Interspeech 2018
2-6 September 2018, Hyderabad

Speaker Activity Detection and Minimum Variance Beamforming for Source Separation

Enea Ceolini, Jithendar Anumula, Adrian Huber, Ilya Kiselev, and Shih-Chii Liu

Institute of Neuroinformatics, University of Zurich and ETH Zurich
Zurich, Switzerland

enea.ceolini@ini.uzh.ch, anumula@ini.uzh.ch, huberad@ini.ethz.ch, kiselev@ini.uzh.ch, shih@ini.ethz.ch

Abstract

This work proposes a framework that renders minimum variance beamforming blind allowing for source separation in real world environments with an ad-hoc multi-microphone setup using no assumptions other than knowing the number of speakers. The framework allows for multiple active speakers at the same time and estimates the activity of every single speaker at flexible time resolution. These estimated speaker activities are subsequently used for the calibration of the beamforming algorithm. This framework is tested with three different speaker activity detection (SAD) methods, two of which use classical algorithms and one that is event-driven. Our methods, when tested in real world reverberant scenarios, can achieve very high signal-to-interference ratio (SIR) of around $20\,\mathrm{dB}$ and sound quality of 0.85 in short-time objective intelligibility (STOI) close to optimal beamforming results of $22\,\mathrm{dB}$ SIR and 0.89 in STOI.

Index Terms: speech separation, sound localization, beamforming, event-driven computation.

1. Introduction

Multi-channel beamforming algorithms are useful for applications such as video conferencing, speech recognition and human computer interfaces [1, 2, 3]. In a standard formulation, they require knowledge of transfer functions between each source and each microphone. Beamforming algorithms including minimum variance distortionless response (MVDR), the more generalized linearly-constrained minimum variance (LCMV) [4] and multi-channel Wiener filter (MWF) [5], can be formulated in such a way that they do not need any information about the geometry of the array, but only the knowledge of the acoustic transfer functions (ATF). Methods exist [6, 7] that can be used to estimate the ATFs directly from the input but they require knowledge about speaker activity.

Some efforts have been made towards rendering beamforming a blind solution by estimating speaker activity. A method has been proposed [8] based on a voice activity detection (VAD) scheme that performs speaker identification from direction of arrival (DOA) clustering. This information is then used by a maximum SNR beamformer. The speakers are assumed to be intermittent which is not the case in most realistic scenarios where source separation has to be applied. Another method proposed recently uses generalized cross-correlation with phase transform (GCC-PHAT) and time difference of arrival (TDOA) together with a clustering algorithm to estimate steering vectors for the speakers, allowing for estimation of speaker activity patterns [9]. However, they did not apply their solution to beamforming for source separation. None of the aforementioned work demonstrates a complete framework for blind beamforming with an ad-hoc (i.e. does not rely on a particular geometry) multi-microphone setup.

This work proposes such a framework with the only assumption of knowing the number of speakers. First, speaker activity over short time frames of 20ms is estimated using different algorithms such as GCC-PHAT [10], multiplicative non-negative independent components analysis (M-NICA) [11] and spike separation (SPS) [12]. From the results of the SAD, time frames assigned to one speaker are pooled together and the ATFs are then estimated using a noise-covariance whitening based ATF estimation method [6]. In particular it is possible to show that the beamforming calibration works even when the speakers' activity highly overlaps. Results are presented for speaker separation on real-world recordings in a reverberant environment with an ad-hoc multi-microphone setup. The quality of the separation is assessed using both BSSEval metrics [13] and perceptual quality measures such as perceptual evaluation of audio source separation (PEASS) [14] and STOI [15]. Ultimately the work demonstrates the potential of using speaker activity (SA) in short time frames to calibrate a beamforming algorithm.

The paper is organized as follows. We shortly describe LCMV in Section 2. We introduce the different methods for the extraction of SAD in Section 3 and present the results in Section 4.3. We shortly conclude in Section 5.

2. Minimum variance beamforming

The problem of minimum variance beamforming is to formulate a beamformer which can minimize the output noise power while leaving non-distorted the desired speech signals. This problem can be seen as a constrained convex optimization problem which has a well-known optimal closed form solution [4].

