UNDERDETERMINED METHODS FOR MULTICHANNEL AUDIO ENHANCEMENT WITH PARTIAL PRESERVATION OF BACKGROUND SOURCES

Ryan M. Corey and Andrew C. Singer

University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

ABSTRACT

Multichannel audio enhancement and source separation traditionally attempt to isolate a single source and remove all background noise. In listening enhancement applications, however, a portion of the background sources should be retained to preserve the listener's spatial awareness. We describe a time-varying spatial filter designed to apply a different gain to each sound source with minimal distortion of the source spectra and background spatial cues. The filter, inspired by methods from underdetermined source separation, adjusts its parameters at each time-frequency bin to preserve dominant sources. The time-varying filter enjoys greater flexibility than a fixed filter, while partial background preservation improves robustness to noise and errors.

Index Terms— Beamforming, audio source separation, speech enhancement, hearing aids, microphone arrays

1. INTRODUCTION

Wearable listening devices, such as hearing aids, are used to amplify desired sounds and attenuate unwanted noise. Devices with multiple microphones can use beamforming to distinguish between signals coming from different locations [1–3]. Sophisticated beamformers can be designed to cancel individual sources and reduce noise without distorting the signal of interest [4–6]. Unfortunately, hearing aid microphones are closely spaced and lack the spatial diversity to reliably separate more than a few simultaneous sources. Furthermore, if a beamformer is designed to isolate signals from a certain direction, then any remaining interference and noise at the output can appear to come from that same direction, giving incorrect localization cues to the listener [7].

Recently developed binaural beamforming methods [8–15] address both of these problems by retaining portions of background sounds. This residual interference preserves the listener's spatial awareness and relaxes interference rejection constraints on the beamforming filters. Early studies, which simply added a fraction of the unprocessed audio to the beamformer output, found that listeners benefited from the preserved cues [8,9]. More recent proposals aim to preserve the statistical coherence of diffuse noise [10,11]. To apply different gain levels to individual sources, devices can use linearly constrained minimum variance (LCMV) beamformers, which apply a distortionless constraint to each source and therefore preserve binaural cues [12, 15]. Unfortunately, binaural listening devices cannot realistically apply more than a few such constraints while still reducing background noise. Researchers have proposed an interaural transfer function constraint [13–15] that uses fewer degrees of freedom by constraining only the interaural transfer function, but such constraints allow arbitrary spectral distortion of the source signals.

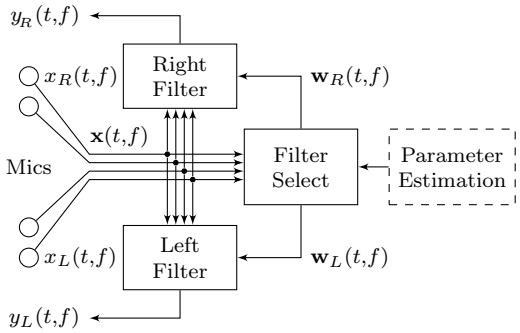

Figure 1: The proposed mutlichannel audio enhancement system. The left and right reference mics are near the corresponding outputs.

Most of these methods assume that the background sources are stationary; however, many real-world signals, such as speech, are nonstationary and sparse. Here, we propose a method that takes advantage of speech sparsity to further relax the constraints on binaural beamformers, allowing them to process more sources than they could with a stationary model. Our approach is inspired by underdetermined source separation [6, 16, 17]. In mixtures with more sources than sensors, these algorithms must rely on assumptions about the signal structure, such as sparsity. Since speech mixtures are approximately disjoint in the time-frequency (T-F) domain [18], mask-based methods assign each T-F bin to one source [19–21]. More general subset-based methods assume that a few sources are active in each bin and switch between several beamformers to separate them [22–26]. More computationally intensive methods, such as informed filtering [27, 28], expectation maximization [29, 30], and non-negative matrix factorization [31], incorporate statistical and spectral models to generate a new filter in each T-F bin.

We propose a spatial filtering system that uses sparsity assumptions to separate several sources from a mixture. Unlike traditional source separation systems, which seek to estimate a single source, the proposed system estimates a pair of modified mixtures with different gains applied to each source. The binaural system, shown in Figure 1, decides which sources are active in each T-F bin and applies strong gain constraints to them while applying weaker gain constraints to the inactive sources. Although the source separation model is relatively simple compared to the state of the art, we find that it improves performance for background-preserving binaural audio enhancement.

