High/Low Model for Scalable Multimicrophone Enhancement of Speech Mixtures

Ryan M. Corey and Andrew C. Singer
University of Illinois Urbana-Champaign

***Abstract*—Many speech separation and enhancement methods take advantage of time-frequency sparsity by assuming that only one speech source in a mixture has nonzero power at each time and frequency. This "on/off" model is valuable for systems with more sources than microphones, but many methods that use it do not benefit from the spatial diversity of systems with large numbers of microphones. This work considers the high/low model, in which one source is strongest at each time-frequency index but all sources have nonzero power. A time-varying enhancement method using the high/low model combines the benefits of sparsity and spatial diversity and scales automatically with the number of microphones, resembling a time-frequency mask for underdetermined systems and a linear filter for overdetermined systems. The model is demonstrated using real-room data with up to 10 speech signals and between 1 and 160 microphones.**

***Index Terms*—Microphone arrays, speech enhancement, source separation**

I. Introduction

In many audio signal processing applications, a system must process individual speech signals from a mixture. Speech separation and enhancement algorithms [1] typically belong to one of two categories: Systems with one or a few microphones rely on data-driven models of speech, while systems with many microphones can separate sounds spatially. Relatively few algorithms use both signal models and spatial information, especially for large arrays. As microphones proliferate in human environments, there is a need for methods that can scale from one to hundreds of sensors.

In principle, if the number of microphones exceeds the number of sources, and if the acoustic channel is known and fixed, then the source signals can be recovered using a linear time-invariant filter. In practice, it is helpful to have more microphones to improve robustness against noise, reverberation, and parameter estimation errors. If the number of microphones is smaller than the number of sources—or too small to be robust—then linear time-invariant filters are not enough.

Many underdetermined algorithms rely on sparsity. In the time-frequency (TF) domain, speech mixtures exhibit W-disjoint orthogonality [2]: At every TF index, most of the energy of the short-time Fourier transform (STFT) of the mixture can be attributed to one source. Therefore, a separation system can assign each TF sample to a single talker, a process known as time-frequency masking. Underdetermined methods differ in how they classify the dominant source. Single-microphone systems typically rely on compositional models [3] or machine learning algorithms [4] to decompose magnitude spectra. For small arrays, the DUET method [5] and its variants [6]–[9] use spatial information to cluster samples. Some authors have used spatial features as inputs to learning-based classifiers [10]. For mixtures of more than a few sources, multimicrophone methods can model multiple simultaneously active sources at each TF index [11]–[14].

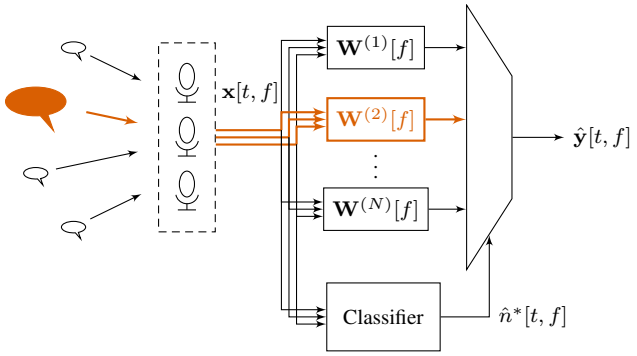

Fig. 1. A time-varying multichannel filter switches between several static filters at each time-frequency index. Each filter is designed for one "high" source and several "low" sources.

In the literature, sparse models are most popular for underdetermined separation systems with more sources than microphones, and have sometimes been used to estimate source statistics for linear beamforming [15], [16]. However, even large microphone arrays could benefit from sparsity: If a set of sources can be ignored at a given TF index, then more degrees of freedom are available to improve robustness against noise and channel estimation errors. Likewise, time-varying methods could leverage spatial diversity to reduce the distortion and artifacts that occur when signals do not obey the W-disjoint orthogonality assumption and when classifiers make errors. As large microphone arrays become more practical and data-driven TF separation methods become more powerful, it is important to develop algorithms that combine the benefits of both sparisty and spatial diversity.

This work presents the high/low model, a generalization of the on/off signal model that motivates TF masks and other sparse methods. Instead of assuming that one sound source contributes all energy at each TF index, the model

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 at 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.

assumes that one source contributes more than the others. This subtle distinction has negligible effect in single-microphone systems, but a large impact as the number of microphones increases. A time-varying system using the model prioritizes the dominant source, but also attempts to process the weaker sources. The proposed system, shown in Fig. 1, resembles a conventional TF mask in single-microphone systems and a linear time-invariant filter for many-microphone systems, scaling gracefully between the two extremes.