LCMV has been shown to be a robust beamformer for reverberant environments [16, 17]. LCMV is preferred to MWF since it preserves more the signal quality, avoiding the introduction of distortion, by sacrificing signal-to-noise ratio (SNR). The main assumption in LCMV is to know the steering vector of the beamformer. Estimating this vector is non-trivial in an ad-hoc microphone array of unknown geometry. This work relies on the technique of signal subspace introduced in [7] with which it is possible to estimate the impulse response of a source to the microphones (and thus the steering vector) by extracting the eigenvector corresponding to the maximum eigenvalue of the covariance matrix estimated during period of activity of this desired source. This method is also known as noise-covariance whitening based ATF estimation [6]. The speaker activity is thus the only information that a system implementing LCMV needs, so that it can be considered blind. In the reminder of the paper, we will address methods of extracting this information in an unsupervised way, making the system blind.

10.21437/Interspeech.2018-1606

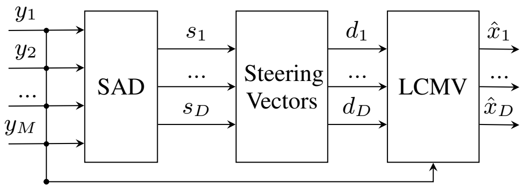

Figure 1: *Flow of the framework. The microphone signals are used to estimate the speakers' activities. Once estimated, the activity information can be used to estimate the steering vectors to extract each speaker with LCMV.*

3. Speaker activity detection for beamforming calibration

The framework used in this work is shown in Figure 1. The short-time Fourier transforms (STFT) of the microphone signals $\mathbf{y}(f,t)$ are considered where f and t denote frequency and time frame respectively. Assuming D speakers, we look for a partitioning where every frame t is assigned to a class $\{c_1, c_2, ..., c_D\} \bigcup \{u\}$ where the first set corresponds to the set of speakers and the second corresponds to no active speaker. Ideally $\mathbf{y}(f,t) \in c_i$ means that speaker i is active alone in frame t. Once *enough* frames of one speaker have been collected (the meaning of *enough* will be addressed later), they are pooled together and used for estimating the ATF. Some heuristics are used to ensure that *enough* frames are collected for a precise calibration. First, only frames in consecutive sets of at least 200 ms (see Figure 3) are collected, shorter collections of frames are less likely to contain useful information about the location. Also it is important to ensure that the total collection of frames can provide a full-rank covariance matrix which is well-conditioned and can be successfully used for ATF estimation. Once the ATFs are estimated, the optimal LCMV solution can be calculated and applied to the already available STFT of the microphone array. Later in this section three methods will be shown that allow us to estimate SA for each speaker at the same time even when they highly overlap. Once the activity pattern $s_1(t), ..., s_D(t)$ of each speaker has been estimated, a single speaker or no speaker is assigned to each frame t using the following procedure

$$\mathbf{y}(f,t) \in c_i \Rightarrow s_i(t)T_{th} > s_k(t) \quad \forall k \in \mathcal{D}^{-i} \tag{1}$$

where $\mathcal{D}^{-i} = \{1, ..., D\} \setminus \{i\}$ and $T_{th} \in [0,1]$ is a threshold that defines how much stronger one source has to be w.r.t. the others in one frame to be assigned to that frame.

After this step, a binary decision is taken for each frame and for each speaker that defines if the frame belongs or not to that speaker as shown in Figure 3. Three methods are considered for the unsupervised estimation of the speaker activity. The first is the classical generalized cross-correlation (GCC) which uses information about location. The second is M-NICA introduced for blind source separation (BSS) of source envelopes. The third is a recently proposed localization method that uses the timing of asynchronous samples from an event-driven binaural audio sensor [12]. We considered this third method because of its possible low computational complexity.

3.1. GCC

GCC can be used to estimate multiple source locations with arbitrary resolution. In particular GCC-PHAT is the most robust variant against reverberation. Similarly to a recently proposed method [18], the GCC-PHAT result calculated at each time frame, consists of a vector of locations corresponding to the number of bins used for the fast Fourier transform (FFT). The correlation output is then averaged over time to determine which bins have the maximum correlation, thus solving the localization problem. Once the bins corresponding to the positions of the sources have been selected, their time evolution can be considered and used as SA. Finally the time evolutions of each selected location can be compared as shown in (1), thus giving the SAD output shown in Figure 3.