This work was supported in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number DGE-1144245.

2. PROPOSED METHOD

2.1. Definitions and signal model

Consider an array of $M \geq 2$ microphones with at least one microphone located in or near each ear. Let $\mathbf{x}(t,f) = [x_1(t,f),\ldots,x_M(t,f)]^T$ be the vector of microphone signals in the short-time Fourier transform (STFT) domain, where t is the frame index and f is the frequency bin. Two of its elements, designated $x_L(t,f)$ and $x_R(t,f)$, serve as reference microphones for the left and right ears. The left and right outputs are given by

$$y_L(t,f) = \mathbf{w}_L^H(t,f)\mathbf{x}(t,f) \quad (1)$$

$$y_R(t,f) = \mathbf{w}_R^H(t,f)\mathbf{x}(t,f), \quad (2)$$

where $\mathbf{w}_L$ and $\mathbf{w}_R$ are length-M vectors of time-varying complex coefficients. Note that each filter uses all M microphones.

Let $\mathbf{s}(t,f) = [s_1(t,f),\ldots,s_N(t,f)]^T$ be a vector of source signals. Let $\mathbf{z}(t,f) \in \mathbb{C}^M$ be a vector of undesired noise signals, including both diffuse environmental noise and sensor noise, received by the microphone array. In this work, we restrict our attention to the stationary narrowband mixing model,

$$\mathbf{x}(t,f) = \mathbf{A}(f)\mathbf{s}(t,f) + \mathbf{z}(t,f), \quad (3)$$

where $\mathbf{A}(f) \in \mathbb{C}^{M\times N}$ is the mixing matrix. Each element $a_{m,n}(f)$ is the acoustic transfer function between source n and microphone m. If the frame size is large compared to the length of the impulse response between source n and microphone m, then the acoustic transfer function characterizes the effects of delay, reverberation, and filtering on the source signal. The transfer functions for the reference microphones, $a_{L,n}(f)$ and $a_{R,n}(f)$, capture the spatial cues—interaural time and level differences and head-related acoustic filtering—that we wish to preserve in the output. For simplicity, we assume that the transfer functions are fixed and that an estimate of them is available. This assumption is reasonable if the spatial parameters change slowly compared to the source statistics.

Let $g_1(t,f),\ldots,g_N(t,f)$ be real-valued amplitude gains used to define the desired levels of the sources in the output. These gains may vary over frequency, such as to apply source-specific equalization, or over time, for example as part of a dynamic range compression system [32]. For each source n, the desired output component at the left ear is

$$d_{L,n}(t,f) = g_n(t,f)a_{L,n}(f)s_n(t,f). \quad (4)$$

The desired output signal is the sum of the individual components:

$$d_L(t,f) = \textstyle\sum_{n=1}^N g_n(t,f)a_{L,n}(f)s_n(t,f) \quad (5)$$

$$= \tilde{\mathbf{g}}_L^H(t,f)\mathbf{s}(t,f), \quad (6)$$

where $\tilde{\mathbf{g}}_L(t,f) = \left[g_1(t,f)a_{L,1}^*(f),\ldots,g_N(t,f)a_{L,N}^*(f)\right]^T$ is the complex gain vector incorporating spatial cues. The desired output at the right ear, $d_R(t,f)$, is defined equivalently.

2.2. Time-varying weighted filter

A number of statistical beamformers are commonly used in audio enhancement: the minimum variance distortionless response (MVDR) beamformer minimizes noise power subject to a distortionless constraint on a single target source; the LCMV beamformer applies constraints to multiple sources; and the multichannel Wiener filter (MWF) minimizes mean square error (MSE). We use a formulation known as the multiple speech distortion weighted MWF (MSDW-MWF) [6, 33], which minimizes a weighted sum of the MSE between $y_{L,n}(t,f)$ and $d_{L,n}(t,f)$ for $n = 1,\ldots,N$. The MSDW-MWF coefficients are given by

$$\mathbf{w}_L(t,f) = \left(\mathbf{A}(f)\Lambda(t,f)\mathbf{R}_s(f)\mathbf{A}^H(f) + \mathbf{R}_z(f)\right)^{-1} \cdot \mathbf{A}(f)\Lambda(t,f)\mathbf{R}_s(f)\tilde{\mathbf{g}}_L(t,f), \quad (7)$$

