SPEECH SEPARATION USING PARTIALLY ASYNCHRONOUS MICROPHONE ARRAYS WITHOUT RESAMPLING

Ryan M. Corey and Andrew C. Singer

University of Illinois at Urbana-Champaign

ABSTRACT

We consider the problem of separating speech sources captured by multiple spatially separated devices, each of which has multiple microphones and samples its signals at a slightly different rate. Most asynchronous array processing methods rely on sample rate offset estimation and resampling, but these offsets can be difficult to estimate if the sources or microphones are moving. We propose a source separation method that does not require offset estimation or signal resampling. Instead, we divide the distributed array into several synchronous subarrays. All arrays are used jointly to estimate the time-varying signal statistics, and those statistics are used to design separate time-varying spatial filters in each array. We demonstrate the method for speech mixtures recorded on both stationary and moving microphone arrays.

Index Terms— Asynchronous microphone array, ad hoc microphone array, distributed arrays, sampling rate offset, audio source separation, spatial filtering, speech enhancement

1. INTRODUCTION

Microphone arrays are useful for separating and enhancing audio signals because they can isolate sound sources coming from different directions [1]. Over the last few years, microphones have become ubiquitous in consumer electronic devices such as mobile phones, hearing aids and other listening devices, computers, gaming systems, and smart speakers. If many distributed microphones were combined into a single *ad hoc* array, they would provide greater spatial resolution and therefore better separation performance than any one of the devices alone [2–10].

Microphones on different devices are sampled at slightly different rates due to hardware variations. Although negligible in most applications, these offsets can be critical in array processing, which relies on precise phase relationships between microphones. Several asynchronous array processing methods have been proposed in the literature. In [5–8], the systems first estimate the sample rate offsets and resample the signals to a common rate. The resampled signals can then be combined coherently using conventional array processing techniques. Unfortunately, existing sample rate estimation algorithms are known to work poorly for moving sources [5, 6] and often do not work at all for moving microphones, as we will demonstrate in Section 2.1. In [9, 10], the sources are separated using single-channel masks that do not require resampling, but also do not take full advantage of the spatial diversity afforded by arrays. To separate sources in the most challenging environments, we need new asynchronous source separation techniques that do not require resampling and that scale well to devices with many microphones.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number DGE-1144245.

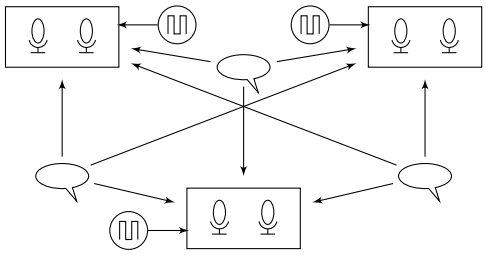

Fig. 1. We wish to separate K sources using M microphone arrays, each with its own sample clock.

In this contribution, we consider *partially asynchronous* microphone arrays in which some of the microphones do share a common sample clock but others do not, as shown in Figure 1. As microphones have become smaller and less expensive, many devices now include at least two. We can take advantage of this partial synchronization to perform multimicrophone source separation without resampling. In our proposed system, each device applies a separate linear time-varying spatial filter [11] to the signals collected by its local microphone array. The filter coefficients are computed using information about the source statistics from the full distributed array. For speech and other sparse signals, this shared information can take the form of source activity probabilities computed using spatial features from each array [10]. We demonstrate the proposed algorithm on real-world recordings of up to eight speech sources using both stationary and moving asynchronous microphone arrays.

2. ASYNCHRONOUS ARRAY PROCESSING

Consider a set of M distributed arrays and let $\mathbf{x}_{\mathrm{c},m}(t)$ be the vector of continuous-time signals captured by array m for $m = 1, \ldots, M$. The arrays need not have the same number of microphones. If the arrays shared a common sample period T, then the sampled discrete-time sequences would be $\tilde{\mathbf{x}}_{\mathrm{d},m}[\tau] = \mathbf{x}_{\mathrm{c},m}(\tau T)$ for integer time indices τ. Instead, each array m has its own sample period T_m, so that the sampled data vectors are $\mathbf{x}_{\mathrm{d},m}[\tau] = \mathbf{x}_{\mathrm{c},m}(\tau T_m)$ for $m = 1, \ldots, M$. The received signals are due to K independent sound sources, so that

