



Cooperative Speech Separation With a Microphone Array and Asynchronous Wearable Devices

Ryan M. Corey, Manan Mittal, Kanad Sarkar, and Andrew C. Singer

University of Illinois Urbana-Champaign, Urbana, Illinois, United States

corey1@illinois.edu

Abstract

We consider the problem of separating speech from several talkers in background noise using a fixed microphone array and a set of wearable devices. Wearable devices can provide reliable information about speech from their wearers, but they typically cannot be used directly for multichannel source separation due to network delay, sample rate offsets, and relative motion. Instead, the wearable microphone signals are used to compute the speech presence probability for each talker at each time-frequency index. Those parameters, which are robust against small sample rate offsets and relative motion, are used to track the second-order statistics of the speech sources and background noise. The fixed array then separates the speech signals using an adaptive linear time-varying multichannel Wiener filter. The proposed method is demonstrated using real-room recordings from three human talkers with binaural earbud microphones and an eight-microphone tabletop array.

Index Terms: speech separation, distributed microphone array, asynchronous microphone array, wearable devices

1. Introduction

Speech separation, the task of isolating one or more speech signals from a noisy mixture, is widely used in speech recognition, conferencing, and hearing enhancement systems [1, 2]. Microphone arrays can use multichannel filtering to separate speech signals arriving from different directions, but they must first learn the reverberant acoustic channels between the sources and microphones, which is difficult in the noisy environments where speech separation is most useful. In this work, we consider a conversation between several talkers of interest positioned near a microphone array in a noisy environment, such as a cafeteria, that also includes many unwanted sound sources.

Because individual devices often struggle in challenging acoustic environments, researchers have proposed distributed systems that aggregate data from multiple devices [3–6]. For example, smartphones have been used to augment a microphone array in a meeting room [7]. Wearable devices such as headphones are especially promising for speech separation because they are physically attached to talkers and move with them. The most direct approach to distributed source separation is joint beamforming, in which signals from all microphones in the network are coherently combined to form a spatial filter [4, 8, 9]. Unfortunately, wireless devices can suffer from network delays, packet loss, and sample rate offsets that make it difficult to

This research was supported by the National Science Foundation under Grant No. 1919257 and by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at the University of Illinois Urbana-Champaign, administered by Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence.

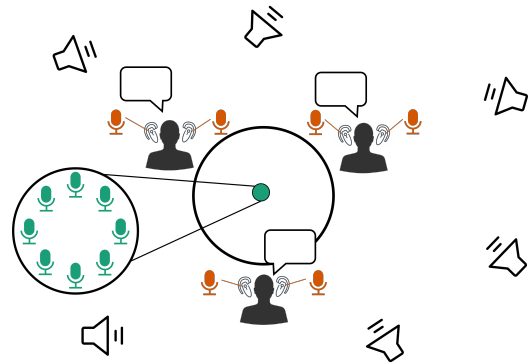


Figure 1: Microphones in wearable devices such as earbuds can complement traditional microphone arrays in noisy environments. In the experiment, three talkers with wearable microphones sat around a circular table with an eight-microphone array while loudspeakers generated background noise.

coherently combine signals from different devices [7, 10–14]. Wearable devices also move relative to each other, causing severe phase uncertainty at high frequencies [15].

Although wearable devices might not be useful for coherent spatial filtering, they can still help to estimate parameters. Time-frequency magnitude and speech presence probability (SPP) features are robust against small sample rate offsets and motion and so can be used for nonlinear source separation in an asynchronous array [16–18]. In real-time applications such as hearing enhancement, a distributed array can be used to estimate parameters for linear separation filters within each device [19]. Similarly, a well-positioned external microphone can be used to estimate the relative transfer function between a source of interest and the microphones of an array [20–22].

In this work, we propose a cooperative speech separation system that combines the strengths of a fixed array and of wearable microphones. The array has excellent achievable performance but needs help to learn the acoustic channel model. The wearable devices have poor signal-to-noise ratio (SNR) on their own, especially at high frequencies, but are useful for distinguishing between different sources because of their known positions relative to the talkers. The proposed system uses the wearable devices to estimate SPP values, which are then used to learn the second-order statistics for each source at the microphones of the fixed array. The array separates the sources using an adaptive linear time-varying spatial filter suitable for real-time applications. This work combines the cooperative architecture of [19], the distributed SPP method of [18], and the motion-robust modeling of [15]. The system is implemented adaptively and demonstrated using live human talkers.