3.2. M-NICA

The M-NICA method [11] is a BSS approach developed to separate a mixture of multiple non-negative sources which in this case correspond to the envelopes of the mixed sources. Once the envelopes for each speaker have been estimated, (1) is used to obtain a SA for each source. The assumptions for M-NICA are not exactly fulfilled in the scenarios described in our experiments because the method was developed for linear and instantaneous mixtures, but we expect the recovered envelopes to be sufficient for a SAD.

3.3. SPS: an event-based SAD

A SAD algorithm based on an data-driven form of encoding information is also considered. This data encoding format is used in the asynchronous outputs of event-driven audio sensors [19] and can be promising for applications that benefit from low-power and low-latency system specifications.

While conventional sampling schemes use a fixed sampling rate, event-driven sampling schemes produce an output only when a fixed change is detected in the input signal [20]. Here, we focus on one particular scheme as described below.

3.3.1. Event-based data and audio sensor

In an event-driven audio sensor, the microphone input signal is filtered through a set of bandpass filters with different center frequencies. The output of a filter, $x(t)$, is first rectified, $x^+(t) = \max(x(t), 0)$ and then integrated, $x_{int}(t) = \int_0^t x^+(s)\,ds$. Samples are generated from this integrated signal using a send-on-delta sampling scheme [21] with a predefined delta threshold Δ. This event generation mechanism is an approximation of the sampling scheme used in the binaural silicon cochlea sensor [19], which roughly models the properties of the early stages of the biological cochlea [22, 23]. This sensor is used in the multi-speaker experiments described in a later section.

3.3.2. Event-based localization algorithm

It was recently shown how the outputs of this event-based cochlea sensor could be used to localize multiple active sources simultaneously [12]. Each output event of the sensor is assigned a probability of it being produced by a source at a particular location $l \in \mathcal{L} = \{1, ...L\}$ where L is the number of possible locations. Because the events are an indirect measure of the signal energy, we also use an 'event count feature' as a measure of the energy of each source in the mixture. The event count feature corresponds to a moving average of events collected in a defined time window. We use this feature as an estimate of the speech envelope [12]. Once an envelope estimate is obtained for each location, the locations of the active sources have to be selected. Similarly as for the other two methods, knowing the number of speakers in the mixture is necessary. In this case, a score q_l can be computed for each location l based on the cumulative probability and the total number of events E_l assigned to location l:

$$q_l = N \sum_{e_i \in E_l} p(e_i = l) \tag{2}$$

where $N = |E_l|$ and $p(e_i = l)$ is the probability that event e_i is produced by the source in location l. The first active location l_1 is assigned as the location with the highest score q_l. The other active locations are iteratively assigned by looking at the envelope estimates of non-assigned locations which least correlated with those of the already assigned active locations. The pairwise correlations $\rho_l^{l_i}$ can be calculated between the assigned location $\mathcal{L}^i = \{l_1, ..., l_i\}$ and the residual locations $\mathcal{L}^{-i} = \mathcal{L} \backslash \mathcal{L}^i$. Then the scores $\hat{q}_l$ can be calculated as follows:

$$\hat{q}_l = \frac{1}{q_l \sum_{l_k \in \mathcal{L}^i} \rho_l^{l_k}} \tag{3}$$

and the selected location of the active speaker is given by $\arg\max_{l \in \mathcal{L}^{-i}} \hat{q}_l$.

We extend the work in [12] by using the estimated envelopes in a SAD system. By comparing the various envelopes using (1), we construct a SAD similar to the M-NICA method.

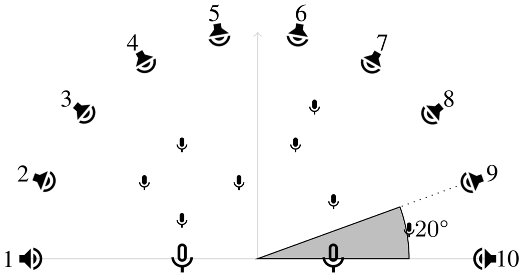

Figure 2: *Experimental Setup*

4. Experiments and Results

In this section, a full evaluation of the framework is presented. This evaluation is carried out on a self recorded dataset made in a typical office space.

