COOPERATIVE AUDIO SOURCE SEPARATION AND ENHANCEMENT USING DISTRIBUTED MICROPHONE ARRAYS AND WEARABLE DEVICES

Ryan M. Corey, Matthew D. Skarha, and Andrew C. Singer

University of Illinois at Urbana-Champaign

ABSTRACT

Augmented listening devices such as hearing aids often perform poorly in noisy and reverberant environments with many competing sound sources. Large distributed microphone arrays can improve performance, but data from remote microphones often cannot be used for delay-constrained real-time processing. We present a cooperative audio source separation and enhancement system that leverages wearable listening devices and other microphone arrays spread around a room. The full distributed array is used to separate sound sources and estimate their statistics. Each listening device uses these statistics to design real-time binaural audio enhancement filters using its own local microphones. The system is demonstrated experimentally using 10 speech sources and 160 microphones in a large, reverberant room.

Index Terms— Distributed microphone array, audio source separation, speech enhancement, augmented listening, hearing aids

1. INTRODUCTION

An important application of audio signal processing is to help people hear better in crowded, noisy environments. Augmented listening (AL) systems, such as hearing aids and augmented reality headsets, alter human perception by processing sound before it reaches the auditory system. Microphone arrays, which are used to filter signals spatially [1], can improve the performance of AL systems by separating sounds coming from different directions [2]. Large arrays can help AL systems to reduce noise more effectively, operate with lower delay [3], and preserve a listener's spatial awareness [4, 5].

Large wearable devices with microphones spread across the body can perform better than small earpieces [6]. Distributed arrays with sensors placed around a room could perform better still [7]. Microphone arrays are common in mobile and wearable devices, teleconferencing equipment, and smart-home appliances. If these devices could be aggregated into room-scale arrays, as shown in Fig. 1, they could dramatically improve the spatial diversity of listening systems. There has been significant recent research interest in distributed microphone arrays, including bandwidth-efficient distributed beamforming algorithms [8, 9], distributed blind source separation methods [10–13], and blind synchronization to compensate for sample rate mismatch between devices [14–17].

Unfortunately, distributed arrays often cannot be used directly for AL applications. In addition to bandwidth and computational limitations, listening devices are subject to severe delay constraints: they must process sound within a few milliseconds to avoid disturbing distortion or echoes [18, 19]. Even if remote microphones cannot be used directly for spatial filtering, however, they can still provide valuable information to the listening device. Spatial filters rely on estimates of the spatial characteristics of sound sources, such as their cross-correlation sequences or acoustic transfer functions [20]. Blind source separation [21] and channel estimation [22] methods are unreliable in large, reverberant spaces with many competing sound sources—the very environments in which humans most need help hearing. To reliably separate signals and estimate parameters in challenging environments, a single device is not enough.

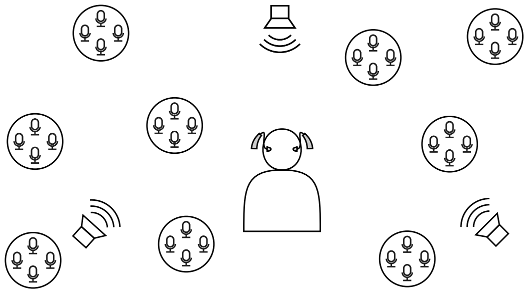

Fig. 1. A room might contain many audio devices, each with multiple microphones, that can cooperate to improve performance.

In this work, we show how AL devices can cooperate with each other and with other devices to improve performance in a real-time listening enhancement task. Due to delay, computation, and bandwidth constraints, a listening device might not be able to use data from remote devices to perform spatial filtering. Instead, the distributed array is used to separate signals and estimate their space-time statistics, as shown in Fig. 2. Each AL device uses these estimated parameters to design a real-time multimicrophone audio enhancement filter for its local microphones.

Because the proposed system would be quite complex to implement, we make several simplifying assumptions in this work. First, we assume that the sources and microphones do not move. We also assume that all devices are perfectly synchronized; for asynchronous array processing methods, see [23, 24]. To emphasize the benefits of spatial diversity from large arrays, we ignore the temporal structure of the source signals; for distributed methods that leverage speech signal sparsity, see [24–26]. Finally, although the reverberant acoustic channel is unknown, we assume that the number and rough locations of the sources are known.