where $\mathbf{R}_s = \mathrm{Cov}(\mathbf{s})$ and $\mathbf{R}_z = \mathrm{Cov}(\mathbf{z})$ are the covariance matrices of the sources and noise and $\Lambda = \mathrm{diag}[\lambda_1,\ldots,\lambda_N]$ is the matrix of speech distortion weights. Each λ_n controls the trade-off between noise reduction and spectral distortion of source n. If $\Lambda = I$, the beamformer is a MWF that minimizes overall MSE between $y_L(t,f)$ and $d_L(t,f)$ for each f. In the limit as $\lambda_n \to \infty$, a distortionless constraint is applied to source n, forcing $\mathbf{w}_L^H(t,f)\mathbf{a}_n(f) = \tilde{g}_{L,n}^*(t,f)$. The right beamformer $\mathbf{w}_R$ is defined equivalently, with $\tilde{\mathbf{g}}_R$ replacing $\tilde{\mathbf{g}}_L$.

Since listening devices have relatively poor spatial diversity, a fixed beamformer is not sufficient to separate more than a few sources and we must use time-varying methods. Because speech is nonstationary, many time-varying source separation methods use different estimates of $\mathbf{R}_s$ and $\mathbf{R}_z$ in each T-F bin [30, 31]. Here, since we would like to estimate a distortion-constrained mixture of multiple active sources, we instead fix the covariance matrices, for example using long-term average source spectra, and vary the distortion weights $\lambda_n(t,f)$ in each bin. In [34], the distortion weight of a single speech source was varied based on the speech presence probability. For our multiple-source filtering problem, we let each of the N weights vary. Thus, the proposed time-varying weighted filter (TVWF) minimizes a cost function that includes the distortion of all N sources in every T-F bin, but prioritizes those sources that are most perceptually relevant in each bin.

Since multivariate probabilistic models are computationally expensive, we simplify the problem using the W-disjoint orthogonality assumption [18, 19], which states that one speech source dominates in each T-F bin:

$$|s_{n^*(t,f)}(t,f)|^2 \gg |s_n(t,f)|^2, \quad n \neq n^*(t,f). \quad (8)$$

This assumption, which is widely used in underdetermined source separation, allows us to select the weights using a computationally efficient classifier. Here, we use a maximum likelihood classifier to select the source whose steering vector best matches the observed signal [22, 35]:

$$\hat{n}(t,f) = \arg\min_n \left(\mathbf{x}(t,f) - \mathbf{P}_n(f)\mathbf{x}(t,f)\right)^H \mathbf{R}_z^{-1}(f) \cdot \left(\mathbf{x}(t,f) - \mathbf{P}_n(f)\mathbf{x}(t,f)\right), \quad (9)$$

where $\mathbf{P}_n(f) = \mathbf{a}_n(f)(\mathbf{a}_n^H(f)\mathbf{R}_z^{-1}(f)\mathbf{a}_n(f))^{-1}\mathbf{a}_n^H(f)\mathbf{R}_z^{-1}(f)$ is the whitened projection matrix onto $\mathbf{a}_n(f)$. In noisy environments, we add a state representing no active sources—that is, the noise is dominant—for a total of $N+1$ possible states. This model can be generalized to multiple simultaneously active sources [26], but our experiments suggest that there is little benefit for the small arrays typically encountered in listening devices.

The TVWF is given by (7) with distortion weights that take one of two values:

$$\lambda_n(t,f) = \begin{cases} \alpha_1, & \text{if } n = \hat{n}(t,f), \\ \alpha_0, & \text{otherwise.} \end{cases} \quad (10)$$

The value of α_1 controls the tradeoff between source distortion and noise reduction, like the traditional speech distortion weight. The weight for the inactive state, α_0, dampens the transients that cause musical noise in rapidly-varying filters such as binary masks. If $\alpha_0 = \alpha_1$, we have a conventional fixed MWF. If $\alpha_0 = 0$ and $\alpha_1 > 0$, we obtain a T-F mask. Because there are only $N+1$ possible filters for each bin, they can be computed offline and stored in memory rather than recomputed for every sample.