The proposed model is conceptually related to soft masks, which multiply each TF sample by a value between 0 and 1, for example based on estimated speech presence probability [17]. However, soft masks are often single-channel postfilters, whereas the proposed time-varying filter changes the spatial pattern of the beamformer at each TF index. It shares the computational advantage of masks since it switches between a finite number of states rather than calculating a new filter at each TF index. The high/low model has previously been applied to underdetermined systems to better preserve binaural cues [18] and to aggregate information from asynchronous distributed sensors [19]. However, these papers focused on applications rather than the model itself and did not consider larger arrays. This paper motivates the high/low model based on empirical data, uses it to derive a time-varying discrete-state speech enhancement filter, and analyzes performance scaling with array size in real-room experiments using both ideal parameters and estimates based on established blind source separation techniques.

II. Multimicrophone speech enhancement

A. Time-frequency mixing and enhancement

Consider a mixture of N speech signals captured by M microphones. Let $\mathbf{s}[t,f] = [s_1[t,f],\dots,s_N[t,f]]^T$ be the vector of STFTs of the speech signals, let $\mathbf{x}[t,f] \in \mathbb{C}^M$ be the vector of STFTs of the microphone signals, and let $\mathbf{z}[t,f] \in \mathbb{C}^M$ be the vector of STFTs of the non-speech noise signals at the microphones. For simplicity, this work assumes the multiplicative transfer function model,

$$\mathbf{x}[t,f] = \mathbf{A}[f]\mathbf{s}[t,f] + \mathbf{z}[t,f], \tag{1}$$

for all frequency indices f, where $\mathbf{A}[f] \in \mathbb{C}^{M\times N}$ is a matrix of acoustic transfer functions or relative transfer functions. We assume that an estimate of $\mathbf{A}[f]$ is available, for example from a set of pilot signals or a blind source separation algorithm.

The desired output $\mathbf{y}[t,f] \in \mathbb{C}^J$ of the enhancement system is a linear time-invariant combination of the source signals:

$$\mathbf{y}[t,f] = \mathbf{G}[f]\mathbf{s}[t,f], \tag{2}$$

where $\mathbf{G}[f] \in \mathbb{C}^{J\times N}$ is a matrix of desired responses. In a source separation system, $\mathbf{G}$ would be an $N \times N$ identity matrix, while in a binaural remixing system [20], it would be a $2 \times N$ matrix whose columns are head-related transfer functions. The system estimates $\mathbf{y}[t,f]$ from $\mathbf{x}[t,f]$ using a time-varying filter $\mathbf{W}[t,f] \in \mathbb{C}^{J\times M}$. The output $\hat{\mathbf{y}}[t,f]$ is

$$\hat{\mathbf{y}}[t,f] = \mathbf{W}[t,f]\mathbf{x}[t,f], \tag{3}$$

For brevity, we omit the frequency index f in the remainder of the paper; all variables are functions of frequency.

In this work, we restrict our attention to the linear minimum-mean-square-error estimator, also known as a multichannel Wiener filter (MWF). Suppose that each TF sample $s_n[t]$ is a zero-mean random variable with time-varying variance $r_n[t]$, that $\mathbf{z}[t]$ is a zero-mean random vector with full-rank covariance matrix $\mathbf{R_z}$, and that the speech signals and noise are mutually uncorrelated. Let $\mathbf{R_s}[t] = \mathrm{diag}\{r_1[t],\dots,r_N[t]\}$. If the channel matrix and variances are known, the MWF is

$$\mathbf{W}[t] = \mathbf{G}\mathbf{R_s}[t]\mathbf{A}^H\left(\mathbf{A}\mathbf{R_s}[t]\mathbf{A}^H + \mathbf{R_z}\right)^{-1}. \tag{4}$$

B. The on/off model and time-frequency masks

The spectra of speech signals vary rapidly over time, so a system using the time-varying MWF (4) would need to estimate a different set of $r_n[t]$ parameters and recompute the filter at each TF index. We can simplify the problem using the orthogonality property [2]: At each $[t,f]$ there exists a dominant source $n^*[t] \in \{1,\dots,N\}$ such that

$$\left|s_{n^*[t]}[t]\right|^2 \gg |s_n[t]|^2, \quad n \neq n^*[t]. \tag{5}$$

Assuming that the non-speech noise is also much weaker than the dominant signal, this property implies that

$$\mathbf{x}[t] \approx \mathbf{A}_{n^*[t]}s_{n^*[t]}[t], \tag{6}$$

where $\mathbf{A}_n$ is the column of $\mathbf{A}$ corresponding to source n.