2. REAL-TIME LISTENING ENHANCEMENT

Consider a room, like that in Fig. 1, with N sound sources and a total of M microphones spread across several devices, including at least one binaural listening device. All M microphones can be used to separate the sound sources and estimate the parameters of the acoustic channel. However, due to delay constraints, only $M_{\text{local}} \geq 2$ microphones are available to the listening device for real-time audio

978-1-7281-5549-4/19/$31.00 ©2019 IEEE

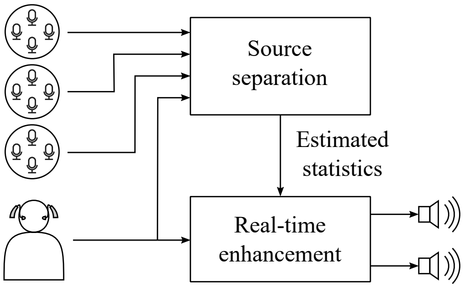

Fig. 2. Data from many devices is used to perform source separation and estimate signal statistics. Listening devices perform real-time filtering using only their own microphones.

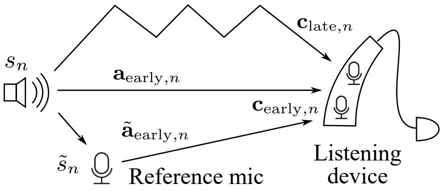
s_n
$\mathbf{a}_{\mathrm{early},n}$
$\mathbf{c}_{\mathrm{late},n}$
$\mathbf{c}_{\mathrm{early},n}$
$\tilde{\mathbf{a}}_{\mathrm{early},n}$
$\tilde{s}_n$
Reference mic
Listening device

Fig. 3. The signal due to source n is modeled as an early component and late reverberation.

enhancement. By convention, microphone 1 is in or near the left ear and microphone 2 is in or near the right ear.

Let $\mathbf{x}_{\mathrm{local}}[t] \in \mathbb{R}^{M_{\mathrm{local}}}$ be the signal received by these local microphones. It is modeled as a mixture of N source images, $\mathbf{c}_1[t], \ldots, \mathbf{c}_N[t]$ and noise $\mathbf{z}[t]$, so that

$$\mathbf{x}_{\mathrm{local}}[t] = \textstyle\sum_{n=1}^{N} \mathbf{c}_n[t] + \mathbf{z}[t]. \tag{1}$$

In a reverberant environment, each source image $\mathbf{c}_n[t]$ can be split into an early component $\mathbf{c}_{\mathrm{early},n}[t]$, which includes the direct path and early reflections, and late reverberation $\mathbf{c}_{\mathrm{late},n}[t]$, as shown in Fig. 3 [27]. There is no precise boundary between the early and late components, but it is assumed that each early component can be modeled by an acoustic impulse response $\mathbf{a}_{\mathrm{early},n}[k]$ so that

$$\mathbf{c}_{\mathrm{early},n}[t] = \textstyle\sum_{k=0}^{\infty} \mathbf{a}_{\mathrm{early},n}[k] s_n[t-k], \tag{2}$$

where $s_n[t]$ is the signal emitted by source n, for $n = 1, \ldots, N$.

2.1. Source remixing for augmented listening

Human augmented listening differs from other audio enhancement applications in two important ways. First, binaural devices must preserve spatial perception by maintaining the interaural cues between the left and right outputs for each source [4,5]. Second, processing delay must be no more than a few milliseconds to avoid perceptible distortion or echoes [18,19]. This delay constraint limits the achievable performance of the system [3].

The AL device enhances the user's perceived auditory scene by adjusting the levels of different sound sources, that is, by remixing them. For simplicity, suppose that the desired response for each source n is a scalar gain $g_n \geq 0$. The desired output signals $y_{\mathrm{L}}[t]$ at the left ear and $y_{\mathrm{R}}[t]$ at the right ear are

$$y_{\mathrm{L}}[t] = \textstyle\sum_{n=1}^{N} g_n \mathbf{e}_1^T \mathbf{c}_{\mathrm{early},n}[t] \quad \text{and} \tag{3}$$

$$y_{\mathrm{R}}[t] = \textstyle\sum_{n=1}^{N} g_n \mathbf{e}_2^T \mathbf{c}_{\mathrm{early},n}[t], \tag{4}$$

where $\mathbf{e}_m^T$ is the unit vector with a 1 in position m. Applying the same processing to the signals at the left and right ears ensures that interaural cues are preserved.