$$\mathbf{x}_{\mathrm{d},m}[\tau] = \sum_{k=1}^{K} \mathbf{c}_{\mathrm{d},m,k}[\tau] \quad \text{for } m = 1, \ldots, M, \tag{1}$$

where $\mathbf{c}_{\mathrm{d},m,k}[\tau]$ is the response of array m to source k, which is often called the source image [12]. The sources may include both directional sound sources and diffuse noise. Our goal is to estimate one or more of the source images $\mathbf{c}_{\mathrm{d},m,k}[\tau]$ from the mixtures $\mathbf{x}_{\mathrm{d},1}[\tau], \ldots, \mathbf{x}_{\mathrm{d},M}[\tau]$.

978-1-5386-8151-0/18/$31.00

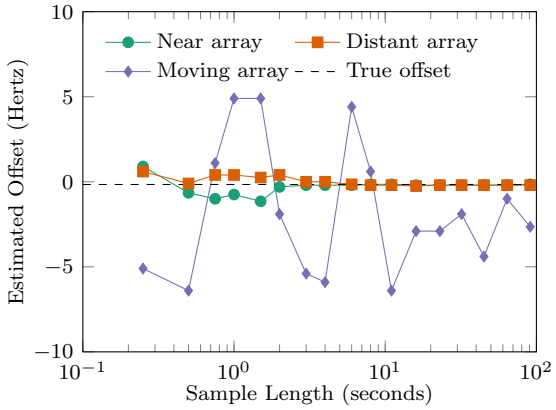

Fig. 2. Estimated sample rate offsets between closely spaced, distant, and moving arrays in an eight-talker cocktail party scenario (see Section 3.2) using handheld recorders and the two-stage correlation maximization algorithm [6].

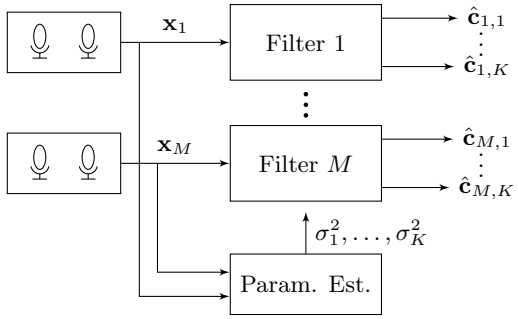

Fig. 3. Each device estimates each source image using its local microphones. The source powers are estimated using all M arrays.

2.1. Sample rate offset model

Let $\mathbf{c}_{m,k}[n,f]$, $\tilde{\mathbf{x}}_m[n,f]$, and $\mathbf{x}_m[n,f]$ be the short-time Fourier transform (STFT) vectors of the corresponding discrete-time sequences, where n is the frame index and f is the frequency index. Since each array has a different sample rate, the $[n,f]$ indices of each sequence $\mathbf{x}_m[n,f]$ correspond to slightly different continuous-time intervals and frequencies. We assume that the sample times are coarsely synchronized and that the sample rate offsets are sufficiently small that the sequences $\mathbf{x}_{\mathrm{d},m}[\tau]$ are offset from each other by much less than one STFT frame length over the period of interest. We can model the effect of those offsets by [6]

$$\mathbf{x}_m[n,f] = e^{j\alpha_m[n,f]}\tilde{\mathbf{x}}_m[n,f], \tag{2}$$

where $\alpha_m[n,f]$ is a phase shift due to the small sample rate offset at array m. Then, assuming that the sequences are zero-mean random processes, the across-array correlations are given by

$$\mathbb{E}\left[\mathbf{x}_m[n,f]\mathbf{x}_l^H[n,f]\right] = \mathbb{E}\left[e^{j(\alpha_m[n,f]-\alpha_l[n,f])}\tilde{\mathbf{x}}_m[n,f]\tilde{\mathbf{x}}_l^H[n,f]\right], \tag{3}$$

where $\mathbb{E}$ denotes expectation and H the Hermitian transpose. If the sample rate offsets are sufficiently small and time-invariant over the period of interest, then each $\alpha_m[n,f]$ is approximately proportional to $nf(T^{-1}-T_m^{-1})$ [6].