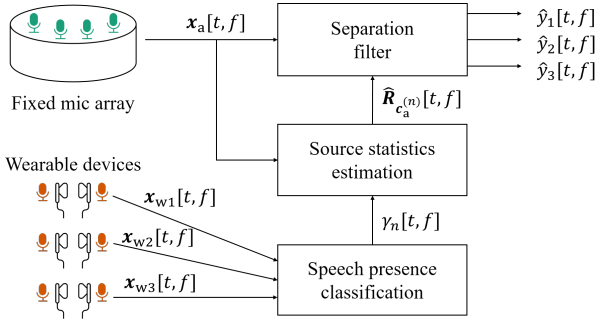


Figure 2: The proposed system uses a fixed microphone array for adaptive linear separation. The parameters are updated using speech presence probabilities estimated with a set of asynchronous wearable devices.

2. Time-frequency signal model

A group of N talkers of interest, each wearing a device containing one or more microphones, converse with each other near a fixed microphone array, as shown in Fig. 1. Let $\mathbf{x}[t, f] \in \mathbb{C}^M$ be the vector of short-time Fourier transform (STFT) signals captured from all M microphones, where t is the time index and f is the frequency index. The input vector can be partitioned into vectors captured by each of the devices, as shown in Fig. 2: $\mathbf{x}^T[t, f] = [\mathbf{x}_a^T[t, f], \mathbf{x}_{w1}^T[t, f], \dots, \mathbf{x}_{wN}^T[t, f]]$, where $\mathbf{x}_a[t, f]$ is from the fixed microphone array and $\mathbf{x}_{wm}[t, f]$ is from wearable device m .

The observed signal is a linear combination of speech signals from the N talkers and unwanted noise. Let $\mathbf{c}^{(n)}[t, f] \in \mathbb{C}^M$ be the component due to talker n for $n = 1, \dots, N$ and let $\mathbf{c}^{(N+1)}[t, f]$ be the noise, so that

$$\mathbf{x}[t, f] = \sum_{n=1}^{N+1} \mathbf{c}^{(n)}[t, f]. \quad (1)$$

The goal of source separation is to extract the signal component due to each talker from the mixture $\mathbf{x}[t, f]$. The desired outputs are $y_n[t, f] = \mathbf{e}_1^T \mathbf{c}^{(n)}[t, f]$ for $n = 1, \dots, N$, where $\mathbf{e}_1^T = [1, 0, \dots, 0]$.

In multichannel source separation, it is common to use linear estimators of the form $\hat{y}_n[t, f] = \mathbf{w}_n^H[t, f] \mathbf{x}[t, f]$. In this work we restrict our attention to the linear minimum-mean-square-error (LMMSE) estimator, also known as a multichannel Wiener filter (MWF), which is derived from the second-order statistics of the signals. Each mixture component $\mathbf{c}^{(n)}$ is modeled as a zero-mean time-varying random process with covariance $\mathbf{R}_{\mathbf{c}^{(n)}}[t, f]$. The sources and noise are assumed to be mutually uncorrelated so that the covariance of the mixture $\mathbf{x}[t, f]$ is given by

$$\mathbf{R}_{\mathbf{x}}[t, f] = \sum_{n=1}^{N+1} \mathbf{R}_{\mathbf{c}^{(n)}}[t, f]. \quad (2)$$

Assume that $\mathbf{R}_{\mathbf{x}}[t, f]$ is invertible, which is typically true in practice due to spatially diffuse noise. The MWF estimates are given by

$$\hat{y}_n[t, f] = \mathbf{e}_1^T \mathbf{R}_{\mathbf{c}^{(n)}}[t, f] \mathbf{R}_{\mathbf{x}}^{-1}[t, f] \mathbf{x}[t, f], \quad n = 1, \dots, N. \quad (3)$$

To implement the MWF, the system must estimate the source and mixture covariance matrices, which depend on the reverberant acoustic paths between each source and each microphone.

3. Second-order statistics

The challenges introduced by wearable devices, including sample rate offsets and relative motion, can be characterized by their effects on the second-order statistics of the signals. When using synchronous, nonmoving microphone arrays, it is common to assume a rank-one covariance model for each source of interest, where $\mathbf{R}_{\mathbf{c}^{(n)}}$ is proportional to the outer product of an acoustic transfer function vector that depends on the source and microphone geometry and room acoustics. In such a rank-one matrix, the signals from a given source observed at all microphones are perfectly correlated with one another. Sample rate offsets and relative motion reduce this coherence.