2.2. Delay-constrained listening enhancement

For brevity, we henceforth restrict our attention to the left output. A causal order-K finite impulse response filter $\mathbf{w}_{\mathrm{L}}[k] \in \mathbb{R}^{M_{\mathrm{local}}}$ produces an output signal

$$\hat{y}_{\mathrm{L}}[t] = \textstyle\sum_{k=0}^{K} \mathbf{w}_{\mathrm{L}}^T[k] \mathbf{x}_{\mathrm{local}}[t-k]. \tag{5}$$

Let the allowable delay be α samples so that $\hat{y}_{\mathrm{L}}[t]$ is an estimate of $y_{\mathrm{L}}[t-\alpha]$. To derive a minimum-mean-square-error (MMSE) estimator for the desired output, model $\mathbf{x}_{\mathrm{local}}[t]$ and $y_{\mathrm{L}}[t]$ as zero-mean wide-sense-stationary random processes. Let $\mathbf{r}_{xx}[k] = \mathbb{E}\left[\mathbf{x}_{\mathrm{local}}[t]\mathbf{x}_{\mathrm{local}}^T[t-k]\right]$ and $\mathbf{r}_{xy_{\mathrm{L}}}[k] = \mathbb{E}\left[\mathbf{x}_{\mathrm{local}}[t] y_{\mathrm{L}}[t-k]\right]$ be their auto- and cross-correlation functions, where $\mathbb{E}$ denotes expectation. Then the linear MMSE filter that estimates $y_{\mathrm{L}}[t-\alpha]$ given $\mathbf{x}_{\mathrm{local}}[t]$ is the time-domain multichannel Wiener filter [28]

$$\begin{bmatrix} \mathbf{w}_{\mathrm{L}}[0] \\ \mathbf{w}_{\mathrm{L}}[1] \\ \vdots \\ \mathbf{w}_{\mathrm{L}}[L] \end{bmatrix} = \begin{bmatrix} \mathbf{r}_{xx}[0] & \mathbf{r}_{xx}[1] & \cdots & \mathbf{r}_{xx}[K] \\ \mathbf{r}_{xx}[-1] & \mathbf{r}_{xx}[0] & & \\ \vdots & & \ddots & \\ \mathbf{r}_{xx}[-K] & & & \mathbf{r}_{xx}[0] \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{r}_{xy_{\mathrm{K}}}[\alpha] \\ \mathbf{r}_{xy_{\mathrm{L}}}[\alpha-1] \\ \vdots \\ \mathbf{w}_{\mathrm{L}}[\alpha-K] \end{bmatrix}. \tag{6}$$

The cross-correlation matrix can be decomposed in terms of the source images. From (3), we have

$$\mathbf{r}_{xy_{\mathrm{L}}}[k] = \textstyle\sum_{n=1}^{N} g_n \mathbf{r}_{xc_n}[k]\mathbf{e}_1, \tag{7}$$

where $\mathbf{r}_{xc_n}[k] = \mathbb{E}\left[\mathbf{x}_{\mathrm{local}}[t]\mathbf{c}_{\mathrm{early},n}^T[t-k]\right]$ for $n = 1, \ldots, N$. The listening device can easily estimate $\mathbf{r}_{xx}[k]$, but it would be difficult for it to estimate the source statistics $\mathbf{r}_{xc_n}[k]$ on its own.

3. COOPERATIVE PARAMETER ESTIMATION

Although remote devices cannot be used for real-time processing, they can be used to estimate the spatial statistics of the sound sources. If the remote devices are spread around and among the sound sources, then the distributed array will have far greater spatial resolution than the listening device alone. Furthermore, parameter estimation does not have a strict delay constraint, so the system can use several seconds or more of audio data.