If the $\tilde{\mathbf{x}}_m[n,f]$ are approximately stationary over a long time interval, then the relative sample rate offsets can be estimated based on these cross-correlations [5–7] and the $\mathbf{x}_{\mathrm{d},m}[\tau]$ sequences can be resampled to obtain estimates of $\tilde{\mathbf{x}}_{\mathrm{d},m}[\tau]$. Correlation-based methods are known to be sensitive to source motion [5, 6]. Movement of the microphones themselves is fatal, since sample rate offsets and constant-velocity motion induce nearly identical linear phase shifts [13]. Figure 2 shows the performance of a blind sample rate estimation algorithm [6] in a cocktail party scenario. It works well when the microphones are stationary, even if they are far apart, but it fails when one microphone moves relative to the other. Thus, these algorithms are poorly suited to cocktail party scenarios with microphones worn or carried by moving humans.

Here, we consider a worst-case scenario in which we know little about the phase offsets between arrays. In particular, we model each $\alpha_m[n,f]$ as an independent random variable uniformly distributed from $-\pi$ to π. Under this model, since $\mathbb{E}\left[e^{j\alpha_m[n,f]}\right] = 0$, by linearity of expectation we have

$$\mathbb{E}\left[\mathbf{x}_m[n,f]\mathbf{x}_l^H[n,f]\right] = 0 \quad \text{for } m \neq l. \tag{4}$$

The captured sequences are thus uncorrelated across arrays. Assuming that the source images are uncorrelated with each other, their linear minimum mean square error estimators are given by the multichannel Wiener filters

$$\hat{\mathbf{c}}_{m,k}[n,f] = \mathbf{R}_{m,k}[n,f]\left(\sum_{k=1}^{K}\mathbf{R}_{m,k}[n,f]\right)^{-1}\mathbf{x}_m[n,f], \tag{5}$$

for $m = 1,\ldots,M$ and $k = 1,\ldots,K$, where each $\mathbf{R}_{m,k}[n,f] = \mathbb{E}\left[\mathbf{c}_{m,k}[n,f]\mathbf{c}_{m,k}^H[n,f]\right]$ is the time-varying source image covariance matrix. Since the images are due to both directional and diffuse sources, we assume that $\sum_{k=1}^{K}\mathbf{R}_{m,k}[n,f]$ is nonsingular for all m, n, and f. Thus, the linear estimators for the source images at each array use only the local microphones in that array. If each array has only a few microphones, then these filters might perform quite poorly compared to those for a synchronous distributed array.

2.2. Distributed spatial filtering

The multichannel Wiener filter (5) is often implemented using time-varying estimates $\hat{\mathbf{R}}_{m,k}[n,f]$ of the nonstationary source covariances [4,14–17]. Separation algorithms rely on good covariance estimates, and that is where we can take advantage of the asynchronous arrays. Although the sequences $\mathbf{x}_m[n,f]$ and $\mathbf{x}_l[n,f]$ are uncorrelated for $m \neq l$ due to their assumed-random phase shifts, they are not independent: both are generated by the same set of sources. Thus, we can use information from all M arrays to estimate the time-varying source statistics, then use those statistics to create M time-varying spatial filters. The proposed system is shown in Figure 3.

We will apply a variant of the full-rank local Gaussian model [16], in which we assume that each source image $\mathbf{c}_{m,k}[n,f]$ has zero mean and a conditional normal distribution given its covariance

$$\mathbf{R}_{m,k}[n,f] = \sigma_k^2[n,f]\bar{\mathbf{R}}_{m,k}[f], \tag{6}$$

where $\sigma_k^2[n,f]$ is the time-varying source spectrum and $\bar{\mathbf{R}}_{m,k}[f]$ is the frequency-dependent spatial covariance, which depends on the source and array geometry and room acoustics. For simplicity, we assume here that each $\bar{\mathbf{R}}_{m,k}[f]$ is time-invariant and that the full-rank covariance matrix accounts for uncertainty due to motion of the array. As is typically done with the local Gaussian model, we

assume that the $\mathbf{c}_{m,k}[n,f]$ are conditionally independent across n, f, and k given the source spectra $\sigma_1^2[n,f],\ldots,\sigma_K^2[n,f]$. Here, we further assume conditional independence across m, which reflects the uncorrelatedness of the array signals from (4).