3.1. Sample rate offsets

In a distributed microphone array comprising multiple devices, each device uses a slightly different sample rate, causing relative time compression or expansion. Because these offsets tend to be small, on the order of parts per million (ppm), they can be modeled in the STFT domain as phase shifts [10, 11]. Suppose that the devices are coarsely time-aligned, e.g. to within several sample periods, so that the real-world time intervals of their corresponding STFT frames strongly overlap. Let $\mathbf{x}_{wm}^{\text{sync}}[t, f]$ be the signal that would have been captured by wearable device m if it used the same sample clock as the fixed array. The asynchronous signal is well approximated by

$$\mathbf{x}_{wm}^{\text{async}}[t, f] \approx e^{j2\pi f \alpha_m [t]} \mathbf{x}_{wm}^{\text{sync}}[t, f], \quad (4)$$

where $\alpha_m[t]$ is proportional to the sample rate offset of device m relative to the array.

In this asynchronous system, the covariance between the signals from wearable device m and from the array becomes

$$\mathbf{R}_{\mathbf{x}_a \mathbf{x}_{wm}}^{\text{async}}[t, f] \approx \mathbb{E} \left[e^{-j2\pi f \alpha_m [t]} \right] \mathbf{R}_{\mathbf{x}_a \mathbf{x}_{wm}}^{\text{sync}}[t, f], \quad (5)$$

where \mathbb{E} denotes statistical expectation. In some cases the offset can be estimated and this phase shift can be explicitly tracked [7, 13, 23]. If the offset is unknown, however, then it reduces the coherence between devices, especially at high frequencies. In particular, if the phase shift is treated as a uniform random variable on $(0, 2\pi)$, then the expectation in (5) is equal to zero. Then the between-device covariance becomes

$$\mathbf{R}_{\mathbf{x}_a \mathbf{x}_{wm}}^{\text{async}}[t, f] = \mathbf{0}, \quad m = 1, \dots, N. \quad (6)$$

By the same argument, the signals from different wearable devices are also uncorrelated with one another.

3.2. Relative motion

Wearable devices are also susceptible to unpredictable motion. Doppler effects are mathematically equivalent to sample rate offsets, but they vary unpredictably over time [14]. It has been shown that even small motion such as breathing can harm microphone array performance at high frequencies [15]. Consider an anechoic environment with a single sound source and assume that relative motion between microphones causes the time differences of arrival between devices to vary according to a normal distribution with variance σ^2 . Then it can be shown that the inter-device covariance is given by

$$\mathbf{R}_{\mathbf{x}_a \mathbf{x}_{wm}}^{\text{moving}}[t, f] = e^{-(2\pi f)^2 \sigma^2} \mathbf{R}_{\mathbf{x}_a \mathbf{x}_{wm}}^{\text{rigid}}[t, f], \quad (7)$$

where $\mathbf{R}^{\text{rigid}}$ is the covariance if the microphones did not move [15]. As with sample rate offsets, the effects of relative motion are stronger at high frequencies. If the uncertainty due to

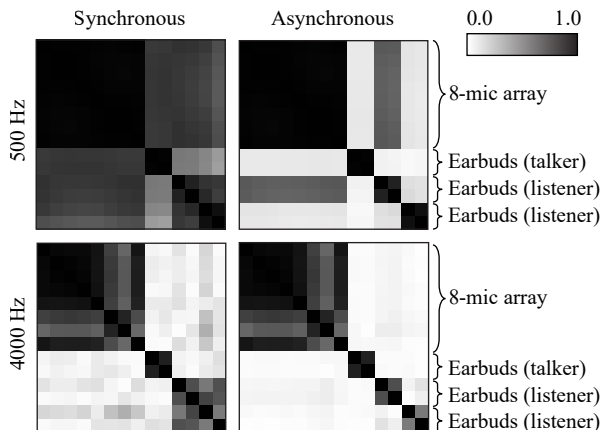


Figure 3: Magnitudes of sample correlation coefficients between microphones in a fixed array and three pairs of earbuds.

motion between two devices is large compared to the acoustic wavelength, then the signals become uncorrelated and the devices cannot be used for LMMSE estimation.