3.1. Source separation using reference microphones

To learn the acoustic channel parameters, we would like to estimate the source signals $s_n[t]$ for $n = 1, \ldots, N$. Unfortunately, due to permutation and scale ambiguities, we cannot directly recover the sound produced by the source. Instead, we will estimate each source as observed by a nearby reference microphone.

A key advantage of distributed arrays is that some devices are much closer to some sources than others. Source separation algorithms can exploit this spatial diversity by, for example, assigning different sources to different devices [11–13]. Suppose that for each source n, there is a unique reference microphone m_n^* that is closest to it. Each reference microphone enjoys a higher signal-to-noise ratio and direct-to-reverberant ratio than more distant microphones for its corresponding source. Let $\tilde{s}_1[t], \ldots, \tilde{s}_N[t]$ be the set of source signals as observed by their respective reference microphones.

Let $\hat{s}_1[t], \ldots, \hat{s}_N[t]$ be estimates of these reference signals produced by a source separation algorithm. The cooperative system of Fig. 1 can be used with any multimicrophone source separation method. It remains an important open problem to develop scalable source separation algorithms that can take full advantage of massive-scale arrays in strongly reverberant environments with many sources. To assess the impact of source separation on the performance of the augmented listening system, the experiments in Sec. 4 compare three methods with different levels of separation performance:

1. A baseline unprocessed estimate, which is the input mixture at the nearest microphone to each source.
2. A blind source separation method known as independent vector analysis (IVA), which attempts to maximize the statistical independence between sources. We use the algorithm of [29] initialized with the nearest-microphone estimate.
3. An ideal linear MMSE filter that estimates each $\tilde{s}_n[t]$ using ground-truth acoustic channel parameters.

3.2. Estimation of second-order statistics

To compute the source-remixing filter derived in Sec. 2.2, we must find the second-order statistics of the early source images $\mathbf{c}_{\text{early},1}[t], \ldots, \mathbf{c}_{\text{early},N}[t]$. Because the true source signals $s_n[t]$ are not available, we cannot use the convolutional model (2) directly. Instead, we will use the *relative* early impulse responses (REIRs) [27] $\tilde{\mathbf{a}}_{\text{early},n}[k]$ with respect to the reference microphones:

$$\mathbf{c}_{\text{early},n}[t] = \textstyle\sum_{k=-\infty}^{\infty} \tilde{\mathbf{a}}_{\text{early},n}[k]\tilde{s}_n[t-k], \qquad (8)$$

for $n = 1, \ldots, N$. Notice that REIRs are noncausal in general, but if the reference microphone is close to its source, then the REIR should closely resemble the true early impulse response.

Because many source separation algorithms, including IVA, operate in the time-frequency domain, it will be convenient to compute signal statistics using the periodogram method. Let $\mathbf{X}_{\text{local}}[\tau, f]$ be the short-time Fourier transform (STFT) of $\mathbf{x}_{\text{local}}[t]$ and let $\hat{S}_n[\tau, f]$ be the STFT of $\hat{s}_n[t]$ for $n = 1, \ldots, N$. The sample statistics are

$$\hat{R}_{s_n s_n}[f] = \text{mean}_\tau \left|\hat{S}_n[\tau, f]\right|^2, \quad n = 1, \ldots, N, \qquad (9)$$

$$\hat{\mathbf{R}}_{x s_n}[f] = \text{mean}_\tau \mathbf{X}_{\text{local}}[\tau, f]\hat{S}_n^*[\tau, f], \quad n = 1, \ldots, N, \qquad (10)$$

$$\hat{\mathbf{R}}_{xx}[f] = \text{mean}_\tau \mathbf{X}_{\text{local}}[\tau, f]\mathbf{X}_{\text{local}}^H[\tau, f]. \qquad (11)$$

Note that these sample statistics are only correct if the signals are wide-sense stationary, which is not true in practice. By using the long-term average statistics, we ignore the temporal nonstationarity of the source signals and rely on spatial diversity alone.

