters are re-estimated, it becomes larger on average as shown in the third plot. With the re-estimation and smoothing the fit of the GMM to the data is significantly improved. Moreover, the audio results as well as objective measurements validate the utility of re-estimation of GMM parameters during movement. Tab. 1 shows the performance of the above three different methods in terms of instrumental measures (see also Section 4). According to this table, results are improved when the GMM parameters are re-estimated in each frame and are recursively smoothed using (10).

	method 1	method 2	method 3
PESQ	1.41	1.90	2.23
Δ-PESQ	-0.17	0.21	0.53
SIR	10.0	12.4	17.4
SDR	-0.66	-0.11	3.67

Table 1. Comparison of the performance of three different methods. Array rotations with an angular speed of $15^\circ/s$. Method 1: without IMU. Method 2: with IMU information and fixed GMM weight and variance parameters. Method 3: with IMU information and re-estimated and smoothed GMM parameters.

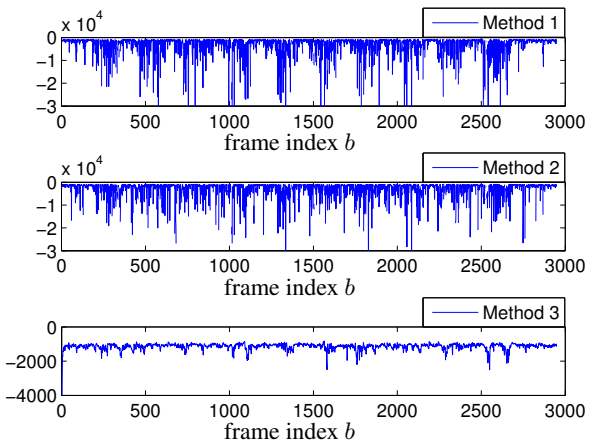


Fig. 3. Log-likelihood measure for three different methods. Methods 2 and 3 make use of the IMU. Method 3 includes GMM re-estimation and smoothing.

4. EXPERIMENTAL EVALUATION

Experiments were conducted in an acoustically treated room measuring 7.5 x 6 x 3 meters and a $T_{60} = 0.5s$. Genelec 2029BR speakers were placed at $\pm 30^\circ$ at a height of 1.2 m and a distance of 1.5 meters from a five microphone linear array. The microphone spacing was 3, 5, 7 and 10 cm. The array was mounted at a height of 1.2 m on a wooden pole in the centre of a Brüel and Kjær Type 3921 turntable. The audio interface was an RME Hammerfall DSP Multiface II. A Sparkfun 9-axis IMU (SEN-10736) was attached to the pole directly above the microphone array. The relative position of the microphone array was measured every 0.02 s using the 9-axis IMU with open-source firmware [18], however, only the most recent values recorded by the 9-axis IMU was saved with each (1536 sample) audio buffer. Recordings were made at 48 kHz and later downsampled to 8 kHz. By selecting the correct buffer size, synchronization of the audio and position data was more easily maintained after downsampling. Recordings were made using three different movement speeds: $15^\circ/s$, $30^\circ/s$ and $45^\circ/s$. Each cycle started with the array positioned at $+30^\circ$, perpendicular (broadside) to the female speaker and end up through 60° to -30°, perpendicular (broadside) to the male speaker as depicted in Fig. 1. Speech material was taken from the Telecommunications & Signal Processing Laboratory (TSP) 2 speech database [19]. Sentences were randomly concatenated for each recording. The total recording time was approximately 4.5 minutes.

The performance of algorithm has been evaluated in terms of the perceptual evaluation of speech quality (PESQ) [20], the PESQ improvement (Δ-PESQ) with the center microphone as the reference signal, as well as signal to interference ratio (SIR) and signal to distortion ratio (SDR) taken from BSS EVAL toolbox [21]. The experiments were conducted on three mixing conditions, i.e no additive background noise, white noise added at microphones with 0 and 10 dB SNR.

Tab. 2 shows the results obtained by new algorithm and the baseline method [9]. According to this table, the proposed method produces an improvement as compared to the baseline method [9] for all SNRs and over all angular speeds. The results show less improvement for a faster angular speed ($45^\circ/s$), despite acceptable audio results. Informal listening test reveal indeed a significantly improved audio quality: While the baseline system leads to inconsistent and distorted outputs the proposed approach is able to lock onto the target source and eliminate the effects of the array rotation.

