MOTION-TOLERANT BEAMFORMING WITH DEFORMABLE MICROPHONE ARRAYS

Ryan M. Corey and Andrew C. Singer

University of Illinois at Urbana-Champaign

ABSTRACT

Microphone arrays are usually assumed to have rigid geometries: the microphones may move with respect to the sound field but remain fixed relative to each other. However, many useful arrays, such as those in wearable devices, have sensors that can move relative to each other. We compare two approaches to beamforming with deformable microphone arrays: first, by explicitly tracking the geometry of the array as it changes over time, and second, by designing a time-invariant beamformer based on the second-order statistics of the moving array. The time-invariant approach is shown to be appropriate when the motion of the array is small relative to the acoustic wavelengths of interest. The performance of the proposed beamforming system is demonstrated using a wearable microphone array on a moving human listener in a cocktail-party scenario.

Index Terms— Microphone arrays, array processing, audio enhancement, hearing aids, wearables

1. INTRODUCTION

Microphone arrays can be used to spatially localize and separate sound sources from different directions [1–4]. Small arrays, typically with up to eight microphones spaced a few centimeters apart, are widely used in teleconferencing and speech recognition. A promising application is in hearing aids and other augmented listening devices [5], where arrays could improve intelligibility in noisy environments. However, the arrays in listening devices are tiny: typically only two microphones a few millimeters apart.

Arrays with microphones spread across the body can can perform better than listening devices with only a few microphones near the ears [6]. There is a major challenge in using such arrays, however: humans move. The microphones in a wearable array not only move relative to sound sources, *but also move relative to each other*, as shown in Figure 1. Because array processing typically relies on phase differences between sensors, even small deformations can harm the performance of a spatial sound capture system.

There has been little prior work on deformable microphone arrays. In [7], a robot with microphones on movable arms was adaptively repositioned to improve beamforming performance. In [8], microphones were placed along a hose-shaped robot and used to estimate its posture. In [9], wearable arrays were placed on three human listeners in a cocktail party scenario and aggregated using a sparsity-based time-varying filter. That paper applied the full-rank covariance model for deformation that is presented here.

In contrast, the problem of tracking moving *sources* has received significant attention. Most solutions combine a localization method, such as steered response power or multiple signal classification, with a tracking algorithm, such as Kalman or particle filtering [10–15]. Others use blind source separation techniques that adapt over time as the sources move [16, 17]. Sparse signal models can improve performance when there are multiple competing sound sources [9,18–21]. These time-varying methods are necessary when the motion of the sources or microphones is large. However, tracking algorithms are computationally complex and time-varying filters can introduce disturbing artifacts. For small motion, such as breathing or nodding with a wearable array, it may be possible to account for motion using a linear time-invariant filter instead.

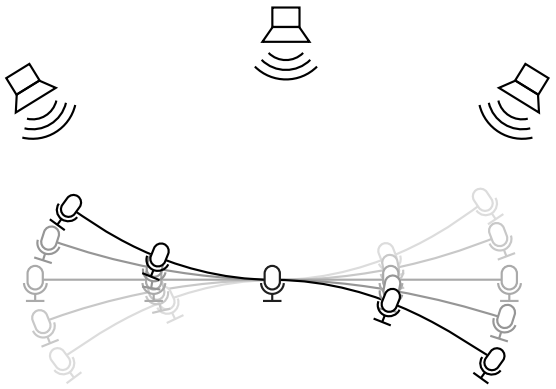

Figure 1: In a deformable microphone array, the sensors can move relative to the sound sources and also relative to each other.

The design of spatial filters that are robust to small perturbations is well studied. Mismatch between the true and assumed positions of the sensors can be modeled as uncorrelated noise and addressed using diagonal loading on the noise covariance matrix or using a norm constraint on the beamformer coefficient vector [22]. Other approaches include derivative constraints that ensure the beam pattern does not change too quickly [23] and distortion constraints within a region or subspace [24]. For far-field beamformers, these methods widen the beam pattern and therefore reduce array gain compared to non-robust beamformers.

In this work, we explore the impact of deformation on the performance of multimicrophone audio enhancement systems. If motion is small enough that it can be effectively modeled using second-order statistics, then the signals can be separated using linear time-invariant filters. Larger motion destroys the spatial correlation structure of the sources and therefore requires more complex time-varying methods. We compare the performance of different beamforming strategies on two deformable arrays: a linear array of microphones hanging from a pole, the motion of which is straightforward to model, and a wearable array on a human listener with more complex movement patterns. We find that the effects of deformation are dramatic at high frequencies but manageable at the low frequencies for which large arrays have the greatest benefit.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1144245.

