published in: *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM), July 2016, Rio de Janeiro, Brazil*

Multi-microphone Speech Enhancement Informed by Auditory Scene Analysis

Axel Plinge

Department of Computer Science
TU Dortmund University
Dortmund, Germany

Sharon Gannot

Department of Engineering
Bar Ilan University
Ramat Gan, Israel

ABSTRACT

A multitude of multi-microphone speech enhancement methods is available. In this paper, we focus our attention to the well-known minimum variance distortionless response (MVDR) beamformer, due to its ability to preserve distortionless response towards the desired speaker while minimizing the output noise power. We explore two alternatives for constructing the steering vectors towards the desired speech source. One is only using the direct path of the speech propagation in the form of delay-only filters, while the other is using the entire room impulse response (RIR). All beamforming methods requires some control information to be able to accomplish the task of enhancing a desired speech signal. In this paper, an acoustic event detection method using biologically-inspired features is employed. It can interpret the auditory scene by detecting the presence of different auditory objects. This is employed to control the estimation procedures used by beamformer. The resulting system provides a blind method of speech enhancement that can improve intelligibility independently of any additional information. Experiments with real recordings show the practical applicability of the method. Significant gain in fwSNRseg is achieved. Compared to using the direct path only, the use of the entire RIR proves beneficial.

***Index Terms*—** microphone array, auditory scene analysis, blind beamformer for speech enhancement

1. INTRODUCTION

For speech enhancement using multiple microphones, a multitude of beamforming methods exist [1]. They employ a variety of optimization criteria. One basic, but yet robust, type is the delay-and-sum beamformer, that uses only time delays to steer the spatial filter to the source direction [2], [3]. Better enhancement can be gained by the so-called data-dependent beamformers that apply some constrained minimization criterion. One such spatial filter is the minimum variance distortionless response (MVDR) beamformer that steers a "beam" towards the desired source while minimizing sounds from all other directions [4]. This can be split in two parallel processing paths in the well-established generalized sidelobe canceler (GSC) implementation [5]: A fixed beamformer (FBF) focusing on the source and a blocking matrix (BM) that blocks it and provide noise reference signals to the subsequent adaptive noise canceler (ANC), c.f Fig. 1. Similarly, the linearly constrained minimum variance (LCMV) criterion minimizes the noise power while satisfying a set of linear constrains on the responses of multiple sources of interest. For both MVDR and LCMV, the fixed response for the desired source is often a delay-only steering vector, thus only using the direct path. In reverberant enclosures, the room impulse response (RIR) of the source consists of many reflections. Hence, methods using an estimate of the entire RIR provide better speech quality [6], [7]. Another family of beamformers is based on blind source separation (BSS) concepts that aims at imposing independence between the sources [8], [9].

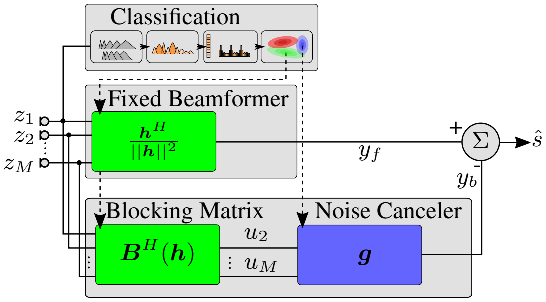


Fig. 1. Proposed method: A classifier identifies speech or suitable noise segments that respectively update a FBF and BM or ANC.

All the above mentioned methods benefit from and sometimes require a control mechanism that informs them when to estimate the filters and allows them to distinguish between sources. Speech activity can be used to adapt the different paths of the GSC [6]. The direction of the desired source can be utilized in delay-and-sum, MVDR, and LCMV beamformers [3], [10], [11]. Directional information can also be incorporated into BSS beamformers [12], [13].

Recently, signal processing based on physiological insights has become increasingly popular [14]. One of the most influential theories is the auditory scene analysis (ASA) that describes how the human listeners is able to segregate speech by interpreting the acoustic scene as composed of auditory objects [15]. A common concept in computational ASA (CASA) is the estimation of a time-frequency mask for each speaker [16], which can be applied to beamforming in the form of a postfilter [17]. Using CASA frontends has also been shown to be beneficial in the higher level tasks of speaker localization and identification [18]–[20]. A higher-level task is recognition of types of auditory objects [21]. Classification based approaches have shown to be more robust than simpler measures for the detection of speech in noise [22]. The use of CASA-inspired features has shown potential in [23].