The proposed estimation method is as follows:

1. Estimate the spatial parameters $\bar{\mathbf{R}}_{m,k}[f]$ using any suitable method. We show experimental results in Section 3 using both a blind method and a method based on training.
2. Find estimates $\hat{\sigma}_k^2[n,f]$ of the time-varying source spectra $\sigma_k^2[n,f]$ using the observations from all M arrays. We propose an estimator for sparse mixtures in Section 2.3.
3. Use the estimated source spectra and spatial parameters in (5) to estimate the source images at each array:

$$\hat{\mathbf{c}}_{m,k}[n,f] = \hat{\sigma}_k^2[n,f]\bar{\mathbf{R}}_{m,k}[f]\left(\sum_{s=1}^{K}\hat{\sigma}_s^2[n,f]\bar{\mathbf{R}}_{m,s}[f]\right)^{-1}\mathbf{x}_m[n,f]. \tag{7}$$

2.3. Joint spectral estimation for sparse sources

There are many methods to estimate time-varying source spectra, such as those based on expectation maximization [15, 16] and nonnegative matrix factorization [17]. Since here we are interested in speech sources, we will demonstrate a classification method that takes advantage of the time-frequency sparsity of speech [18]. The W-disjoint orthogonal model, which is most often used for single-channel methods such as time-frequency masks [19] but has also been applied for underdetermined multimicrophone separation [4, 20], assumes that for every $[n,f]$, we can assign a state $s[n,f] \in \{1,\ldots,K\}$ such that $\sigma^2_{s[n,f]}[n,f] \gg \sigma_k^2[n,f]$ for $s[n,f] \neq k$. To account for periods with no active directional sources, we include at least one stationary diffuse noise source in the model.

Let $\sigma^2_{k|s}[f]$ denote the variance of source k at frequency index f when the system is in state s. We model the variance as taking one of two values for each source, depending on the state:

$$\sigma^2_{k|s}[f] = \begin{cases} \sigma^2_{k,\text{high}}[f], & \text{if } k = s \\ \sigma^2_{k,\text{low}}[f], & \text{if } k \neq s. \end{cases} \tag{8}$$

Typical mask-based systems choose $\sigma^2_{k,\text{low}} = 0$, but since microphone arrays can steer multiple nulls at once, it is advantageous to include all sources in the model. Here, we choose $\sigma^2_{k,\text{high}}[f]$ and $\sigma^2_{k,\text{low}}[f]$ to be respectively 10 dB above and 10 dB below the long-term average source spectrum, which we have found to work well for speech sources [11]. The diffuse noise source has the same assumed spectrum in every state, and its magnitude can be tuned to improve the conditioning of the matrices in (7). In our experiments in Section 3, we use a spatially uncorrelated spectrum similar in power to that of the directional speech sources.

Under the local Gaussian model, the log-likelihood of the observations in state s is given by

$$\begin{aligned}\log p_s[n,f] = &-\sum_{m=1}^{M}\mathbf{x}_m^H[n,f]\left(\sum_{k=1}^{K}\sigma^2_{k|s}[f]\bar{\mathbf{R}}_{m,k}[f]\right)^{-1}\mathbf{x}_m[n,f] \\ &-\sum_{m=1}^{M}\log\det\left(\pi\sum_{k=1}^{K}\sigma^2_{k|s}(f)\bar{\mathbf{R}}_{m,k}[f]\right)\end{aligned} \tag{9}$$

Assuming uniform priors over all states, the posterior probability of state s is given by $\gamma_s[n,f] = p_s[n,f]/\left(\sum_{k=1}^{K}p_k[n,f]\right)$. Finally, the Bayesian estimate of each source power sequence is given by

$$\hat{\sigma}_k^2[n,f] \;=\; \sum_{s=1}^{K}\gamma_s[n,f]\sigma^2_{k|s}[f]. \tag{10}$$

Type	Mics	Resampled $K=3$	Resampled $K=4$	Not Resampled $K=3$	Not Resampled $K=4$
Unprocessed		-3.0	-5.0	-3.0	-5.0
Static	2	0.7	0.3	0.7	0.3
	8	**8.2**	**2.9**	2.1	0.1
Varying	2	1.3	0.5	1.3	0.5
	2×4	5.5	2.2	**5.5**	**2.2**

Table 1. Mean SDR performance, in dB, of several filters on the SiSEC ASY dev2 dataset [21].