2. TIME-FREQUENCY BEAMFORMING

Let $\mathbf{X}[t,f] = [X_1[t,f], X_2[t,f], \ldots, X_M[t,f]]^T$ be the vector of short-time Fourier transforms (STFT) of the signals captured at microphones 1 through M, where t is a time index and f is a frequency index. Assuming linear mixing, the received signal can be modeled as the sum of components $\mathbf{C}_1[t,f], \ldots, \mathbf{C}_N[t,f]$ due to N sources and diffuse additive noise $\mathbf{V}[t,f]$:

$$\mathbf{X}[t,f] = \sum_{n=1}^{N} \mathbf{C}_n[t,f] + \mathbf{V}[t,f]. \tag{1}$$

The components $\mathbf{C}_1, \ldots, \mathbf{C}_N$ are sometimes called source spatial images [25]. Assume that the source images and noise are zero-mean random processes that are uncorrelated with each other and that the diffuse noise is wide-sense stationary. Let $\mathbf{R}_n[t,f] = \mathbb{E}\left[\mathbf{C}_n[t,f]\mathbf{C}_n^H[t,f]\right]$ be the time-varying STFT covariance matrix of source image n for $n = 1, \ldots, N$, where $\mathbb{E}$ denotes expectation, and let $\mathbf{R}_v[f]$ be the time-invariant covariance of $\mathbf{V}[t,f]$.

The output $\mathbf{Y}[t,f] = \mathbf{W}[t,f]\mathbf{X}[t,f]$ of the audio enhancement system is a linear transformation of the microphone input signals in the time-frequency domain. The beamforming weights $\mathbf{W}[t,f]$ may vary over time and may produce one or several outputs. In this work, we restrict our attention to the multichannel Wiener filter (MWF) [2], which minimizes mean squared error between the output and a desired signal $\mathbf{D}[t,f]$:

$$\mathbf{W}[t,f] = \mathrm{Cov}\left(\mathbf{D}[t,f], \mathbf{X}[t,f]\right) \mathrm{Cov}\left(\mathbf{X}[t,f]\right)^{-1}. \tag{2}$$

Here we choose $\mathbf{D}[t,f] = \left[\mathbf{e}_1^T\mathbf{C}_1[t,f], \ldots, \mathbf{e}_1^T\mathbf{C}_N[t,f]\right]^T$ where $\mathbf{e}_1 = [1,0,\ldots,0]^T$; that is, we estimate each source signal as observed at microphone 1. In a listening device, this reference microphone might be the one nearest the ear canal so that head-related acoustic effects are preserved [26]. The MWF beamforming weights are given by

$$\mathbf{W}[t,f] = \begin{bmatrix} \mathbf{e}_1^T\mathbf{R}_1[t,f] \\ \vdots \\ \mathbf{e}_1^T\mathbf{R}_N[t,f] \end{bmatrix} \left(\sum_{n=1}^{N} \mathbf{R}_n[t,f] + \mathbf{R}_v[f]\right)^{-1}. \tag{3}$$

2.1. Statistical models

Many audio source separation and enhancement methods [3, 4] use time-varying STFT beamformers similar to (3). Time-varying covariance matrices capture the nonstationarity of natural signals such as speech and adapt to source and microphone movement. Because the focus of this paper is on the spatial separability of sound sources with deformable arrays, we will ignore the temporal statistics of the sound sources. Any variation of $\mathbf{R}_n[t,f]$ with respect to t is assumed to be due to motion of the microphones.

Let $\tilde{\mathbf{R}}_n[f;\theta]$ be the source covariance matrix corresponding to state $\theta \in \mathcal{X}$ for $n = 1, \ldots, N$, where $\mathcal{X}$ is a set of states that represent the positions and orientations of the microphones. Assume that the motion of the array is slow enough that each frame has a single corresponding state $\Theta[t]$ and that the effects of Doppler can be neglected. Then the sequence of covariance matrices is $\mathbf{R}_n[t,f] = \tilde{\mathbf{R}}_n[f;\Theta[t]]$ for $n = 1, \ldots, N$.

While it is often assumed that each $\mathbf{R}_n$ is a rank-one matrix proportional to the outer product of a steering vector, here we adopt the full-rank STFT covariance model [27]. Although originally developed to compensate for long impulse responses, the full-rank model is also useful for modeling uncertainty due to deformation.