Figure 3 demonstrates the adverse effects of sample rate offsets and relative motion for wearable devices on human subjects. Each matrix shows the magnitudes of the sample correlation coefficients between signals at a fixed array of eight microphones and three pairs of earbuds for a single talker, averaged over a 60 second recording. The left column shows synchronous recordings and the right column shows the same recordings with simulated sample rate offsets between -50 and $+50$ ppm between the four devices. The eight microphones of the fixed array are always strongly correlated with one another, as are the two microphones in each pair of earbuds. However, the coherence between devices depends on frequency and synchronization. At high frequencies, relative motion and sample rate offsets reduce the between-device coherence to nearly zero so that the covariance matrix is approximately block diagonal.

3.3. Linear estimation with uncorrelated devices

To ensure robustness against sample rate offsets and motion, we adopt a conservative statistical model and set the between-device covariances to zero:

$$\mathbf{R}_{\mathbf{x}_a \mathbf{x}_{w_m}}[t, f] = \mathbf{0}, \quad m = 1, \dots, N \quad (8)$$

$$\mathbf{R}_{\mathbf{x}_{w_\ell} \mathbf{x}_{w_m}}[t, f] = \mathbf{0}, \quad \ell \neq m. \quad (9)$$

The overall covariance matrix $\mathbf{R}_{\mathbf{x}}$ is therefore block diagonal, with each block corresponding to one device. Applying (8) and (9) to the MWF (3), the filter coefficients are zero for the microphones of the wearable devices. Therefore, each source estimate is given by

$$\hat{y}_n[t, f] = \mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_a^{(n)}}[t, f] \mathbf{R}_{\mathbf{x}_a}^{-1}[t, f] \mathbf{x}_a[t, f], \quad (10)$$

which uses only the microphones of the fixed array. Because it does not use the wearable microphone signals to generate the output, this system is also less susceptible to network latency and packet loss compared to a coherent distributed beamformer, so it is appropriate for real-time applications such as hearing enhancement.

The wearable microphone signals cannot be used directly by the linear estimator, but they are still useful. Speech signals

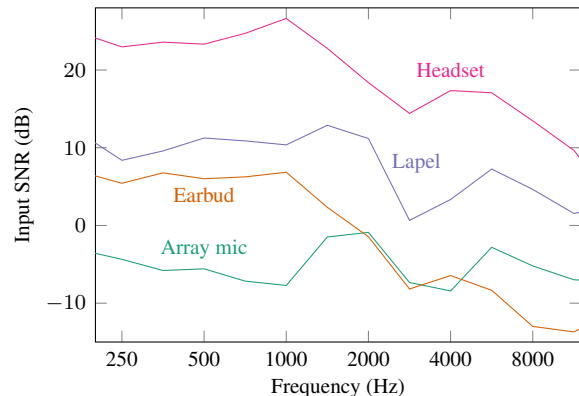


Figure 4: Average input SNR of wearable microphones for the wearer's speech.

are not Gaussian, so although the signals of different devices are uncorrelated, they are not independent. For example, the STFT magnitudes are robust against sample rate offsets and motion and convey useful information about speech presence at each time-frequency index. In [18], an offline algorithm aggregated SPP estimates from several wearable arrays to perform nonlinear source separation. Here we are interested in online processing, so the wearable microphone data is instead used to update the time-varying parameters of a linear separation filter.

4. Cooperative parameter estimation

If the wearable devices had excellent SNR for their respective talkers, then the wearable microphone signals could be used directly as source references to learn acoustic channel parameters [20–22]. In the group conversation scenario considered here, however, the users are close together and so the wearable devices suffer from significant cross-talk, especially at high frequencies. Figure 4 shows the input SNR at one microphone of the array and at headset, lapel, and earbud microphones on the talker of interest. Only the headset microphone, which is mounted just in front of the mouth, would be suitable as a reference signal for an adaptive separation filter. Our experiment used earbuds, which are ubiquitous among consumers, but at high frequencies the earbud microphones receive more speech energy from other talkers than they do from the wearer.