Figure 1 shows the structure of the system proposed in this paper. The single channel bag of features (BoF) classifier introduced in [23] is used to classify time segments. This is used to estimate the components of an MVDR beamformer implemented in a GSC structure [6]. From speech segments, the FBF and BM blocks are updated. Stationary noise segments are used to adapt the ANC. When nonstationary noise is detected, neither are updated.

This work was supported by a fellowship within the FITweltweit programme of the German Academic Exchange Service (DAAD).

2. METHOD

The system consist of an MVDR beamformer and a control unit in the form of a classifier. A dedicated training procedure allows the latter to blindly distinguish between speech, noise and nonstationary noise. The estimation of the beamformer components is guided by the detection of these categories.

2.1. Beamformer

In the short time Fourier transform (STFT) domain the signal received by the M microphones can be written in vector form as $\boldsymbol{z}(t,k) = [z_1(t,k), \ldots, z_M(t,k)]^T$, with t denoting the frame and k the frequency indexes. This way we can write the signal model as the sum of the source signal s filtered by the time varying acoustic transfer function (ATF) $\boldsymbol{a}(t,k) = [a_1(t,k), \ldots, a_M(t,k)]^T$ plus noise signals $\boldsymbol{v}_i(t,k)$ with their respective ATF $\boldsymbol{b}_i(t,k)$:

$$\boldsymbol{z}(t,k) \approx \boldsymbol{a}(t,k)\, s(t,k) + \sum_i \boldsymbol{b}_i(t,k)\, \boldsymbol{v}_i(t,k)\,. \qquad (1)$$

The method introduced in [6] estimates the relative transfer function (RTF) by a least squares (LS) fit utilizing speech nonstationarity. The overall GSC estimation procedure is summarized in algorithm 1. The operator $\overset{\mathrm{FIR}}{\leftarrow}$ stands for the operation of constraining the support of the filter in time-domain. f_s denotes the sampling frequency.

Algorithm 1 GSC estimation

```
1a. using all speech time segments, estimate RTF
        $\boldsymbol{h}(t,k) = \frac{\boldsymbol{a}(t,k)}{a_1(t,k)}$        or
1b. estimate DoA based TF
        $h_m(t,k) = \exp(-\imath 2\pi k/K(\tau_m(\theta) - \tau_1(\theta))f_s)$
2. compute FBF
        $y_F(t,k) = \frac{\boldsymbol{h}^H(t,k)\,\boldsymbol{z}(t,k)}{||\boldsymbol{h}||^2}$
3. compute noise reference
        $\boldsymbol{u}(t,k) = \boldsymbol{B}^H(\boldsymbol{h})(t,k)\,\boldsymbol{z}(t,k)$ with
        $\boldsymbol{B}(\boldsymbol{h}) := \begin{bmatrix} -\boldsymbol{h}^H_{2:M} \\ \boldsymbol{I}_{M-1\times M-1} \end{bmatrix}$
4. compute output
        $\hat{s}(t,k) = y_F(t,k) - \boldsymbol{g}^H(t,k)\,\boldsymbol{u}(t,k)$
5. update ANC for each noise frame
        $\tilde{\boldsymbol{g}}(t+1,k) = \boldsymbol{g}(t,k) + \mu\frac{\boldsymbol{u}(t,k)\,\hat{s}^*(t,k)}{p(t,k)}$
        $\boldsymbol{g}(t+1,k) \overset{\mathrm{FIR}}{\leftarrow} \tilde{\boldsymbol{g}}(t+1,k)$ with
            $p(t,k) = \lambda p(t-1,k) + (1-\lambda)||\boldsymbol{z}(t,k)||^2$
```

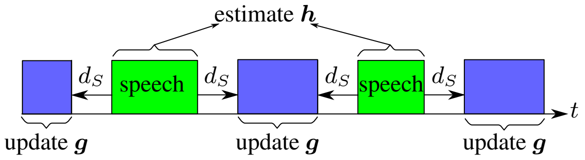


Fig. 2. Updating strategy from classifier output. The frames classified as speech are used together to compute the fixed beamformer coefficients $\boldsymbol{h}$. The frames classified as noise farther away than a guard margin d_S from speech are used to continuously update the ANC coefficients $\boldsymbol{g}$.