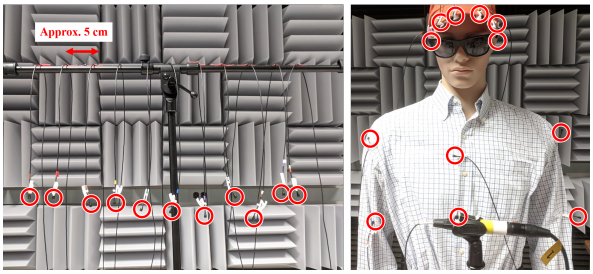

Figure 2: Deformable linear array (left) and wearable array (right).

2.2. Static and dynamic beamformers

This work will compare the performance of two separation methods, one static and one dynamic. For the static method, assume a prior distribution p_θ on θ. Because $\mathbf{C}_n[t,f]$ is assumed to have zero mean, the ensemble covariance matrices $\bar{\mathbf{R}}_n[f]$ are given by

$$\bar{\mathbf{R}}_n[f] = \mathbb{E}\left[\mathrm{Cov}\left(\mathbf{C}_n \mid \Theta\right)\right] = \int_{\mathcal{X}} p_\theta(\theta)\tilde{\mathbf{R}}_n[f;\theta]\,\mathrm{d}\theta, \tag{4}$$

for $n = 1, \ldots, N$. The static beamformer is computed by substituting $\bar{\mathbf{R}}_n[f]$ for $\mathbf{R}_n[t,f]$ in (3). In the static beamforming experiments presented here, the states are never explicitly defined. Instead, each $\bar{\mathbf{R}}_n[f]$ is estimated by the sample covariance over a set of training data. This is equivalent to an empirical measure over Θ.

For the dynamic method, assume that an estimate $\hat{\Theta}[t]$ of the state sequence is available, for example from a tracking algorithm. Then the estimated covariance matrices are

$$\hat{\mathbf{R}}_n[t,f] = \tilde{\mathbf{R}}_n[f;\hat{\Theta}[t]], \quad n = 1, \ldots, N. \tag{5}$$

In the results presented here, the set of states is manually determined for each experiment based on the range of motion of the array. For example, the linear array has discrete states representing different angles of rotation. To ensure that the results are as general as possible, we do not use a blind state estimation or tracking algorithm. Instead, we measure the states using near-ultrasonic pilot signals that are played back alongside the source speech signals. The source statistics within each discrete state are estimated by the sample covariance of the training data for time frames in that state.

3. SECOND-ORDER STATISTICS

Because the MWF depends on the second-order statistics of the observed signals, it will be instructive to analyze the effects of deformation on the covariance structure of the acoustic source images.

Since the source images are assumed to have full rank, they do not occupy different subspaces and the separability of different sources must be analyzed statistically. For example, the Kullback-Leibler divergence between two zero-mean multivariate Gaussian distributions with covariances $\mathbf{R}_1$ and $\mathbf{R}_2$ is [28]

$$D(\mathbf{R}_1, \mathbf{R}_2) = \frac{1}{2}\left[\mathrm{trace}\left(\mathbf{R}_1\mathbf{R}_2^{-1} - \boldsymbol{I}\right) - \ln\frac{\det\mathbf{R}_1}{\det\mathbf{R}_2}\right]. \tag{6}$$

This quantity is largest for pairs of matrices whose principal eigenvectors are orthogonal and zero for identical matrices. Although the signals captured by deformable arrays do not have Gaussian distributions, the divergence expression (6) will be useful in quantifying the impact of deformation on their second-order statistics.

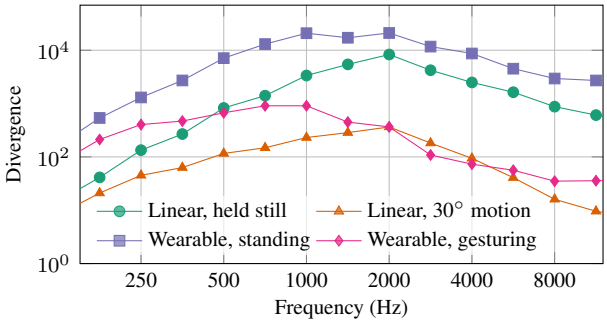

Figure 3: Average divergence between source covariance matrices.