3. EXPERIMENTS

3.1. Experimental setup and metrics

We conducted experiments using a database of behind-the-ear hearing aid impulse responses recorded in an anechoic chamber and a university courtyard [36]. Our experiments use two microphones from each earpiece, for a total of $M = 4$. Up to eight sources are placed throughout the room surrounding the head. For each of 50 trials, the 20 second source signals are drawn at random from a set of spoken English sentences from the TIMIT corpus [37]. The STFT uses a raised cosine window of length 1024 samples (64 ms) and step size of 256 samples (16 ms). The diffuse background noise is approximately cylindrically isotropic with a stationary spectrum similar to that of the speech data.

We use the following metrics, defined here for the left output, to evaluate the performance of the beamformer. The index τ indicates time-domain rather than STFT-domain signals. The signal-to-interference-plus-noise ratio (SINR) measures the isolation of a single target source, assumed to be s_1:

$$\mathrm{SINR}_L = 10\log_{10}\frac{\sum_\tau y_{L,1}^2(\tau)}{\sum_\tau (y_L(\tau) - y_{L,1}(\tau))^2}, \tag{11}$$

where $y_{L,1}(\tau)$ is the output component due to the target source. The signal-to-distortion ratio (SDR) measures the error between each output component and its target value:

$$\mathrm{SDR}_{L,n} = 10\log_{10}\frac{\sum_\tau d_{L,n}^2(\tau)}{\sum_\tau (y_{L,n}(\tau) - d_{L,n}(\tau))^2}, \tag{12}$$

where $y_{L,n}(\tau)$ is the output component due to source n. We also consider the norm of the filter vectors, which has been shown to predict sensitivity to parameter estimation errors [38]:

$$\mathrm{Sensitivity}_L(f) = 10\log_{10}\mathbf{w_L}^H(f)\mathbf{w_L}(f). \tag{13}$$

For the TVWF, we report the maximum sensitivity over the $N+1$ filter vectors for each bin. The above metrics are defined equivalently for the right output. All experimental results are averaged over the left and right outputs.

To assess binaural cues, we measure the error in interaural phase difference (IPD) and interaural level difference (ILD):

$$\Delta\mathrm{IPD}_n(f) = \frac{\sum_t |d_{R,n}(t,f)|^2 \left| \angle\frac{d_{L,n}(t,f)}{d_{R,n}(t,f)} - \angle\frac{y_{L,n}(t,f)}{y_{R,n}(t,f)} \right|}{\sum_t |d_{R,n}(t,f)|^2} \tag{14}$$

$$\Delta\mathrm{ILD}_n(f) = \frac{\sum_t |d_{R,n}(t,f)|^2 \left| 20\log_{10}\left|\frac{d_{L,n}(t,f)}{d_{R,n}(t,f)}\right| - \left|\frac{y_{R,n}(t,f)}{y_{L,n}(t,f)}\right| \right|}{\sum_t |d_{R,n}(t,f)|^2}. \tag{15}$$

Our metrics differ from those in [9–11, 14, 15] in that we weight the errors by the signal power to avoid penalizing the filters for errors during periods of silence. Since those papers use stationary noise, weighting is not necessary. The reported IPD, ILD, and sensitivity results are averaged over frequency.

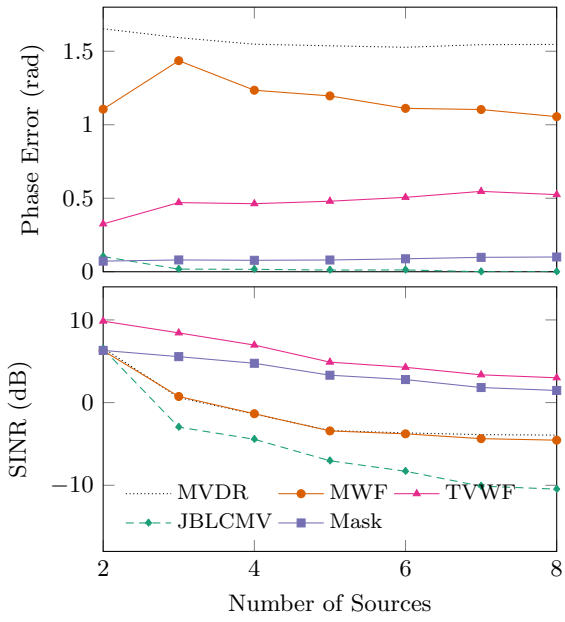

Figure 2: Comparison of enhancement methods for four-microphone binaural hearing aids in an anechoic chamber. The phase error is averaged over all background sources.