4.1. Dataset

The setup used to record the dataset is shown in Figure 2. It was placed on a table within a room of 6 m × 5 m with a reverberation time of 0.4 s. The voices of two males were taken from online free audiobooks[1]. For each trial, the talkers were played from two of 10 possible locations around a half circle of radius 2 m. The minimum and maximum angular distances between two active loudspeakers in a trial are 20° and 180° respectively. The talkers in each trial are mixed with a random SNR$\in [-6, 6]$ dB. Each trial is 1 minute long comprising 15 s segments of only one active speaker and 30 s for the simultaneous active talkers. Recordings are done with two different audio platforms. The first is a wirelessly synchronized multi-microphone platform called WHISPER [24] which allows for recordings from up to 8 microphones arranged in an ad-hoc manner (🎙). The second audio platform [19] is described in Section 3.3.1, and its microphones (🎙) are placed in the center of the half circle of loud-speakers at a distance of 20cm. The recordings for the different sensor platforms were synchronized with both an acoustic and a digital trigger. All the recordings [2] and the code used for the analysis [3] are available online.

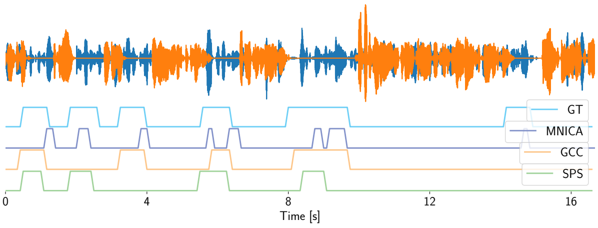

Figure 3: *Speaker activity detection of a particular speaker (blue) in the two speaker mixture waveform presented in the top panel. The bottom panel shows the ground truth (GT) SA and the estimated SA using GCC, M-NICA and SPS.*

4.2. Metrics for real world BSS evaluation

The quality of the separated sources is evaluated using both BSSEval metrics and speech intelligibility metrics. However, these metrics assume that the ground truth is available. In real world scenario recordings, we can only compare the quality of the estimated separated signals with the signals that are sent to the loudspeakers. Therefore the recorded sound from the microphones include the degradation due to the quality of the loudspeakers and the microphones. To address this, we define three measures:

- **BEST**: Obtained without mixing the sources therefore showing the quality loss of the signal simply by recording it with a microphone setup in the reverberant room.
- **ELIM**: Represents the empirical upper limit to the quality of the beamforming which is calibrated with the method [6], but having access to a segment of 15 s of each speaker alone, which is provided in every trial.
- $*$**-SAD**: Obtained by doing the calibration based on the information given by the estimated SA using one of the 3 methods ($*$), GCC, M-NICA or SPS.

4.3. Results

Table 1 shows the BSSEval, STOI and PEASS results of the three SAD methods computed over 18 trials. These results are obtained using the defined set of angular separations used in our experimental setup. All three SAD methods produced very similar SIR results close to the empirical limit set by LCMV. Figure 4 shows an example of mixed and enhanced speech. Suppression of the undesired source is clear, in particular in the panel *Enhanced source 2*. The signal-to-artifact ratio (SAR) and signal-to-distortion ratio (SDR) values are lower than the values obtained by state-of-the art algorithms [18, 25] due to the problems described in Section 4.2. However, these values are only around 2dB lower that the best possible numbers, proving that most of the degradation is already present in the microphone recordings. The STOI scores are high, especially from GCC-SAD which reaches the empirical limit of LCMV. The same is true of the PEASS measure which reflects the results obtained

[1] https://librivox.org/

[2] https://goo.gl/RZCe6H

[3] https://github.com/SensorsAudioINI/SpikeSeparation

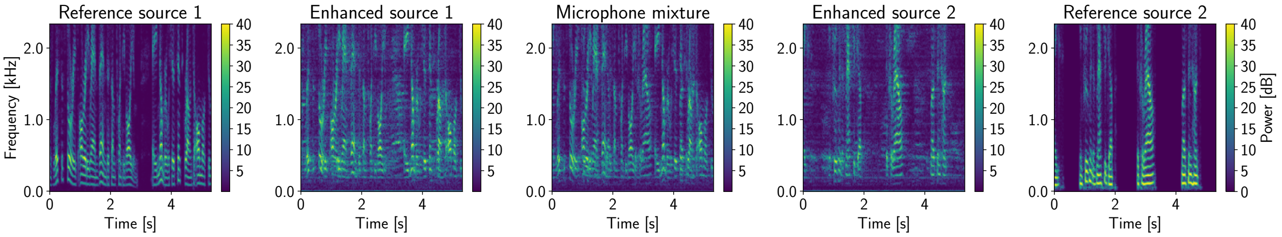

Figure 4: *Spectrograms of original and separated sources in a mixture sample from real-world recordings. SPS-SAD is used.*