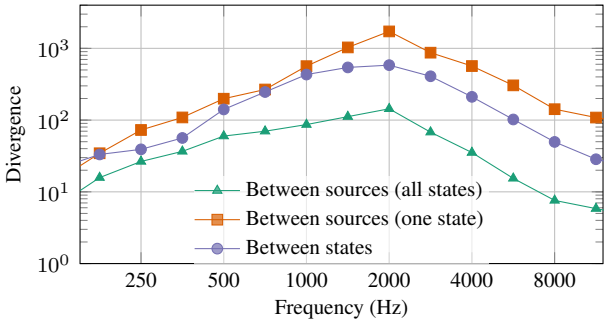

Figure 4: Divergence between sources and states for the hanging linear array. The between-source curves are the average divergence of the outer four sources with respect to the central source. The between-state curve is for the central source with the array at opposite ends of its range of motion, about 90° apart.

3.1. Ideal far-field array

Consider an array of ideal isotropic sensors observing N far-field sources from different angles. Suppose that the sources all have power spectral density $\sigma_n^2[f] = 1$. Then the STFT covariance matrices are $\mathbf{R}_n[f] = \mathbf{a}_n[f]\mathbf{a}_n^H[f]$ for $n = 1, \ldots, N$, where $\mathbf{a}_n[f]$ is a steering vector with $a_{n,m}[f] = e^{j\Omega_f \tau_{n,m}}$ for $m = 1, \ldots, M$, Ω_f is the continuous-time frequency corresponding to frequency index f, and $\tau_{n,m}$ is time delay of arrival for source n at microphone m.

Now suppose that the positions of the microphones are randomly perturbed so that $a_{n,m}[f] = e^{j\Omega_f(\tau_{n,m}+\Delta_{n,m})}$. If $\Delta_{n,m}$ have independent Gaussian distributions with zero mean and variance σ^2, then the off-diagonal elements of the ensemble average covariance matrices are attenuated:

$$\bar{\mathbf{R}}_{n,m_1,m_2}[f] = \mathbb{E}\left[e^{j\Omega_f(\tau_{n,m_1}-\tau_{n,m_2}+\Delta_{n,m_1}-\Delta_{n,m_2})}\right] \quad (7)$$

$$= \mathbf{R}_{n,m_1,m_2}[f]\mathbb{E}\left[e^{j\Omega_f(\Delta_{n,m_1}-\Delta_{n,m_2})}\right] \quad (8)$$

$$= \mathbf{R}_{n,m_1,m_2} e^{-\Omega_f^2\sigma^2}, \quad (9)$$

where the last step comes from the moment-generating function. Because all off-diagonal elements are scaled equally, we have

$$\bar{\mathbf{R}}_n[f] = e^{-\Omega_f^2\sigma^2}\mathbf{R}_n[f] + (1 - e^{-\Omega_f^2\sigma^2})\boldsymbol{I}. \quad (10)$$

Substituting (10) into (6) and applying the Sherman-Morrison formula, it can be shown that the Gaussian divergence between two source covariance matrices with these Gaussian random offsets is

$$D(\bar{\mathbf{R}}_1[f], \bar{\mathbf{R}}_2[f]) = \frac{M^2 - \left|\mathbf{a}_1^H[f]\mathbf{a}_2[f]\right|^2}{2(e^{\Omega_f^2\sigma^2} - 1)(e^{\Omega_f^2\sigma^2} - 1 + M)}. \quad (11)$$

From this expression, the second-order statistics of the two sources become more similar to each other as their unperturbed steering vectors become closer together, as the uncertainty due to motion increases, and as the frequency increases. Motion should have little impact if $\Omega_f\sigma$ is small, that is, if the scale of the motion is small compared to a wavelength. At high audible frequencies, where acoustic wavelengths might be just a few centimeters, deformable arrays will be quite sensitive to motion.

3.2. Experimental measurements

The derivation above assumed independent motion of all microphones. To confirm the predicted trends—that spatial diversity decreases with frequency and with amount of deformation—for real arrays with more complex deformation patterns, the second-order statistics of several deformable arrays were measured. Sample STFT covariance matrices were computed using 20-second pseudorandom noise signals produced sequentially by $N = 5$ loudspeakers about 45° apart in a half-circle around arrays of $M = 12$ omnidirectional lavalier microphones. One set of experiments used a linear array of microphones hanging on cables from a pole that was manually rotated in a horizontal plane. The hanging microphones swung by several millimeters relative to each other as they were moved. A second array used microphones affixed to a hat and near the ears, chest, shoulders, and elbows of a human subject who moved in different patterns. The arrays are shown in Figure 2.