3. SPEECH SEPARATION EXPERIMENTS

We demonstrate the performance of the proposed method in two scenarios using two different parameter estimation methods. We report the results using the signal-to-distortion ratio (SDR) criterion [12]:

$$\text{SDR}_{m,k} = 10\log_{10}\frac{\sum_\tau|\mathbf{c}_{\text{d},m,k}[\tau]|^2}{\sum_\tau|\hat{\mathbf{c}}_{\text{d},m,k}[\tau]-\mathbf{c}_{\text{d},m,k}[\tau]|^2}. \tag{11}$$

3.1. SiSEC ASY

To understand the performance of the proposed resampling-free source separation method, we first compare it to resampling-based methods. In this section, we describe our contribution to the 2018 Signal Separation Evaluation Campaign (SiSEC) asynchronous source separation (ASY) task [21]. We show results for Task 2, which is to separate either $K=3$ or $K=4$ talkers from recordings made by $M=4$ portable recorders with two microphones each.

Because the sources and microphones are fixed in this scenario, it is possible to estimate the sample rate offsets and correct for them before applying ordinary synchronous blind source separation techniques. Two of the three contributions to SiSEC 2015 ASY, which used the same data set, adopted this approach [22]. Our baseline resampling implementation combines these two approaches from SiSEC 2015: first, we use two-stage correlation maximization [6] to estimate the sample rate offsets, then correct them using Lagrange interpolation [23]. The sources are blindly separated using offline independent vector analysis [24], and we infer the sources' rank-one covariance matrices from the resulting unmixing filters. We use these blindly estimated covariance matrices to design the four separation filters compared in the rows of Table 1: separate static two-channel Wiener filters for each recorder; a single static Wiener filter using all eight microphones; separate time-varying two-channel filters for each recorder; and finally the proposed method, with four time-varying two-channel filters designed using a common set of estimated source power sequences. Each filter is tested with and without resampling the signals.

When the signals are resampled before separation, the synchronous eight-channel filter outperforms all other methods. When we restrict the filters to use two microphones, the separation problem is underdetermined, so the time-varying filters perform better than the static filter. In fact, when using the other recorders to classify the active source, the two-channel filter performs nearly as well as

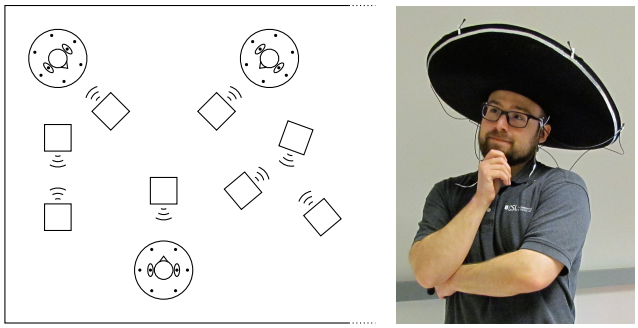

Fig. 4. Left: Cocktail party layout with eight loudspeakers and three human listeners. Right: Each human listener wears two in-ear microphones and a hat with six additional microphones.

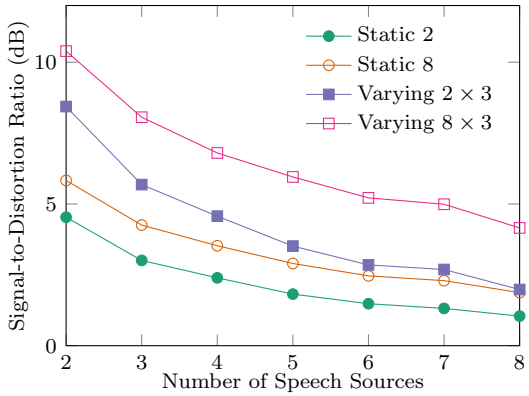

Fig. 5. Experimental results for a cocktail party scenario with moving wearable microphone arrays. The SDR is averaged over the left and right ears of all three listeners and over all sources.

the static eight-channel filter. Next, we test the four filters without resampling the signals. The two-channel filters are not affected, since the two microphones of each recorder are synchronously sampled. The eight-channel filter performs much worse since it relies on across-array coherence. The proposed asynchronous time-varying filter performance is identical with or without resampling, suggesting that it is resilient to sample rate offsets.

3.2. Cocktail party scenario with moving wearable arrays

The proposed method performs worse than previously proposed methods for the SiSEC ASY data set, which is amenable to resampling, but it should be better suited to moving arrays for which resampling is difficult or impossible. We now consider a listening enhancement experiment in which microphone arrays are attached to moving human listeners in a cocktail party scenario. In this scenario there are up to eight simultaneous speech sources. Since current blind source separation techniques are poorly suited to such large mixtures, and since we wish to demonstrate the achievable performance of an asynchronous array system, we use measured rather than estimated spatial parameters for this experiment.