2.2. Control

The control information is provided by the soft supervised BoF acoustic event classification method introduced in [23]. From the single channel signal at the first microphone, both mel frequency cepstral coefficients (MFCCs) and gammatone frequency cepstral coefficients (GFCCs) features and their first derivatives are calculated. The input is classified by the BoF system in a sliding window of 1 s. The output is reduced to the decision ‘speech’, ‘noise’, or ‘non-stationary noise’. The input signal is divided into consecutive time segments of these three types.

2.3. Training

In order to deal with different noise types and speech mixed with noise, a dedicated training strategy was devised: First, for different types of noises, several examples are recorded with the device. It is distinguished between four types of noise, ordered by increasing nonstationarity: 1) very stationary noises such as white noise or fan sounds, 2) mechanical noises, 3) speech-like babble noise, and 4) nonstationary noise like keyboard typing. In order to estimate a good representation for speech, the speech samples are mixed with samples of each of the different noise types at a high SNR of 18 dB to train the speech class Ω_0. For each of the four noise classes, a class $\Omega_1 \ldots \Omega_4$ is trained individually using its examples. Additionally, for each of them a mixture class $\Omega'_1 \ldots \Omega'_4$ is trained by mixing noise types of the same level of stationarity or lower, e.g. Ω'_2 is trained by mixing with different noise types from the categories Ω_1 and Ω_2.

2.4. Estimation

The time segments classified as speech are used to estimate the FBF and BM. The noise segments are used to update the ANC. In the transitions between speech and noise, especially at low SNRs, an underestimation of speech existence might occur. The updating of the ANC in speech would lead to a serious deterioration of the performance due to speech distortion. Therefore, a guard boundary of $d_S = 0.5\,\mathrm{s}$ around the time segments classified as speech is introduced. The ANC is only updated in step 5 in noise segments that are d_S before or after the speech segments as shown in Fig. 2.

There are two versions of the FBF and BM. They employ different methods of estimating $\boldsymbol{h}$ in the first step of algorithm 1. The first method is estimating the full RTFs. The second method only uses the direct path. h_m is set to a pure phase corresponding to the time difference of arrivals (TDoAs) for each microphone given the direction of arrival (DoA) of the speaker. Multiple methods for estimating the DoA of the sound source towards the microphones exist [24]. In line with the ASA approach, a robust neuro-biologically

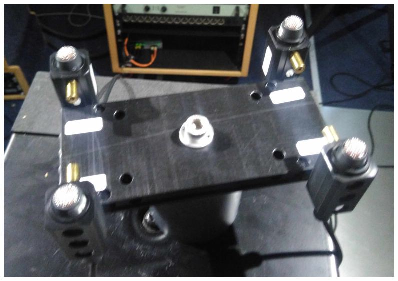


Fig. 3. Smartphone mockup comprised of four microphone mounts attachable to a plastic body.

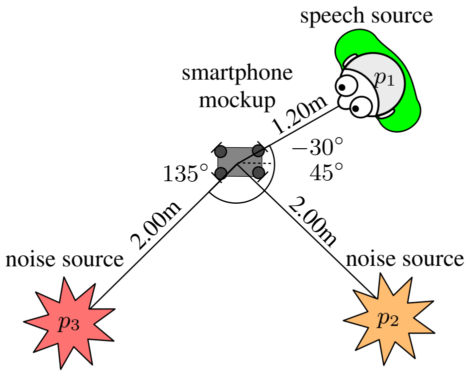


Fig. 4. Recording scenario: Three speakers are placed around the mockup at different angles and distances.

inspired method [19] was employed. The mean of DoA present in time segments classified as speech, but not in time segments classified as noise, was used.