Figure 3 shows the mean Gaussian divergence between the long-term average STFT covariance matrices of the central source and the four other sources for different array and motion types. The nonmoving wearable array provides the greatest spatial diversity between sources. The moving linear array provides the least. For both arrays, motion causes the greatest penalty at higher frequencies, as predicted.

With large deformations, it is difficult to distinguish the two sources based on their long-term average statistics and it would be helpful to use a time-varying model. Figure 4 shows the divergence between ensemble average covariances of two sources over all states, $D(\bar{\mathbf{R}}_1[f], \bar{\mathbf{R}}_2[f])$; the divergence between their covariances in a single state, $D(\hat{\mathbf{R}}_1[f;\theta_1], \hat{\mathbf{R}}_2[f;\theta_1])$; and the divergence between two different states for the same source, $D(\hat{\mathbf{R}}_1[f;\theta_1], \hat{\mathbf{R}}_1[f;\theta_2])$. At high frequencies, the two states are more different from each other than the two sources are on average, suggesting that the ensemble covariance would not be useful for separation. The divergence between sources is an order of magnitude larger within a single state than in the ensemble average.

4. STATIC AND DYNAMIC BEAMFORMING

To demonstrate the impact of deformation on audio enhancement, the two arrays were used to separate mixtures of speech sources using static and dynamic beamformers. For each experiment, the STFT covariance matrices were estimated using 20 seconds of pseudorandom noise played sequentially from each loudspeaker while the array was moved. The source signals are five 20-second anechoic speech clips from different talkers in the VCTK corpus [29].

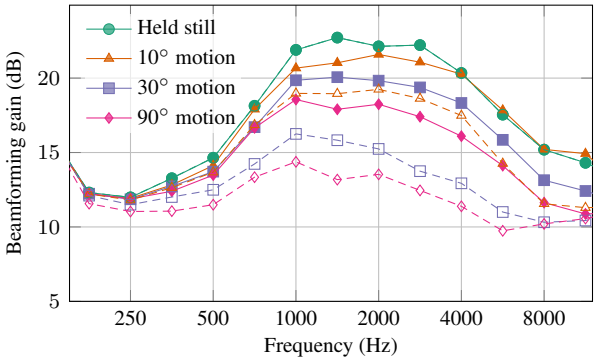

Figure 5: Beamforming performance with a linear array of dangling microphones. Solid curves show dynamic beamforming and dashed curves show static beamforming.

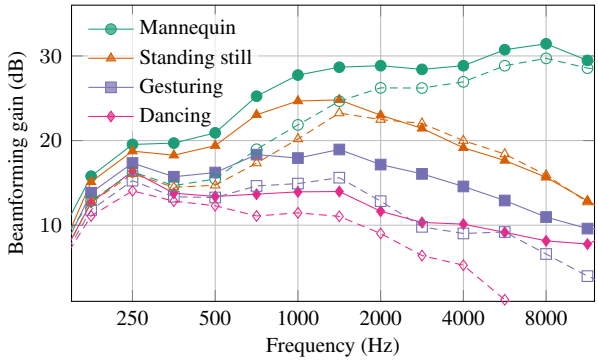

Figure 6: Beamforming performance with a wearable microphone array. Solid curves show full-rank beamformers dashed curves show rank-one beamformers.

The motion patterns produced by the human subject were similar but not identical between the training and test signals.

Speech enhancement performance is measured using the mean improvement in squared error between the input and output:

$$\text{Gain}[f] = \frac{1}{N}\sum_{n=1}^{N} 10\log 10 \frac{\sum_t |X_1[t,f] - D_n[t,f]|^2}{\sum_t |Y_n[t,f] - D_n[t,f]|^2}. \quad (12)$$

Normally, the ground truth signals $D_n[t,f]$ could be measured by recording each source signal in isolation. However, because the motion patterns cannot be exactly reproduced between experiments, it is impossible to know the ground truth signals received by a moving array. To provide quantitative performance measurements, the deformable arrays were supplemented by a nonmoving microphone used as the reference ($m = 1$). To qualitatively evaluate a fully deformable array, the wearable-array experiments were repeated without the fixed microphone using the two microphones near the ears as references; audio clips of these binaural beamformer outputs are available on the first author's website[1].