The recordings were made in the Augmented Listening Laboratory at the University of Illinois at Urbana-Champaign, which has a reverberation time of about $T_{60} = 300$ ms. The cocktail party scenario, shown in Figure 4, consists of eight talkers, which were simulated using loudspeakers playing clips from the VCTK anechoic speech database [25], and three real human listeners. Each human listener wore a head-mounted array of eight omnidirectional lavalier microphones: one in each ear and six affixed to a rigid, wide-brimmed hat with diameter 60 cm. The listeners moved their heads continuously during the recordings, alternately nodding, looking around the room, and shifting from side to side.

The twenty-four signals were recorded on a single interface, sampled at at 16 kHz, and highpass filtered from 100 Hz to remove low-frequency ambient noise. Artificial sample rate offsets of ± 0.3 Hz were applied to two arrays using Lagrange interpolation [23]. The STFT was computed with a length-4096 von Hann window and 75% overlap. The spatial covariance matrices $\tilde{\mathbf{R}}_{m,k}[f]$ were estimated using 5-second training clips from the same talkers and with similar listener motion as the 15-second test clips. Because they are designed for binaural listening devices, the filters produce only the source image estimates for the microphones in the ears, not for those on the hat. To measure the source images, the source signals were recorded individually and then superimposed to form a mixture. This procedure allows us to measure the ground truth SDR, but it is physically unrealistic because the human motion is different in every source recording.[1]

Figure 5 compares the separation performance of four arrays: a static array of two in-ear microphones, a static array of all eight microphones, a time-varying asynchronous array of two microphones per listener, and a time-varying asynchronous array of eight microphones per listener. It is noteworthy that the distributed array of two microphones per listener outperforms the eight-microphone static array, even when there are eight sources. The distributed classifier helps to resolve ambiguities between sources that have similar transfer functions to the individual arrays. It is particularly important for moving arrays: when a listener turns their head from side to side, the classifier can use the other two arrays to decide which source they are hearing. This feature requires no explicit modeling of head motion; it is a consequence of the full-rank spatial covariance model and conditional independence between subarrays.

4. CONCLUSIONS

The experimental results from Section 3 show that the proposed asynchronous separation method can effectively separate speech mixtures even when there are more sources than microphones on each device. The SiSEC results show that it does not perform as well as a synchronized stationary array, but it does outperform a single device and does not require sample rate offset estimation or resampling. The results from the cocktail party scenario show that the time-varying filters and state classifier work with moving microphones and scale well to larger arrays. The distributed classifier is particularly useful for resolving ambiguities when the arrays move or when sources are far away.

The time-varying filters and classifier both rely on accurate estimation of the source spatial covariances. In this work, we have not proposed a method to estimate these parameters without either resampling-based blind source separation or training data, nor do we explicitly model their variation over time; asynchronous parameter estimation and tracking remain important challenges for future work. The proposed asynchronous source separation system is well suited to distributed arrays in which individual devices have multiple microphones, are far apart, and are mobile.

[1] Separated sound samples using real simultaneous recordings are available on the first author's website.

5. REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[2] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 38–51, 2009.

[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)*, pp. 1–6, 2011.

[4] 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.

[5] 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.

[6] 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.

[7] M. H. Bahari, A. Bertrand, M. Moonen, 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.

[8] 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.

[9] 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)*, pp. 203–207, 2014.

[10] 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.

[11] R. M. Corey and A. C. Singer, "Underdetermined methods for multichannel audio enhancement with partial preservation of background sources," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

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

[13] D. Cherkassky and S. Gannot, "Blind synchronization in wireless sensor networks with application to speech enhancement," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 183–187, 2014.

[14] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.

[15] M. A. Dmour and M. Davies, "A new framework for underdetermined speech extraction using mixture of beamformers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 445–457, 2011.

[16] N. Q. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[17] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[18] 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, pp. 529–532, 2002.

[19] Ö. 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.

[20] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[21] A. Liutkus, F.-R. Stoter, and N. Ito, "The 2018 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*, 2018.

[22] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*, pp. 387–395, 2015.

[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] 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)*, pp. 189–192, 2011.

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