We compare the following spatial filtering methods: a conventional MVDR beamformer that applies a distortionless constraint to s_1 and suppresses other sources; the joint binaural LCMV (JBLCMV) beamformer of [13, 14], which constrains the binaural cues but not the spectral distortion of the background sources; a fixed background-preserving MWF, that is, (7) with $\Lambda = I$; a mask that determines the dominant source in each T-F bin and applies the corresponding gain to the reference microphone output,

$$y_L(t,f) = g_{\hat{n}(t,f)}(f)x_L(t,f); \tag{16}$$

and finally the proposed TVWF with $\alpha_0 = 10^{-1}$ and $\alpha_1 = 10^1$. We found these parameters to give consistent results without excessive distortion or musical noise.

3.2. Binaural cues

The first experiment evaluates noise reduction and binaural cue preservation in anechoic mixtures of varying numbers of speech sources and diffuse speech-shaped noise at a comparable level to the sources. The filters in this experiment were designed using the ground truth impulse responses and source and noise statistics used to generate the mixtures. The MWF, mask, and TVWF use target gains $\mathbf{g} = [1, 0.1, \ldots, 0.1]^T$.

Figure 2 shows average background source IPD error and target source SINR of the five filters as functions of N. The ILD error curves are not shown as they are similar in shape to the IPD curves. All methods achieved nearly zero IPD and ILD error for the target source. The JBLCMV is the most effective at preserving binaural cues: the errors, which are due to the narrowband approximation (3), actually decrease with N. The mask is similarly effective

	SINR	SDR-T	SDR-B	IPD-B	ILD-B	Sens.
MVDR	3.9	16.7	−5.9	1.61	8.96	7.8
JBLCMV	−0.4	15.4	−8.9	0.46	3.21	12.0
MWF	2.7	12.4	−0.8	0.86	5.08	3.7
Mask	5.9	13.8	3.2	**0.10**	**0.82**	**0.0**
TVWF	**6.8**	**17.6**	**5.5**	0.45	2.90	2.1
MVDR	−5.1	1.8	−14.1	1.58	8.32	13.6
JBLCMV	−8.4	0.9	−15.5	1.16	6.80	14.3
MWF	−3.9	3.7	−8.6	1.23	7.20	8.9
Mask	**5.1**	**12.9**	**1.9**	**0.11**	**0.93**	**0.0**
TVWF	3.8	8.6	−0.5	0.96	5.95	7.1

Table 1: Reverberant audio enhancement. The top filters are given the true mixing parameters. The bottom filters use an anechoic approximation. SDR-T is the SDR of the target source. SDR-B, IPD-B, and ILD-B are averaged over the background sources.

because it performs no spatial projection; its errors are transients caused by the time-varying gain. The MVDR is worst since it makes no attempt to preserve binaural cues. The background-preserving MWF does better, and the TVWF does better still. Since it projects onto the dominant source at each T-F point, most of the interaural distortion affects the inaudible sources.

The SINR results show that the time-varying filters are more effective than the fixed filters at attenuating background sources because the speech signals are nonstationary. They also have better noise reduction performance because they use fewer degrees of freedom for interference suppression. The JBLCMV beamformer has the worst SINR performance at large N because it applies $N+1$ constraints to its $2M=8$ coefficients. For $N \geq 7$ sources, it cannot filter its inputs at all without distorting binaural cues. Overall, Figure 2 shows the tradeoff between interference reduction, spectral distortion, and binaural cue distortion.

3.3. Reverberation and modeling error

The second experiment shows the robustness of the filtering methods to reverberation and modeling error. The mixtures were generated using the impulse responses for the courtyard, which has a reverberation time of $T_{60} = 900$ ms [36]. There were $N = 5$ speech sources with target gains $\mathbf{g} = [1.0, 0.2, \ldots, 0.2]^T$. To assess sensitivity to parameter mismatch, the filters were computed using both the ground truth impulse responses and erroneous impulse responses, which used the anechoic measurements from similar angles. The diffuse background noise was around 20 dB below the speech sources and the filters were computed with negligible diagonal loading to emphasize the effects of mismatch.