4.1. Adaptive covariance estimation

Because the wearable microphones have poor SNR and low between-device coherence, the source and mixture second-order statistics are tracked only for microphones of fixed array. First, the covariance of the mixture is updated recursively by

$$\hat{\mathbf{R}}_{\mathbf{x}_a}[t, f] = (1 - \mu) \hat{\mathbf{R}}_{\mathbf{x}_a}[t - 1, f] + \mu \mathbf{x}_a[t, f] \mathbf{x}_a^H[t, f], \quad (11)$$

where μ is a tunable step size parameter. In our experiments, it was 0.01, corresponding to a $1/e$ time of about 2 seconds.

To estimate the statistics of the speech sources, we can take advantage of the sparse structure of speech in the time-frequency domain [24]. At each $[t, f]$, the mixture is assumed to be dominated by a single speech source, so that $\mathbf{x}[t, f] \approx \mathbf{c}^{(n^*[t, f])}[t, f]$ for some $n^*[t, f] \in \{1, \dots, N + 1\}$. Single-microphone methods such as time-frequency masks often separate speech by assigning each time-frequency index of the mixture to a single source [25]. Multichannel systems can also take

advantage of this sparsity property, applying the same time-frequency mask to each microphone in the array and using the masked signals to estimate the second-order statistics [7, 9, 26].

Let $\gamma_n[t, f]$ be the SPP for source n at time-frequency index $[t, f]$. Treating the SPP as a soft mask for speech separation, we have $\mathbf{c}^{(n)}[t, f] \approx \gamma_n[t, f]\mathbf{x}[t, f]$. The source covariance matrices at the array microphones are recursively estimated as

$$\hat{\mathbf{R}}_{\mathbf{c}_a^{(n)}}[t, f] = (1-\mu)\hat{\mathbf{R}}_{\mathbf{c}_a^{(n)}}[t-1, f] + \mu\gamma_n[t, f]\mathbf{x}_a[t, f]\mathbf{x}_a^H[t, f], \quad (12)$$

for $n = 1, \dots, N + 1$.

4.2. Speech presence probability

Because they are attached to the talkers, the wearable microphones are well suited to perform SPP estimation. Assume a conditionally Gaussian model so that each $\mathbf{c}^{(n)}[t, f]$ has a normal distribution given its covariance $\mathbf{R}_{\mathbf{c}^{(n)}}[t, f]$. Then the log-likelihood of the observation $\mathbf{x}[t, f]$ given that source n^* is present is given by

$$\ln p(\mathbf{x}|n^*) = -\mathbf{x}^H \mathbf{R}_{\mathbf{c}^{(n^*)}}^{-1} \mathbf{x} - \ln \det(\pi^M \mathbf{R}_{\mathbf{c}^{(n^*)}}). \quad (13)$$

Because the covariance matrix is block-diagonal, the log-likelihood statistics can be computed independently within each wearable device and then summed:

$$\ln p(\mathbf{x}|n^*) = \sum_{m=1}^N -\mathbf{x}_{wm}^H \mathbf{R}_{\mathbf{c}_{wm}^{(n^*)}}^{-1} \mathbf{x}_{wm} - \ln \det(\pi^{M_m} \mathbf{R}_{\mathbf{c}_{wm}^{(n^*)}}), \quad (14)$$

where M_m is the number of microphones in device m . Finally, assuming uniform prior probabilities over the sources, the posterior probabilities are $\gamma_n[t, f] = p(\mathbf{x}[t, f]|n) / \sum_{\ell=1}^{N+1} p(\mathbf{x}[t, f]|\ell)$ for $n = 1, \dots, N + 1$.

The array microphones are excluded from the SPP estimate (14) because their statistics are unknown at this stage. The statistics of the wearable devices can be reliably measured in advance because the microphones are always in the same position relative to their wearers.

5. Experiment

The proposed system was evaluated using three human subjects seated around a circular table, as shown in Fig. 1. Each wore a pair of omnidirectional lavalier microphones on the ears to emulate earbuds. An eight-microphone circular array with diameter 10 cm was positioned in the center of the table. Each subject read aloud for 60 seconds. Background noise was generated by six loudspeakers playing clips derived from the VCTK corpus [27]. All microphones were recorded synchronously at 48 kHz and processed at 24 kHz. Simulated sample rate offsets between -50 and $+50$ ppm were applied to the wearable microphones using the interpolation method of [23].