This property can be incorporated into the statistical mixing model by selecting binary values for $r_n[t]$:

$$r_n[t] = \begin{cases} r_{\mathrm{on},n}, & \text{if } n^*[t] = n, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

We call this the on/off model. The MWF (4) under the on/off model is $\mathbf{W}[t] = \mathbf{W}^{(n^*[t])}$, where

$$\mathbf{W}^{(n)} = \mathbf{G}_n r_{\mathrm{on},n}\mathbf{A}_n^H\left(\mathbf{A}_n r_{\mathrm{on},n}\mathbf{A}_n^H + \mathbf{R_z}\right)^{-1}. \tag{8}$$

In the single-microphone case, the output is

$$\hat{\mathbf{y}}[t] = \mathbf{G}_{n^*[t]}\frac{r_{\mathrm{on},n^*[t]}}{r_{\mathrm{on},n^*[t]} + r_z}x[t]. \tag{9}$$

This system applies a scalar TF mask to the single-microphone input and then applies the desired response for source $n^*[t]$. Because the output at each TF index is parallel to a single column of $\mathbf{G}$, the components of all non-dominant sources present in $x[t]$ will have incorrect processing applied, which can introduce distortion even with an error-free source activity classifier. For example, in a binaural system, the non-dominant source components would be presented with the interaural cues of the dominant source.

III. Speech enhancement with the high/low model

A. The high/low model

Systems designed using the on/off model, even those with multiple microphones, ignore the $N-1$ inactive sources at each TF index. A system with many microphones could attempt to process the non-dominant sources as well. To account

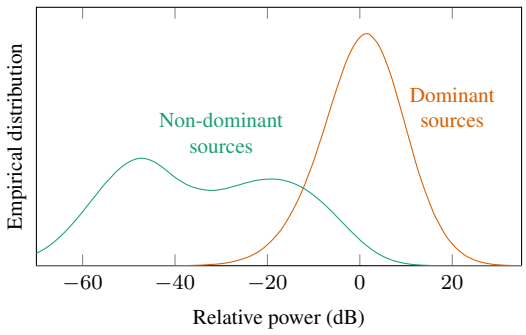

Fig. 2. Empirical distribution of "high" and "low" signal levels at each time-frequency index for an anechoic mixture of eight speech signals. Levels are scaled so that the average mixture power is 0 dB at each frequency.

for all sources while still only performing one classification at each TF index, we can replace the on/off model with the high/low model.

At each TF index, there is still a single dominant source, as in (5). However, the signal variances are given by

$$r_n[t] = \begin{cases} r_{\text{high},n}, & \text{if } n^*[t] = n, \\ r_{\text{low},n}, & \text{otherwise}, \end{cases} \tag{10}$$

where $r_{\text{high},n}$ is the variance of $s_n[t]$ when it is the dominant source and $r_{\text{low},n}$ is its variance when it is not dominant.

To see why the high/low model makes sense, consider the empirical distribution of speech energy shown in Fig. 2. This plot was generated from a mixture of eight quasi-anechoic speech recordings from the VCTK dataset [21] using 64 ms STFT windows. While the dominant source has consistently strong energy, non-dominant sources have a wide distribution of energy. The non-dominant sources are often negligible, fitting the on/off model, but sometimes they contribute substantial energy to the mixture. Further empirical results on high/low ratios are discussed in Sec. IV-A below.

The on/off model is a special case of the high/low model where $r_{\text{low},n} = 0$ for all n. Meanwhile, if $r_{\text{high},n} = r_{\text{low},n}$ for all n, we have a time-invariant signal model and the resulting filter will be time-invariant. Thus, the ratio can be used to tune the system behavior and need not match the empirical high-to-low ratio for the mixture.