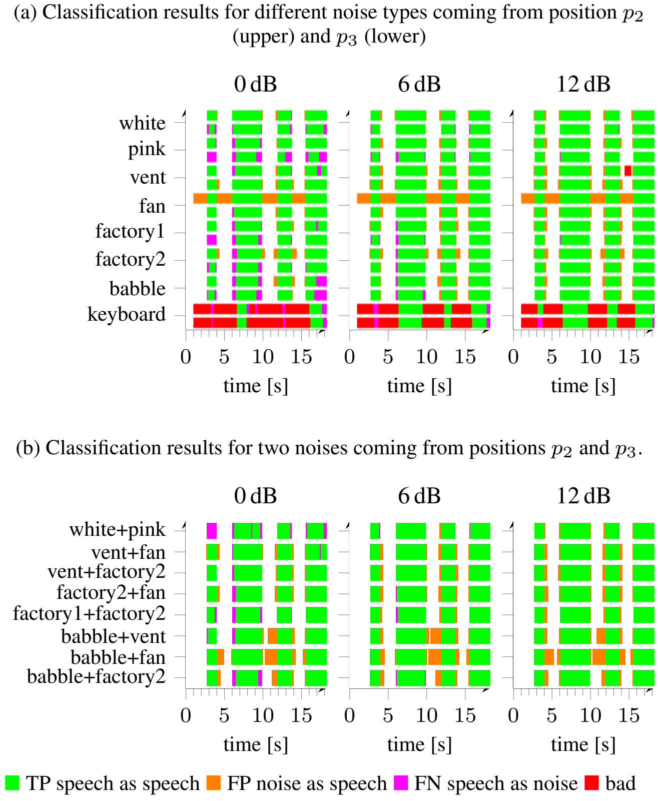
3. EVALUATION

Recordings were done in the acoustic lab at Bar Ilan University [25]. The room's T_{60} was adjusted to 320ms. A smartphone mockup was used for recordings. The mockup consists of a plastic body and four microphones mounted near the edges in an 12x8 cm rectangular pattern, cf. Fig 3.

Three speakers were placed around the mockup as illustrated in Fig. 4. The desired speaker was placed at 1.2 m distance at $-30°$ in position p_1. Two interfering speakers were placed at 2.0 m distance at 45° and 135°, at position p_2 and p_3 respectively. The recordings were executed with 48 kHz sampling rate and 24 bit resolution. This was used for the classifier. The beamformer was applied to the signals downsampled to 16 kHz.

The noise samples 'pink', 'babble', 'factory1' and 'factory2' were taken from the NOISEX-92 database [26]. From the freesound database [27], the following were extracted: 'roaring fan' – a rather loud humming fan, 'ventilation' – air conditioning noise, 'keyboard' – constant keyboard typing, The 'white', 'pink', 'roaring fan', and 'ventilation' noises were used for Ω_1, 'factory1' and 'factory2' for Ω_2,'babble' for Ω_3 and 'keyboard' for Ω_4.

For testing, two different anechoic speech sequences from the same speaker were played. In each sequence, there are four speech segments of 2-4 s. Overall, they were 18.5 s and 16.5 s long, where speech is present half of the total time. Each noise was added to each


(a) Classification results for different noise types coming from position p_2 (upper) and p_3 (lower)

(b) Classification results for two noises coming from positions p_2 and p_3.

Fig. 5. Classification results for the first speech sequence.

speech sequence at SNRs of 0,6,12 dB played individually form each of the noise speaker positions. Additionally, sequences with two different noises, each of them played from one of the two speakers, were used.

The classifier was trained with data from a different recording session using the same mockup placed at a different angle and playing speech or noise from p_1, p_2. Recordings of a 45 s long anechoic speech sequence and the different noise samples played for 120 s were used.

3.1. Classification

Figure 5 shows the classifications for the first speech sequence. Different colors are used to show correct speech detections, speech classified as noise, noise classified as speech, and the detection of nonstationary noise. The speech related performance is expressed as true positive (TP), false positive (FP), and false negative (FN) respectively.

At 0 dB sometimes speech segments are estimated too short or missed, especially with 'pink' and 'factory' noise. The 'fan' noise from position 1 is the only case where noise is classified as speech. Over all sequences, excluding the keyboard, there are 92.8% TP, 14.2% FP, and 7.2% FN relative to the number of speech frames. The keyboard is detected, but it also deteriorates the speech estimations as a lot of speech frames are also classified as keyboard.

For the mixtures of noises from both positions, the results are still good with 95.7% TP, 18.6% FP, and 4.3% FN. In the 0 dB SNR condition, speech is missed again in some cases. In mixtures with babble noise, there is some miss-classification of noise as speech.