If the sources and noise are uncorrelated with each other, then we can estimate the discrete-frequency relative transfer function of $\mathbf{c}_n[t]$ with respect to microphone m_n^* as

$$\hat{\mathbf{A}}_n[f] = \hat{\mathbf{R}}_{x s_n}[f]\hat{R}_{s_n s_n}^{-1}[f], \quad n = 1, \ldots, N. \qquad (12)$$

The relative early transfer function $\hat{\mathbf{A}}_{\text{early},n}[f]$ is obtained by time-domain windowing. The length of this window is a tunable parameter that, based on our experiments, does not appear to have a strong impact on objective performance.

The estimated cross-spectra between the mixture and images are

$$\hat{\mathbf{R}}_{x c_n}[f] = \hat{\mathbf{R}}_{x s_n}[f]\hat{\mathbf{A}}_{\text{early},n}^H[f], \quad n = 1, \ldots, N. \qquad (13)$$

The correlation functions required to compute the remixing filter are obtained by taking the inverse discrete Fourier transform of $\hat{\mathbf{R}}_{xx}[f]$ and $\hat{\mathbf{R}}_{x c_n}[f]$ for $n = 1, \ldots, N$.

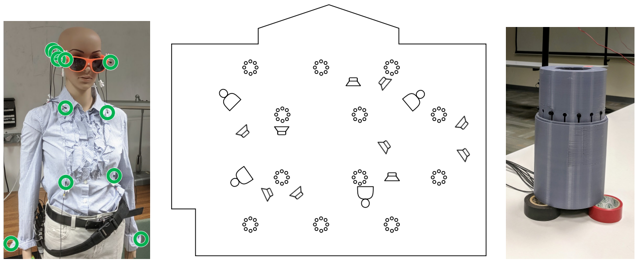

Fig. 4. Four wearable arrays (left), twelve tabletop arrays (right), and ten loudspeakers were spread around a large room.

4. EXPERIMENTS

4.1. Experimental setup

To demonstrate the proposed cooperative source separation and enhancement system in a challenging real-world environment, an experiment was conducted using 10 loudspeakers and a total of 160 omnidirectional microphones in a large (9 m by 13 m), strongly reverberant conference room ($T_{60} \approx 780$ ms), shown in Fig. 4. Twelve enclosures, designed to resemble voice-assistant speakers, held eight microphones each in a circular pattern with diameter 10 cm. The remaining 64 microphones were in wearable arrays on four plastic mannequins. Each had one microphone near each ear canal, four in behind-the-ear earpieces, two on eyeglasses, and eight on a shirt. Due to equipment limitations, recordings were captured using sixteen microphones at a time and the devices were moved between recordings while the ten loudspeakers remained fixed.

The loudspeakers played quasi-anechoic speech samples from ten talkers in the VCTK corpus [30]. To quantify the source separation and enhancement performance of the system, each sound source was played back and recorded separately to capture the source images, which can be added together to form test mixtures. The dataset is available on the Illinois Data Bank [31].

The closest microphone to each source was selected as the reference for source separation using the baseline, blind, and ideal methods described above. Source separation and parameter estimation were performed using 16 seconds of audio data. Different 16-second speech clips from the same talkers were used to evaluate the resulting listening enhancement filters. The filters have a target delay of 16 ms and an impulse response length of 128 ms. The length of the REIRs used to model the target sources' acoustics was 32 ms.

To consistently quantify performance, the enhancement filters (3) were designed to isolate a single source at a time, so that $g_n = 1$ for target source n and 0 for all others. A total of $8N$ single-target enhancement filters were designed, one for each source and each ear. The output signal-to-noise ratio (SNR) for a source separation or enhancement filter $\mathbf{w}$ designed to isolate source n is

$$\text{SNR}_n = 10\log_{10} \frac{\sum_t \left(\sum_k \mathbf{w}^T[k]\mathbf{c}_n[t-k]\right)^2}{\sum_t \left(\sum_{p\neq n}\sum_k \mathbf{w}^T[k]\mathbf{c}_p[t-k]\right)^2}. \qquad (14)$$

Note that separately recording and combining source images has the effect of amplifying ambient noise. For qualitative evaluation of enhancement results under more realistic conditions, the experiment was repeated using a simultaneous recording of all ten sources; binaural samples are available at the first author's website[1].