The performance results, shown in Table 1, illustrate the differing design criteria of the filters: the MVDR beamformer attempts to suppress interference without regard to binaural cues, the JBLCMV maintains those cues at the expense of SINR, and the MWF balances both. For both spatial models, the time-varying methods achieve better SINR and also distort the background sources less than the static filters. The mask performs nearly as well as the TVWF using the exact parameters and is more robust to mismatch, but it does produce more musical noise. A key advantage of background-preserving filters is their reduced sensitivity. The background-preserving MWF has sensitivity about 4 dB lower than the background-suppressing MVDR. The time-varying filters have even lower sensitivity since they apply fewer constraints at any given time. The SINR results show that they are less impacted by mismatch than the fixed beamformers. The SDR metrics are particularly sensitive to mismatch because the erroneous transfer functions do not account for reverberation; the IPD and ILD, which depend primarily on the direct acoustic path, are more robust to error.

4. CONCLUSIONS

Noisy environments with multiple simultaneous speakers are challenging for both human listeners and source separation algorithms. By retaining a portion of the background sources, a listening device can preserve the listener's spatial awareness and also improve the robustness of the spatial filter to noise and mismatch. Processing multiple sources with just a few microphones is challenging for conventional beamformers. Thanks to the sparsity of speech, however, time-varying filters can apply different gain values to many simultaneous speech signals while using relatively few constraints. The extra degrees of freedom can be used to suppress unwanted noise, minimize spectral and binaural cue distortion, and protect against parameter estimation errors.

In this paper, we have shown that a relatively simple system using the narrowband model, W-disjoint orthogonality assumption, and binary distortion weights can outperform a static beamformer in a challenging scenario with many speakers. The performance could be improved by applying state-of-the-art spatial and spectral models and more flexible distortion weights. We did not address source localization and tracking or acoustic parameter estimation, which are particularly challenging problems for wearable listening devices. As listening device technology and source separation methods improve, they will enable more sophisticated spatial filters that can separate and recombine several sources at once with minimal spectral and spatial distortion. Such filters could dynamically alter the sound scene to improve intelligibility and comfort while remaining transparent to the listener.

5. REFERENCES

[1] J. M. Kates, *Digital hearing aids*. Plural publishing, 2008.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," *Handb. array processing sensor networks*, pp. 269–302, 2008.

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

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE AASP magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[5] H. L. Van Trees, *Optimum array processing*. Wiley, 2004.

[6] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[7] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, 2006.

[8] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 360–371, 2009.

[9] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 342–355, 2010.

[10] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 23, no. 12, pp. 2162–2176, 2015.

[11] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van de Par, "Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 12, 2016.

[12] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 24, no. 3, pp. 543–558, 2016.

[13] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 23, no. 12, pp. 2449–2464, 2015.

[14] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, "Improved multi-microphone noise reduction preserving binaural cues," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 460–464, 2016.

[15] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 137–152, 2017.

[16] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007.

[17] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "Convolutive blind source separation methods," in *Springer Handbook of Speech Processing*, pp. 1065–1094, Springer, 2008.

[18] S. Rickard and Ö. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *IEEE Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 529–532, 2002.

[19] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, 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 Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

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

[22] J. Rosca, C. Borss, and R. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 3, pp. iii–877, 2004.

[23] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," *Indep. Compon. Analysis Blind. Signal Sep.*, pp. 652–660, 2004.

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

[25] 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 Trans. Signal Process.*, vol. 55, no. 3, pp. 897–907, 2007.

[26] R. M. Corey and A. C. Singer, "Nonstationary source separation for underdetermined speech mixtures," in *Asilomar Conf. Signals, Syst., Comp.*, pp. 934–938, 2016.

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

[28] M. Taseska and E. A. Habets, "Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1291–1304, 2016.

[29] M. A. Dmour and M. Davies, "A new framework for underdetermined speech extraction using mixture of beamformers," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 445–457, 2011.

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

[31] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[32] R. M. Corey and A. C. Singer, "Dynamic range compression for noisy mixtures using source separation and beamforming," in *IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2017.

[33] S. Markovich-Golan, S. Gannot, and I. Cohen, "A weighted multichannel Wiener filter for multiple sources scenarios," in *IEEE Conv. of Electrical & Electronics Eng. in Israel*, 2012.

[34] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Incorporating the conditional speech presence probability in multi-channel Wiener filter based noise reduction in hearing aids," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, p. 930625, 2009.

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

[36] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. Adv. Signal Process.*, vol. 2009, p. 6, 2009.

[37] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1." Web Download, 1993.

[38] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, 1987.