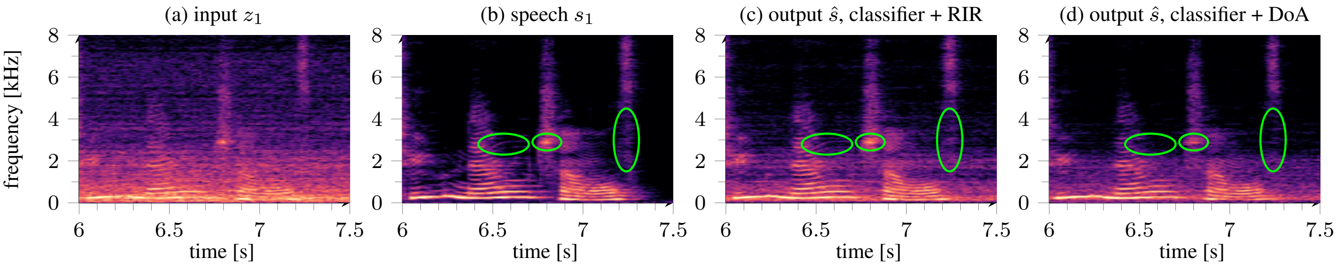


Fig. 6. Spectrograms of speech 'two narrow channels' in speech sequence 2, distorted by 'factory' noise p_2. Input signal (a), speech part (b), output of the proposed method (c), and the DoA variant (d). In (d) some speech parts present in (c) are missing or muffled (green ellipses), both reduce noise similarly.

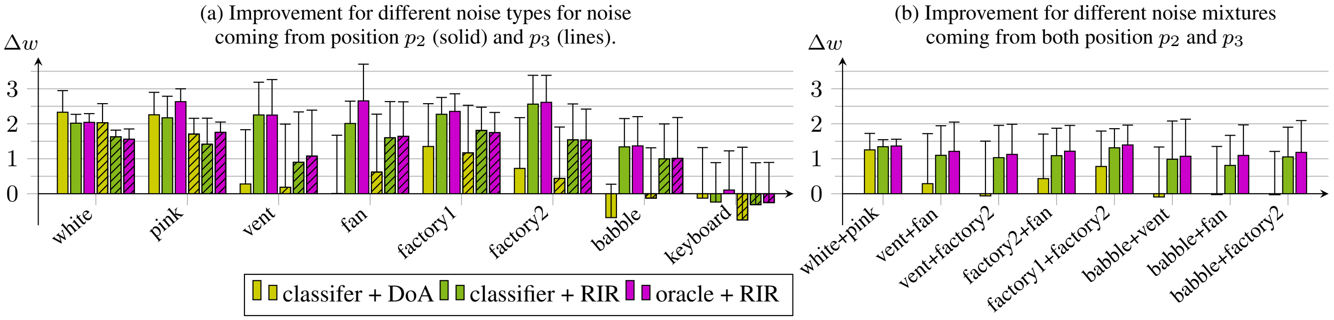


Fig. 7. Improvement for different noise scenarios, mean and standard deviation computed over the three SNRs (0,6,12) and both speech sequences for a fixed DoA, the full RIR estimation using the classifier and the RIR with the oracle annotations of speech.

3.2. System performance

In order to asses the ability to blindly perform speech enhancement by classification of auditory objects guiding the beamforming, the combined system was applied to the test recordings. The proposed method clearly suppresses the noise in all cases while very little distortion is introduced to the speech signal. Figure 6 shows an example application of the method. It can be seen that the noise is greatly reduced. In case of DoA processing, more speech components, especially at high frequencies, are suppressed. From listening to the audio samples[1] the speech sounds a bit muffled.

The frequency weighted segmental SNR (fwSNRseg) [28] was found to be related to subjective listening quality [29]. Therefore, the relative improvement is computed as the difference in the fwSNRseg w between the speech signal as received by the first microphone to the processed output $w(s_1, \hat{s})$ and the fwSNRseg of the speech to the mixed input signal $w(s_1, z_1)$.

$$\Delta w = w(s_1, \hat{s}) - w(s_1, z_1) \tag{2}$$

For a single noise source, the mean Δw over all noise types, excluding the keyboard, is 1.75 ± 0.89 dB. When using an oracle in the form of ground truth annotations instead of the classifier, it is only slightly better with 1.87 ± 0.92 dB. The DoA-based processing achieved a mean improvement of 0.88 ± 1.49 dB This can be explained by the fact that the blocking is less effective and $\boldsymbol{u}$ contains more speech residuals. Although there is some muffling of the speech signal, the figures for the DoA-based steering vector are slightly better for the artificial 'white' and 'pink' noise. In all other cases, the proposed method achieves higher improvement, cf. Fig. 7a. For the babble noise, the fwSNRseg decreases when using the DoA, while the proposed method provides a clear improvement. In the case of 'keyboard' noise, none of the proposed method is able to consistently improve the fwSNRseg.

When two different noises are coming form different direction, the task is more difficult as the ANC has to cancel them both. The improvement in fwSNRseg is 1.09 ± 0.76 dB compared to 1.21 ± 0.74 dB with the oracle. The DoA version performs consistently worse with 0.32 ± 1.26 dB, cf. Fig. 7b.