Table 1: *Mean and standard deviation of BSSEval, STOI and PEASS scores obtained by averaging over different trials of the recorded dataset described in Section 4.1. OPS, APS, TPS and ITS are oracle, target, interference and artifact perceptual scores respectively. All scores are given in dB except for STOI.*

Method	SIR	SAR	SDR	STOI	PEASS - OPS	PEASS - APS	PEASS - TPS	PEASS - IPS
GCC-SAD	21.60 ± 2.21	-1.40 ± 1.75	-1.46 ± 1.75	0.86 ± 0.04	26.72 ± 7.64	19.62 ± 11.13	41.86 ± 11.78	72.72 ± 7.59
M-NICA-SAD	19.04 ± 3.38	-1.30 ± 1.77	-1.44 ± 1.78	0.83 ± 0.06	24.88 ± 5.40	17.10 ± 8.71	38.25 ± 13.60	70.74 ± 6.62
SPS-SAD	20.45 ± 3.02	-1.61 ± 1.68	-1.70 ± 1.70	0.83 ± 0.05	23.92 ± 6.24	16.27 ± 9.65	36.52 ± 11.73	75.60 ± 6.90
ELIM	22.16 ± 4.55	-0.79 ± 1.99	-0.83 ± 1.99	0.86 ± 0.04	27.76 ± 7.49	23.02 ± 11.59	41.97 ± 10.87	76.66 ± 5.71
BEST	27.89	1.17	1.15	0.89	33.21	33.00	67.17	87.11

for STOI and BSSEval showing very high interference rejection. Table 2 shows the average F-scores [26] calculated as

$$F_\beta = (1+\beta^2)\,\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \qquad (4)$$

where $\beta = 0.5$ to give more emphasis to the precision. This table also shows the average time that the SAD selected for calibration (CT in Table 2). The latter is obtained by counting the number of frames that have been assigned to a particular speaker. It is clear that GCC, having the best F-score among the considered methods, can select many correct frames for the calibration and this is also reflected in the BSSEval scores. Nevertheless, it is worth noting that even if the F-scores for M-NICA and SPS are lower then those from GCC, the BSSEval scores are not significantly lower. This is due to the fact that although the recall of the other two methods is not as high as GCC's, the precision is enough to select a sufficient number of correct frames for calibration.

4.4. Computational complexity of SAD

To compare the computational complexity of the three algorithms, we consider a complexity evaluation based on the computation in a time window. In the case of GCC and M-NICA, the computational complexity depends on the number of samples n in a fixed time window. In the case of SPS, the number of samples is variable and hence we consider the average number of samples $\tilde{n}$ in a time window. For the recordings used in this work, the ratio $n/\tilde{n} \approxeq 4.5$. Since the steps in the pipeline after the SAD are common to all three methods, they are not considered in the analysis of the computational complexity.

Table 2: *Mean and standard deviation for F-scores ($F_{0.5}$) of the SAD methods and total time used for calibration (CT).*

	GCC	M-NICA	SPS
$F_{0.5}$	0.75 ± 0.03	0.67 ± 0.04	0.67 ± 0.08
$CT(\mathrm{s})$	6.98 ± 1.58	4.46 ± 1.90	3.83 ± 1.85

In the case of GCC, because the FFT and inverse FFT operations dominate the computations, the complexity of the method is of $\mathcal{O}(n \log n)$. The complexity here is also proportional to the number of frequencies in the FFT, which directly affects the precision of localization. M-NICA, dominated by matrix-vector multiplication, has a complexity of $\mathcal{O}(n^2)$. For SPS, there are two parts. The first part which involves the computation of the ITDs, has a complexity of $\mathcal{O}(\tilde{n})$. The second part which computes the posterior probabilities, has a complexity of $\mathcal{O}(L\tilde{n})$, where L corresponds to the number of possible locations in the space used for the estimation (see Section 3.3.1 and [12]).