B. Discrete-state enhancement filter

The high/low model provides the same computational advantage as the on/off model: Rather than estimating N unconstrained variance parameters at each TF index, the system needs only classify the dominant source $n^*[t]$. Thus, the enhancement filter is given by $\mathbf{W}[t] = \mathbf{W}^{(n^*[t])}$ where

$$\mathbf{W}^{(n)} = \left(r_{\text{high},n} \mathbf{G}_n \mathbf{A}_n^H + \sum_{m \neq n} r_{\text{low},m} \mathbf{G}_m \mathbf{A}_m^H \right) \mathbf{R}_{\mathbf{x},n}^{-1} \tag{11}$$

and

$$\mathbf{R}_{\mathbf{x},n} = r_{\text{high},n} \mathbf{A}_n \mathbf{A}_n^H + \sum_{m \neq n} r_{\text{low},m} \mathbf{A}_m \mathbf{A}_m^H + \mathbf{R}_{\mathbf{z}} \tag{12}$$

for $n = 1, \ldots, N$. The time-varying filter switches between N fixed filters which can be computed in advance, so that the computational complexity of enhancement is no greater than it would be with the on/off model. The system can be used with any source activity classifier, including spatial methods such as DUET and data-driven methods such as neural networks.

C. Scaling with array size

The advantage of the high/low model over the on/off model is its scaling with array size. Let us consider several regimes for different array sizes. First, consider the single-microphone case with $M = 1$ and $\mathbf{A} = \mathbf{1}^T$. The filters become

$$\mathbf{W}^{(n)} = \mathbf{G}_n \frac{r_{\text{high},n}}{r_{\text{high},n} + \sum_{m \neq n} r_{\text{low},m} + r_z} + \sum_{m \neq n} \mathbf{G}_m \frac{r_{\text{low},m}}{r_{\text{high},n} + \sum_{\ell \neq n} r_{\text{low},\ell} + r_z} \tag{13}$$

for $n = 1, \ldots, N$. If the high-low ratio and signal-to-noise ratio are large, then $\hat{\mathbf{y}}[t] \approx \mathbf{G}_{n^*[t]} x[t]$, which is a conventional time-frequency mask. The filter changes dramatically between states, which might introduce distortion and artifacts, especially if the classifier makes errors or the signals do not obey the assumed model.

Next, suppose that there are multiple microphones, but not enough to perfectly separate the speech sources, either because $M < N$ or because $\mathbf{z}[t]$ is nonnegligible. For concreteness, consider a single-target enhancement system with $\mathbf{G} = [1, 0, \ldots, 0]$. At time-frequency indices with $n^*[t] = 1$, we can apply the Sherman-Morrison-Woodbury formula [22] to find

$$\mathbf{W}^{(1)} = r_{\text{high},1}[t] \mathbf{A}_1^H \mathbf{R}_{\mathbf{x},1}^{-1} \tag{14}$$

$$= \frac{\mathbf{A}_1^H \left(\sum_{m=2}^N r_{\text{low},m} \mathbf{A}_m \mathbf{A}_m^H + \mathbf{R}_{\mathbf{z}} \right)^{-1}}{r_{\text{high},1}^{-1} + \mathbf{A}_1^H \left(\sum_{m=2}^N r_{\text{low},m} \mathbf{A}_m \mathbf{A}_m^H + \mathbf{R}_{\mathbf{z}} \right)^{-1} \mathbf{A}_1}. \tag{15}$$

For large $r_{\text{high},1}$, $\mathbf{W}^{(1)}$ resembles a minimum-variance distortionless-response beamformer. Similar analysis can be applied when $n^*[t] \neq 1$ to show that the beamformer attempts to place a null over the dominant source while applying unity gain to the target signal $s_1[t]$. Thus, a classification error would not strongly affect the gain applied to the target. The filter accounts for every sound source in the mixture even though it cannot perfectly separate them.

Finally, consider the overdetermined case where $M > N$ and the noise power is negligible. If $\mathbf{A}$ has full column rank, then we can apply the Woodbury identity to the MWF (4) to find

$$\mathbf{W}[t] = \mathbf{G} (\mathbf{R}_{\mathbf{s}}^{-1}[t] + \mathbf{A}^H \mathbf{R}_{\mathbf{z}}^{-1} \mathbf{A})^{-1} \mathbf{A}^H \mathbf{R}_{\mathbf{z}}^{-1}. \tag{16}$$

If the speech-to-noise ratio is large—that is, in the limit as $r_n[t] \to \infty$ for $n = 1, \ldots, N$—the filter becomes

$$\mathbf{W}[t] = \mathbf{G} \left(\mathbf{A}^H \mathbf{R}_{\mathbf{z}}^{-1} \mathbf{A} \right)^{-1} \mathbf{A}^H \mathbf{R}_{\mathbf{z}}^{-1}, \tag{17}$$