4.1. Dynamic beamforming with a linear array

The rotating linear array is well suited to dynamic beamforming because its state can be roughly described by its angle of rotation, which is easily measured using near-ultrasonic pilot signals. In this experiment, the states formed a discrete set of about ten positions. Note that there is still some uncertainty within each state because the microphones are allowed to swing freely. Figure 5 shows the average beamforming gain achieved by the linear array with different ranges of motion. Even small motion from being held steady in the experimenter's hand causes poor high-frequency performance. With 10° rotation, the static beamformer performs a few decibels worse than the dynamic motion-tracking beamformer. Dynamic beamforming is necessary for large motion because the angle of rotation is larger than the angular spacing between sources.

4.2. Static beamforming with a wearable array

The wearable array is more difficult to track dynamically because there are many degrees of freedom in human motion. Figure 6 compares the performance of two static beamformers: one designed from the full-rank average covariance matrix, and one designed using a rank-one covariance matrix, that is, using an acoustic transfer function measured from the training signals. For comparison with a truly nonmoving subject, the microphones were placed on a plastic mannequin in the same configuration as on the human subject. This motionless array performed well at the highest tested frequencies. The human subject, even when trying to stand still, moved enough to destroy the phase coherence between microphones at several kilohertz. These results suggest that researchers should use caution when testing arrays on mannequins because high-frequency performance might be different with live humans.

The full-rank covariance model outperforms the rank-one model even for the motionless array at low frequencies. It improves robustness against both motion and diffuse background noise. When the subject is gesturing—turning his head, nodding, and lifting and lowering his arms—or dancing in place by moving his arms, head, and torso, the full-rank beamformer outperforms the rank-one beamformer by several decibels at all frequencies. However, at the highest tested frequencies, the moving-array beamformers perform little better than a single-channel Wiener filter, which would provide about 8 dB gain for this five-source mixture.

5. CONCLUSIONS

The results presented here suggest that deformable microphone arrays perform poorly at high frequencies. The full-rank spatial covariance model can improve performance by several decibels compared to a rank-one model, and dynamic beamforming that tracks the state of the array provides even greater benefit. Even so, it seems that deformable microphone arrays, including wearables, are most useful at low and mid-range frequencies. Fortunately, these are the frequencies most important for speech perception.

Deformable arrays are advantageous because they can spread microphones across multiple devices or body parts. Thus, an array might combine rigidly-connected, closely-spaced microphones for high frequencies with deformable, widely-spaced microphones for low frequencies. Furthermore, as shown in [9], the full-rank covariance model can be used in nonlinear, time-varying methods that aggregate data from multiple wearable arrays. Large deformable arrays can provide greater spatial diversity than small rigid arrays and could be an important tool in spatial sound capture applications.

[1] ryanmcorey.com/demos

6. REFERENCES

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

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

[3] S. Makino, ed., *Audio Source Separation*. Springer, 2018.

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

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

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

[7] H. Barfuss and W. Kellermann, "An adaptive microphone array topology for target signal extraction with humanoid robots," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 16–20, 2014.

[8] Y. Bando, T. Mizumoto, K. Itoyama, K. Nakadai, and H. G. Okuno, "Posture estimation of hose-shaped robot using microphone array localization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3446–3451, 2013.

[9] 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.

[10] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 3021–3024, 2001.

[11] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[12] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[13] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with RANSAC and directional statistics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2233–2243, 2014.

[14] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 8, pp. 1408–1423, 2016.

[15] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 281–295, 2018.

[16] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Robust real-time blind source separation for moving speakers in a room," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[17] J. Málek, Z. Koldovskỳ, and P. Tichavskỳ, "Semi-blind source separation based on ICA and overlapped speech detection," in *International Conference on Latent Variable Analysis and Signal Separation (LVA ICA)*, pp. 462–469, 2012.

[18] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

[19] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying Rao-Blackwellised Monte Carlo data association," in *IEEE Workshop on Statistical Signal Processing*, pp. 253–256, 2009.

[20] S. M. Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 201–204, 2010.

[21] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based ONDOA-HMM," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3191–3195, 2014.

[22] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[23] M. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 6, pp. 1378–1393, 1983.

[24] Y. R. Zheng, R. A. Goubran, and M. El-Tanany, "Robust near-field adaptive beamforming with distance discrimination," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 478–488, 2004.

[25] 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.

[26] 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 *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.

[27] 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.

[28] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*. Springer, 2008.

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