Under these considerations, SPS has the lowest computational complexity among all three methods. This is true as long as $n\log(n) > L\tilde{n}$. In fairness, if one does not need high precision in the localization output of GCC, one could reduce n to the point where both GCC and SPS have the same computational complexity and then compare the source separation results.

5. Conclusion

This work describes a framework for testing different implementations of a system that is able to do source separation using minimum variance beamforming. The methods considered for SA include GCC-PHAT, M-NICA, and SPS, an event-based localization method. The beamforming results on recordings in a real-world setting show that the SA methods can be implemented in real time and they lead to very similar BSSEval and STOI scores. While GCC seems to be the best method in terms of separation quality, SPS is a compelling alternative given its lower computational complexity. Finally, we show that it was possible to extract the intermittent "short" frames of speaker activity in a mixture of speakers and to use this combined set for calibrating the beamformer.

6. Acknowledgements

This work was partially supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 644732.

7. References

[1] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 436–443.

[2] J. Schmalenstroeer and R. Haeb-Umbach, "Online diarization of streaming audio-visual data for smart environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 845–856, Oct 2010.

[3] X. Anguera, C. Woofers, and J. Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, Nov 2005, pp. 426–431.

[4] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, Aug 1972.

[5] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel wiener filtering techniques for noise reduction," in *Speech Enhancement*. Berlin/Heidelberg: Springer-Verlag, 2005, pp. 199–228.

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

[7] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 425–437, 1997.

[8] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 1, April 2007, pp. I–41–I–44.

[9] K. Nakamura and T. Mizumoto, "Blind spatial sound source clustering and activity detection using uncalibrated microphone array," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 2438–2442.

[10] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[11] A. Bertrand and M. Moonen, "Energy-based multi-speaker voice activity detection with an ad hoc microphone array," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 85–88.

[12] J. Anumula, E. Ceolini, Z. He, A. Huber, and S. Liu, "An event-driven probabilistic model of sound source localization using cochlea spikes," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018.

[13] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[14] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, Sept 2011.

[15] P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin, "Subjective and objective quality assessment of single-channel speech separation algorithms," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 69–72.

[16] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multispeaker lcmv beamformer and postfilter for source separation and noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 940–951, May 2017.

[17] S. Markovich, S. Gannot, and I. Cohen, "A comparison between alternative beamforming strategies for interference cancellation in noisy and reverberant environment," in *2008 IEEE 25th Convention of Electrical and Electronics Engineers in Israel*, Dec 2008, pp. 203–207.

[18] S. U. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind Speech Separation and Enhancement with GCC-NMF," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 4, pp. 745–755, apr 2017.

[19] S. C. Liu, A. van Schaik, B. A. Minch, and T. Delbruck, "Asynchronous binaural spatial audition sensor with $2 \times 64 \times 4$ channel output," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 4, pp. 453–464, 8 2014.

[20] Y. Tsividis, "Event-driven data acquisition and digital signal processing; a tutorial," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 57, no. 8, pp. 577–581, Aug 2010.

[21] M. Miskowicz, "Send-on-delta concept: An event-based data reporting strategy," *Sensors*, vol. 6, no. 1, pp. 49–63, 2006.

[22] S. A. Shamma, "Speech processing in the auditory system i: The representation of speech sounds in the responses of the auditory nerve," *The Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1612–1621, 1985.

[23] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 7, 1982, pp. 1282–1285.

[24] I. Kiselev, E. Ceolini, D. Wong, A. d. Cheveigne, and S. C. Liu, "Whisper: Wirelessly synchronized distributed audio sensor platform," in *2017 IEEE 42nd Conference on Local Computer Networks Workshops (LCN Workshops)*, Oct 2017, pp. 35–43.

[25] Z. Huang, Z. Cao, D. Ying, J. Pan, and Y. Yan, "Time delay histogram based speech source separation using a planar array," in *INTERSPEECH*, 2017, pp. 1879–1883.

[26] D. M. W. Powers and Ailab, "Evaluation: from precision, recall and R-measure to ROC, informedness, markedness & correlation," *Journal of Machine Learning Technologies ISSN*, vol. 2, no. 1, pp. 2229–3981, 2011.