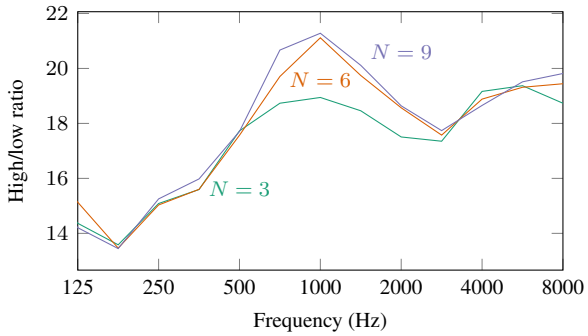

Fig. 3. Empirical high/low ratio for mixtures of different numbers of anechoic speech sources as a function of frequency.

TABLE I
EMPIRICAL HIGH/LOW RATIOS (DB)

Talkers	2	4	6	8	10
Anechoic	19	19	18	19	19
Small room ($T_{60} \approx 250$ ms)	20	19	20	20	20
Large room ($T_{60} \approx 780$ ms)	22	20	20	21	20

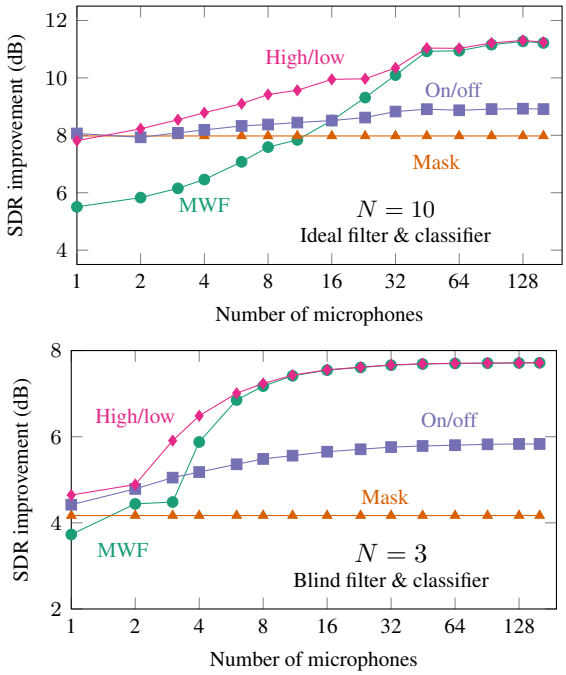

Fig. 4. Speech enhancement performance using arrays of different sizes. Top: Ideal filter and classifier separating ten sources. Bottom: Estimated filter and learning-based classifier separating three sources.

which is a linear time-invariant filter that separates the sources using a left inverse of $\mathbf{A}$. Notably, this filter does not depend on the relative source powers; instead, it fully separates the speech sources and uses any extra degrees of freedom to reduce noise. Because this overdetermined filter changes little between states, it does not depend on accurate source activity classification and does not produce the distortion and artifacts that plague many time-varying enhancement methods.

IV. Experiments

A. Empirical analysis of high/low ratios

The sparsity properties of speech signals in the STFT domain have been well studied. Speech energy is most concentrated with frame size around 60 ms [23]. The W-disjoint orthogonality assumption works well for three or four sources, but there is greater overlap for mixtures of many sources [5]. The accuracy of the high/low model and the ratio between states should similarly depend on the number of talkers.

To study high/low ratios with real speech signals, mixtures were generated using quasi-anechoic speech from the VCTK dataset [21]. Figure 3 shows the ratio between the mean dominant-source power and the mean non-dominant-source power for mixtures of different numbers of sources as a function of frequency with an STFT frame size of 64 ms. The ratio is generally larger at high frequencies, but only by a few decibels, and does not vary strongly with N. Table I shows the ratio averaged across frequencies for the same speech sources convolved with impulse responses recorded in different rooms. Despite the different acoustic conditions, the mean ratio varies little between rooms or with the number of sources.