[1] http://ryanmcorey.com/demos

		Ideal			IVA		
Array	M	$N = 4$	7	10	$N = 4$	7	10
Reference	10	13	12	10	5	4	3
Wearable	64	25	26	23	8	7	3
Tabletop	96	23	23	21	7	6	5
All mics	160	23	24	23	8	7	6

Table 1. Median source separation performance measured by SNR improvement, in dB, between the estimated reference signal and the noisy signal at the reference microphone.

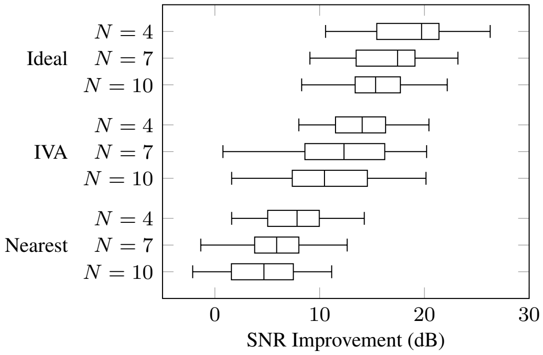

Fig. 5. Listening enhancement performance for different numbers of sources and different separation methods. All 160 microphones were used for separation. The boxes show the quartile statistics over all $8N$ source-output pairs.

4.2. Experimental results

The separation system was evaluated with several array configurations: the 10 reference microphones alone, the 4 wearable arrays, the 12 tabletop arrays, and all 160 microphones together. Table 1 shows the median SNR improvement of the estimated reference signal $\hat{s}_n[t]$ compared to the unprocessed nearest-microphone signal for $N = 4$, 7, and 10 sources. While IVA performs better than the baseline, especially for small N, it does not scale well with increasing array size and there is a large gap between its performance and that of the ideal filter. Notice that the wearable arrays outperform the tabletop arrays despite having fewer total microphones; the acoustically opaque body improves the spatial diversity of these arrays [6].

Figure 5 shows the SNR improvement of the AL device output compared to the unprocessed input at each ear. Within each experiment, lower SNR improvements generally correspond to distant source-listener pairs and larger improvements are for nearby source-listener pairs. For example, for the 7-source mixture using IVA parameters, the listener in the upper left corner in Fig. 4 achieved an 18 dB SNR improvement for the directly adjacent source but only a 1 dB improvement for the source in the opposite corner of the room.

Listening enhancement performance appears to depend strongly on source separation performance. There is a roughly 5 dB performance difference between the filters designed from the unprocessed reference microphone signals and those designed from the IVA estimates, showing that cooperative source separation did improve the performance of the individual AL devices. There is also a 5 dB difference between the IVA-based filters and those designed from ideal estimates, showing that there is room for improvement in distributed source separation.

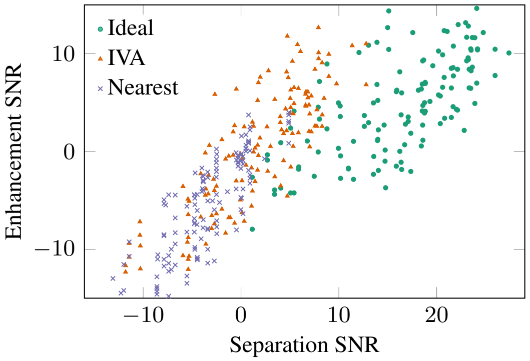

Fig. 6. Enhancement SNR as a function of separation SNR. Each point represents one source-to-ear filter. Results are sampled from all array and source configurations in Table 1.

To further illustrate the relationship between source separation and audio enhancement performance, Fig. 6 shows enhancement SNR as a function of separation SNR for individual source-microphone pairs. For the two non-ideal separation methods, there appears to be a roughly linear relationship: every 1 dB improvement in the separation SNR provides about 1 dB improvement in enhancement SNR. The ideal unmixing filter shows diminishing returns above around 10 dB. The vertical spread in the figure appears to be due to different distances between sources and listeners: nearby listeners achieve larger enhancement gains compared to distant listeners for the same source estimate.