angular speed $15^\circ/s$

	baseline method [9]			proposed method		
SNR [dB]	0	10	(∞)	0	10	(∞)
PESQ	1.14	1.30	1.41	1.53	2.03	2.23
Δ-PESQ	-0.19	-0.26	-0.17	0.16	0.41	0.53
SIR	10.1	11.50	10.00	15.8	17.9	17.4
SDR	-0.4	-0.51	-0.66	2.25	2.81	3.67

angular speed $30^\circ/s$

	baseline method [9]			proposed method		
SNR [dB]	0	10	(∞)	0	10	(∞)
PESQ	1.14	1.23	1.46	1.56	1.97	2.21
Δ-PESQ	-0.25	-0.19	-0.20	0.25	0.43	0.56
SIR	10.22	10.81	11.30	15.1	16.2	17.32
SDR	-0.29	-0.44	-0.04	2.90	3.32	3.60

angular speed $45^\circ/s$

	baseline method [9]			proposed method		
SNR [dB]	0	10	(∞)	0	10	(∞)
PESQ	1.12	1.22	1.35	1.48	1.92	2.17
Δ-PESQ	-0.35	-0.25	-0.20	0.17	0.41	0.53
SIR	7.95	8.01	8.54	10.9	11.1	12.8
SDR	-2.25	-2.51	-2.10	1.51	1.68	3.11

Table 2. Comparison of the performance of proposed method (with IMU and GMM re-estimation) and the baseline method [9] for angular speed $15^\circ/s$, $30^\circ/s$ and $45^\circ/s$ in top, middle and bottom respectively.

5. CONCLUSION

In this contribution we presented a novel multi-channel algorithm for the separation of concurrent speakers which is suitable for a moving microphone array. For instance, a head-mounted microphone array would be subject to turns of the listeners head during conversations and would require an adaptation of the relative positions of the sources. With this motivation, we investigated effect of rotational movements of the array on the quality of the output signal. We utilize an additional sensor to measure these rotations and use the output of this inertial measurement unit (IMU) to adapt the estimated direction of arrivals of the sources to the actual array position. The localization information is captured in a Gaussian mixture model (GMM) which is then used to compute posterior probabilities of source activity. These probabilities then control a fully adaptive generalized sidelobe canceler. Besides the mean directions of arrival it turns out that all parameters of the GMM should be re-estimated in each signal frame and should be smoothed to achieve a good audio quality. Results averaged over different angular speeds show improvements of 5.9 dB SIR, 4.39 dB SDR and 0.8 PESQ with respect to the baseline method when no ambient noise is added. Thus, using a localization algorithm followed by a statistical model whose parameters are re-estimated based on information delivered by an IMU and a subsequent smoothing process helps to improve the robustness of the localization and the quality of the audio signals.

6. REFERENCES

[1] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Equivalence between Blind Source Separation and Adaptive Beamformers," 2001.

[2] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The Fundamental Limitation of Frequency Domain Blind Source Separation for Convolutive Mixtures of Speech," vol. 11, no. 2, pp. 109–116, 2003.

[3] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A Versatile Framework for Multichannel Blind Signal Processing," 2004.

[4] J. Bourgois and W. Minker, *Time-Domain Beamforming and Blind Source Separation*, Lecture Notes in Electrical Engineering. Springer-Verlag, 2009.

[5] L.J Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *Antennas and Propagation, IEEE Transactions on*, vol. 30, no. 1, pp. 27–34, Jan 1982.

[6] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix Using Constrained Adaptive Filters," vol. 47, no. 10, pp. 2677–2683, Oct. 1999.

[7] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *Signal Processing, IEEE Transactions on*, vol. 49, no. 8, pp. 1614–1626, Aug 2001.

[8] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1071–1086, Aug 2009.

[9] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 1900–1912, Sept 2011.

[10] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P.N. Garner, and Weifeng Li, "Beamforming with a maximum negentropy criterion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 5, pp. 994–1008, July 2009.

[11] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wolfel, "Adaptive beamforming with a minimum mutual information criterion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2527–2541, Nov 2007.

[12] D.B. Ward, E.A. Lehmann, and R.C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 826–836, Nov 2003.

[13] N. Madhu and R. Martin, "A Scalable Framework for Multiple Speaker Localization and Tracking," in *in Proc. Intl. Workshop for Acoustic Echo Cancellation and Noise Control (IWAENC)*, 2008.

[14] Y. Lacouture-Parodi and E.A.P. Habets, "Application of particle filtering to an interaural time difference based head tracker for crosstalk cancellation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE*, May 2013, pp. 291–295.

[15] J.H DiBiase, H.F Silverman, and M.S Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180. Springer, 2001.

[16] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[17] Christopher M Bishop et al., *Pattern recognition and machine learning*, vol. 1, springer New York, 2006.

[18] P. Bartz, "Razor attitude and head rotation sensor," 2012, "https://github.com/ptrbrtz/razor-9dof-ahrs", accessed on 19.09.2014.

[19] Peter Kabal, "TSP speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

[20] A.W Rix, J.G Beerends, M.P Hollier, and A.P Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, (ICASSP), 2001 IEEE International Conference on. IEEE*, 2001, vol. 2, pp. 749–752.

[21] Cedric Févotte, Rémi Gribonval, Emmanuel Vincent, et al., "Bss_eval toolbox user guide–revision 2.0," 2005.