B. Multimicrophone speech enhancement

To study the performance scaling of speech separation and enhancement systems using the high/low model, experiments were performed using real-room data with ten sound sources captured by 160 microphones in wearable and tabletop array devices throughout a large conference room [24]. Speech data from the VCTK dataset was convolved with 32 ms truncated room impulse responses and mixed with spatially uncorrelated speech-shaped noise. The mixtures were processed by several separation systems: a static MWF, a binary mask, a discrete-state multimicrophone filter using the on/off model (8), and a discrete-state filter using the high/low model (11) with a ratio of 15 dB. The desired response was a separating matrix $\mathbf{G} = \mathrm{diag}(A_{1,1}, \ldots, A_{1,N})$ using microphone 1 as a reference. The filter sequence was applied separately to the different signal components so that the contribution of each source to the output could be quantified.

Figure 4 shows performance measured using the average output signal-to-distortion-ratio (SDR):

$$\mathrm{SDR} = \frac{1}{N}\sum_{n=1}^{N} 10\log_{10}\frac{\sum_{t,f}|y_n[t,f]|^2}{\sum_{t,f}|\hat{y}_n[t,f]-y_n[t,f]|^2}. \quad (18)$$

The top panel shows results for an ideal filter separating all 10 sources. It was designed using measured transfer functions and a ground-truth classifier $n^*[t,f] = \arg\max_n |s_n[t,f]|^2$. The bottom panel shows results for a non-ideal filter separating 3 sources, averaged over 100 random combinations of sources and permutations of microphones. It was designed using transfer function estimates from the AuxIVA blind source separation algorithm [25], which was initialized using the nearest microphone to each source. To classify the dominant source at each TF index, the filter uses the Asteroid [26] imple-

mentation of the single-microphone deep clustering algorithm trained on the wsj0-3mix dataset [4]. Note that the model was trained under different acoustic conditions, so it makes relatively frequent errors; the average SDR improvement using the classifier alone was around 4 dB. On the other hand, because it does not use spatial features for classification, it is not affected by errors in the acoustic channel estimate.

When the number of microphones is smaller than the number of sources, the time-varying methods outperform the static MWF, which cannot separate all sources at once. The performance of the static filter improves with the number of microphones, while the performance of the binary mask—which does not use information from multiple microphones—does not. The spatial filter with the on/off model does improve with M because it performs a projection that reduces noise and interference, but it does not specifically target the interfering speech sources. The filter with the high/low model performs well for both small and large M because it can take advantage of both sparsity and spatial diversity. Notably, the high/low system matches the performance of the static MWF for large M even when using the error-prone learning-based classifier.

V. Conclusions

As large and distributed microphone arrays become widespread, there is a need for source separation and enhancement methods that can scale to take advantage of greater spatial diversity. The high/low model allows systems to take advantage of both spatial diversity from large arrays and time-frequency sparsity. It is a versatile model that can be applied to both small arrays, for which the enhancement system behaves like a mask, and large arrays, for which it behaves like a spatial filter. Because it can be used with any source activity classifier, it is a promising tool to incorporate model-based separation methods into multimicrophone systems.

References

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

[2] 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.

[3] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 2015.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[5] 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.

[6] T. Melia and S. Rickard, "Underdetermined blind source separation in echoic environments using desprit," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–19, 2006.

[7] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[8] 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.

[9] M. Kühne, R. Togneri, and S. Nordholm, "A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation," *Signal Processing*, vol. 90, no. 2, pp. 653–669, 2010.

[10] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 116–120.

[11] J. Rosca, C. Borss, and R. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2004, pp. iii–877.

[12] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," in *Independent Component Analysis and Blind Signal Separation (ICA)*. Springer Berlin/Heidelberg, 2004, pp. 652–660.

[13] M. Togami, T. Sumiyoshi, and A. Amano, "Sound source separation of overcomplete convolutive mixture using generalized sparseness," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.

[14] A. Aissa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani, and Y. Grenier, "Underdetermined blind separation of nondisjoint sources in the time-frequency domain," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 897–907, 2007.

[15] 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.

[16] 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.

[17] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for nonstationary noise environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, pp. 1–10, 2003.

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

[19] ——, "Speech separation using partially asynchronous microphone arrays without resampling," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[20] ——, "Binaural audio source remixing with microphone array listening devices," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[21] 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

[22] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Johns Hopkins University Press, 2013.

[23] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.

[24] R. M. Corey, M. D. Skarha, and A. C. Singer, "Massive distributed microphone array dataset," University of Illinois at Urbana-Champaign, 2019. [Online]. Available: https://doi.org/10.13012/B2IDB-6216881_V1

[25] 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.

[26] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Interspeech*, 2020.