4. CONCLUSIONS

A fully blind system for speech enhancement with multiple microphones in stationary noise was proposed. There is a solid improvement achieved by the proposed method for a single noise source and for two noise signals from different directions even in 0 dB SNR. It is beneficial to use the full RIR in the construction of the beamformer instead of only using the direct path, more so in cases of real noise samples.

The classifier performs very well in most cases. Speech is detected with around 95% TP, 15% FP and 5% FN, better performance is achieved in higher SNRs. While the training strategy seems to generalize well, the robustness to other noise types and room configurations should be investigated further. The performance could be enhanced using multiple microphone information [30], [31]. The system performance comes very close to the performance using the ground truth instead of the classifier. This shows that the classifier is integrated in a practically applicable way.

In the case of non-stationary noise, there is little improvement by the proposed method. Since the classifier detects this situation, the ANC adaptation can be switched off.

[1] available at
`www.eng.biu.ac.il/gannot/speech-enhancement/sam16`

5. REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, Berlin, Heidelberg, 2005.

[2] S. Doclo and M. Moonen, "Design of far-field and near-field broadband beamformers using eigenfilters," *Signal Processing*, vol. 83, no. 12, pp. 2641 – 2673, 2003.

[3] P. Pertila and A. Tinakari, "Time-of-arrival estimation for blind beamforming," in *Int. Conf. on Digital Signal Process.*, Fira, Santorini, Greece, July 2013.

[4] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

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

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

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

[8] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, Montreal, Canda, 2004, vol. 3, p. iii–889.

[9] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, Berlin, Heidelberg, 2007.

[10] M. Taseska and E. A. P. Habets, "Spotforming using distributed microphone arrays.," in *IEEE Workshop on Appl. Signal Proces. Audio and Acoustics*, New Paltz, NY, USA, 2013.

[11] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 22, no. 12, pp. 2182–2196, Dec. 2014.

[12] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann, "Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," in *IEEE Workshop on Appl. Signal Proces. Audio and Acoustics*, New Paltz, NY, USA, Oct 2013.

[13] L. Parra and C. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech and Audio Process.*, vol. 10, no. 6, pp. 352–362, sep 2002.

[14] R. F. Lyon, "Machine Hearing – An Emerging Field," *IEEE Signal Process. Magazine*, Sept. 2010.

[15] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[16] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley, 2006.

[17] S. Araki and T. Nakatani, "Hybrid approach for multichannel source separation combining time-frequency mask with multichannel Wiener filter," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, Prague, Czech Republic, May 2011, pp. 225–228.

[18] X. Zhao, S. Member, Y. Shao, and D. Wang, "CASA-Based Robust Speaker Identification," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 20, no. 5, pp. 1608–1616, 2012.

[19] A. Plinge and G. A. Fink, "Online multi-speaker tracking using multiple microphone arrays informed by auditory scene analysis," in *European Signal Process. Conf.*, Marrakesh, Morocco, Sept. 2013.

[20] T. May, S. van de Par, and A. Kohlrausch, "Binaural localization and detection of speakers in complex acoustic scenes," in *The Technology of Binaural Listening*, J. Blauert, Ed., pp. 397–425. Springer, Berlin, Heidelberg, 2013.

[21] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[22] J.-H. Bach, B. Kollmeier, and J. Anemuller, "Modulation-based detection of speech in real background noise: Generalization to novel background classes," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, Dalla, Texas, USA, 2010, pp. 41–44, IEEE.

[23] A. Plinge, R. Grzeszick, and G. A. Fink, "A Bag-of-Features approach to acoustic event detection," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, Florence, Italy, May 2014.

[24] R. Martin, U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission*, Wiley, 1 edition, 2008.

[25] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Int. Works. on Acoustic Signal Enh.*, Juan les Pins, France, Sept. 2014, pp. 313–317.

[26] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[27] "The freesound database," www.freesound.org.

[28] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, jan 2008.

[29] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, Jan. 2016.

[30] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos, "Multi-microphone fusion for detection of speech and acoustic events in smart spaces," in *European Signal Process. Conf.*, Lisbon, Portugal, Sept. 2014, pp. 2375–2379.

[31] H. Phan, M. Maass, L. Hertel, R. Mazur, and A. Mertins, "A multi-channel fusion framework for audio event detection," in *IEEE Workshop on Appl. Signal Proces. Audio and Acoustics*, New Paltz, NY, USA, 2015.