To evaluate separation performance, the individual talkers and the background noise mixture were recorded separately. These recordings were summed to form the noisy mixture used for parameter estimation. The first 10 seconds of the recordings were used to calibrate the earbuds and the last 50 seconds were used for separation. The output SNR was computed by applying the time-varying separation filter separately to each source component, calculating SNRs in half-octave bands, and averaging over the three target sources. To allow the adaptive filters to converge, the first ten seconds of output were excluded from the SNR calculation.

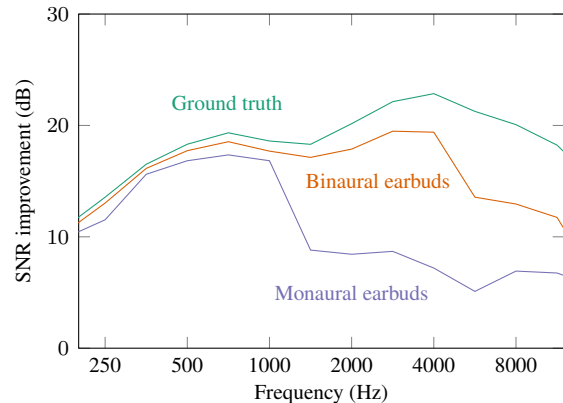


Figure 5: SNR improvement from a linear separation filter using SPP estimates from different devices on each talker.

Figure 5 shows the average SNR improvement of the three-talker separation filter relative to the input SNR at the tabletop array. The curves correspond to different SPP estimates used to update the filter: one using ground-truth source magnitudes, one using the three binaural pairs of earbud microphones, and one using only one earbud per talker. All filters perform similarly at low frequencies, where magnitude differences are reliable features for classification. At high frequencies, the multichannel classifier outperforms the single earbuds because the binaural pairs can better distinguish between speech from the wearer and from other talkers. However, the output SNR of the filter based on single-earbud classifiers still greatly exceeds the input SNR of the earbuds alone.

The experiment was repeated with and without simulated sample rate offsets. The separation results are indistinguishable, showing that the proposed method works without precise synchronization between devices.

6. Conclusions

The experiment demonstrates that wearable devices can be used to estimate acoustic channel parameters for a microphone array in a noisy environment. Noisy wearable devices, such as earbuds, cannot be used directly as source references but can help to compute source presence probabilities. Because the SPP estimates are computed independently within each wearable, the proposed system is robust against small sample rate offsets and relative motion. Furthermore, because the wearable microphones are used only for adaptation, this cooperative architecture is suitable for real-time processing even with severe network delays.

The system combines the strengths of each device: The fixed microphone array has excellent spatial resolution and can achieve strong linear separation performance, but cannot easily learn the source statistics in a complex acoustic environment. The wearable devices have poor individual SNRs at high frequencies, but have known positions relative to the talkers. By working together, the devices can learn the acoustic parameters of the sources and perform reliable separation in noise.

7. References

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

- [3] 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.
- [4] 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.
- [5] M. Taseska and E. A. Habets, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1195–1207, 2014.
- [6] 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.
- [7] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based MVDR beamformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5694–5698.
- [8] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 233–246, 2011.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] M. H. Bahari, A. Bertrand, and M. Moonen, "Blind sampling rate offset estimation for wireless acoustic sensor networks through weighted least-squares coherence drift estimation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 3, pp. 674–686, 2017.
- [14] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Estimation of sampling frequency mismatch between distributed asynchronous microphones under existence of source movements with stationary time periods detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 785–789.
- [15] R. M. Corey and A. C. Singer, "Motion-tolerant beamforming with deformable microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] R. M. Corey, M. D. Skarha, and A. C. Singer, "Cooperative audio source separation and enhancement using distributed microphone arrays and wearable devices," in *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019.
- [20] N. Göbbling and S. Doclo, "RTF-based binaural MVDR beamformer exploiting an external microphone in a diffuse noise field," in *ITG Symposium on Speech Communication*, 2018.
- [21] N. Göbbling, W. Middelberg, and S. Doclo, "RTF-steered binaural MVDR beamforming incorporating multiple external microphones," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 373–377.
- [22] N. Göbbling, D. Marquardt, and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 462–476, 2020.
- [23] 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.
- [24] S. Rickard and Ö. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 529–532.
- [25] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
- [26] T. Nakatani, R. Takahashi, T. Ochiai, K. Kinoshita, R. Ikeshita, M. Delcroix, and S. Araki, "DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6399–6403.
- [27] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017. [Online]. Available: <https://doi.org/10.7488/ds/1994>