5. CONCLUSIONS

The experiments presented above show that a distributed room-scale array can help listening devices to perform useful audio enhancement in challenging reverberant environments where source separation would otherwise be difficult or impossible. In contrast to other distributed methods, data from the distributed array is not used for real-time filtering but for parameter estimation. While a 16-microphone wearable array can provide strong enhancement performance on its own, it cannot reliably estimate the space-time statistics of the sources. The cooperative processing system can.

Because the enhancement filter is designed to match the estimated reference signal, its performance depends strongly on that of the source separation algorithm. Further research is required to find separation methods that can take advantage of massive spatial diversity to reliably separate large numbers of sources in strongly reverberant environments. With these improvements, the proposed cooperative audio separation and enhancement system will allow augmented listening devices to leverage all the microphones in a room, helping users to hear clearly in even the most challenging situations.

6. REFERENCES

[1] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[2] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[3] R. M. Corey, N. Tsuda, and A. C. Singer, “Delay-performance tradeoffs in causal microphone array processing,” in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[4] D. Marquardt, “Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques,” Ph.D. dissertation, Carl von Ossietzky University of Oldenburg, 2016.

[5] A. Koutrouvelis, “Multi-microphone noise reduction for hearing assistive devices,” Ph.D. dissertation, Delft University of Technology, 2018.

[6] R. M. Corey, N. Tsuda, and A. C. Singer, “Acoustic impulse response measurements for wearable audio devices,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[7] A. Bertrand, “Applications and trends in wireless acoustic sensor networks: A signal processing perspective,” in *IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, 2011, pp. 1–6.

[8] A. Bertrand and M. Moonen, “Distributed adaptive node-specific signal estimation in fully connected sensor networks—Part I: Sequential node updating,” *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5277–5291, 2010.

[9] S. Markovich-Golan, S. Gannot, and I. Cohen, “Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 343–356, 2012.

[10] F. Nesta and M. Omologo, “Cooperative Wiener-ICA for source localization and separation by distributed microphone arrays,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 1–4.

[11] I. Himawan, I. McCowan, and S. Sridharan, “Clustered blind beamforming from ad-hoc microphone arrays,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 661–676, 2010.

[12] Y. Hioka and W. B. Kleijn, “Distributed blind source separation with an application to audio signals,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 233–236.

[13] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, “Location feature integration for clustering-based speech separation in distributed microphone arrays,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 354–367, 2014.

[14] S. Markovich-Golan, S. Gannot, and I. Cohen, “Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming,” in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[15] S. Miyabe, N. Ono, and S. Makino, “Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation,” *Signal Processing*, vol. 107, pp. 185–196, 2015.

[16] L. Wang and S. Doclo, “Correlation maximization-based sampling rate offset estimation for distributed microphone arrays,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 571–582, 2016.

[17] D. Cherkassky and S. Gannot, “Blind synchronization in wireless acoustic sensor networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 651–661, 2017.

[18] M. A. Stone and B. C. Moore, “Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses,” *Ear and Hearing*, vol. 20, no. 3, pp. 182–192, 1999.

[19] J. Agnew and J. M. Thornton, “Just noticeable and objectionable group delays in digital hearing aids,” *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

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

[21] S. Makino, Ed., *Audio Source Separation*. Springer, 2018.

[22] Y. Huang, J. Benesty, and J. Chen, “Identification of acoustic MIMO systems: Challenges and opportunities,” *Signal Processing*, vol. 86, no. 6, pp. 1278–1295, 2006.

[23] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, “Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording,” in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 203–207.

[24] R. M. Corey and A. C. Singer, “Speech separation using partially asynchronous microphone arrays without resampling,” in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[25] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, “Distributed microphone array processing for speech source separation with classifier fusion,” in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012.

[26] M. Taseska and E. A. Habets, “Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1291–1304, 2016.

[27] O. Schwartz, S. Gannot, and E. A. Habets, “Multi-microphone speech dereverberation and noise reduction using relative early transfer functions,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2014.

[28] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer, 2008.

[29] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189–192.

[30] C. Veaux, J. Yamagishi, and K. MacDonald, “CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” 2017.

[31] R. M. Corey, M. D. Skarha, and A. C. Singer, “Massive distributed microphone array dataset,” 2019. [Online]. Available: https://doi.org/10.13012/B2IDB-6216881_V1