BINAURAL AUDIO SOURCE REMIXING WITH MICROPHONE ARRAY LISTENING DEVICES

Ryan M. Corey and Andrew C. Singer

University of Illinois at Urbana-Champaign

ABSTRACT

Augmented listening devices, such as hearing aids and augmented reality headsets, enhance human perception by changing the sounds that we hear. Microphone arrays can improve the performance of listening systems in noisy environments, but most array-based listening systems are designed to isolate a single sound source from a mixture. This work considers a source-remixing filter that alters the relative level of each source independently. Remixing rather than separating sounds can help to improve perceptual transparency: it causes less distortion to the signal spectrum and especially to the interaural cues that humans use to localize sounds in space.

***Index Terms*—** Microphone array processing, hearing aids, augmented reality, beamforming, audio source separation

1. INTRODUCTION

Audio signal processing can be used to change the way that humans experience the world around them. Augmented listening (AL) devices, such as hearing aids and augmented reality headsets [1], enhance human hearing by altering the sounds presented to a listener. One of the most important functions of listening devices is to help humans to hear better in noisy environments with many competing sound sources. Unfortunately, many listening devices today perform poorly in noisy environments where users need them most.

One reliable way to reduce unwanted noise and improve intelligibility in noisy situations is with microphone array processing. Arrays of several spatially separated microphones can be used to perform beamforming and isolate sounds from a particular direction [2]. Beamformers are widely used in machine listening systems, for example for automatic speech recognition, to preprocess a sound source of interest. For decades, researchers have tried to incorporate similar beamforming technology into human listening devices [3–6]. Most past studies have taken the same approach used in machine listening: steering a beam to isolate a single sound source of interest and attenuate all others. Although this approach has been shown to improve intelligibility in noise [3, 4], listening devices using large arrays have never been commercially successful.

The human auditory system is different from a machine listening algorithm. Isolating a single sound source, while it might improve intelligibility, would seem unnatural to the listener and it could interfere with the auditory system's natural source separation and scene analysis capabilities. Humans rely on spectral and temporal patterns as well as spatial cues, such as interaural time and level differences, to distinguish between sound sources and remain aware of their environment [7, 8]. Single-target beamformers can distort the spectral and spatial cues of all non-target sound sources [9]. Beamformers can be designed to preserve the interaural cues of multiple sources by deliberately including components of non-target sources in the filter output [10–12]. Early binaural beamformers simply added a fraction of the unprocessed signal into the output [13–15], while more recent methods apply explicit constraints to the filter's response to those signals [16, 17]. It is also possible to constrain only the interaural cues and not the spectral distortion of the sources [18, 19]. In general, the more the background sources are preserved, the less their interaural cues are distorted [15].

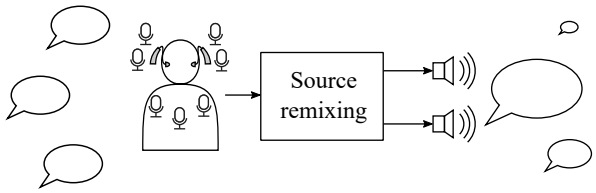

Fig. 1: The proposed source-remixing system uses a microphone array and space-time filters to apply separate processing to many sound sources, altering the auditory perception of the listener.

In this work, we approach augmented listening as a *source remixing* problem. Rather than trying to isolate a single sound source, we use a microphone array to apply different processing to many sound sources, as shown in Fig. 1. The system is analogous to the mixing process in a music or film studio, where each instrument, dialogue track, and sound effect in the mixture is individually adjusted to provide a pleasant listening experience. The processing applied to each source depends on the type of sound, the acoustic environment, and the listener's preferences. For example, a normal-hearing user in a quiet room might prefer no processing at all, while a hearing-impaired user in a noisy restaurant could use aggressive noise reduction similar to single-target beamforming. A similar remixing approach has been proposed for television broadcasts with object-based audio encodings: listeners can tune the relative levels of sources to trade off between immersiveness and intelligibility [20, 21]. Using a powerful microphone array, we could allow listeners to perform that same tuning for real-world sounds.

In this work, we consider the choice of relative levels of sound sources in a source-remixing filter. Further perceptual research and clinical studies will be needed to understand how to select these levels to optimize the listening experience for different AL applications. In the meantime, we can characterize the engineering tradeoffs of such a system. Intuitively, the less we try to separate the sound sources—that is, the more transparent the listening experience—the easier it should be to implement the remixing system and the more

This material is based upon work supported by the National Science Foundation under Grant No. 1919257.

natural it should sound to the listener. Studies of musical source separation have shown that remixing can reduce unpleasant artifacts and distortion compared to complete separation [22]. Applying independent processing to different sound sources can also reduce the distortion introduced by nonlinear processing algorithms such as dynamic range compression [23]. Here we focus on the benefits of remixing for spectral and spatial distortion in a linear time-invariant space-time filter. We show that the less the mixture is altered, the lower the resulting distortion.

2. SPACE-TIME FILTER FOR SOURCE REMIXING

2.1. Signal model

Consider an array of M microphones. By convention, it is assumed that microphone 1 is in or near the left ear and microphone 2 is in or near the right ear; these act as reference microphones for the purposes of preserving interaural cues. The mixture is assumed to include N *source channels.* A source channel is defined as a set of sounds that are to be processed as a group, that is, for which the desired response of the system is the same. A source channel could be an individual talker, a group of talkers, diffuse noise, or all sounds of a certain type, for example.

Let $\mathbf{x}(t) \in \mathbb{R}^M$ be the vector of continuous-time signals observed by the microphones. It is the sum of N source images $\mathbf{c}_n(t)$ due to the source channels:

$$\mathbf{x}(t) = \sum\nolimits_{n=1}^{N} \mathbf{c}_n(t). \tag{1}$$

Let $\mathbf{y}(t) \in \mathbb{R}^2$ be the desired output of the system. By convention, channel 1 is output to the left ear and channel 2 is output to the right ear. For the purposes of this work, assume that the desired processing to be applied to each source channel n is linear and time-invariant so that it can be characterized by an impulse response matrix $\mathbf{g}_n(\tau) \in \mathbb{R}^{2\times M}$. The desired output is therefore

$$\mathbf{y}(t) = \sum_{n=1}^{N} \underbrace{\int_{-\infty}^{\infty} \mathbf{g}_n(\tau)\mathbf{c}_n(t-\tau)\,\mathrm{d}\tau}_{\mathbf{d}_n(t)}. \tag{2}$$

The signal $\mathbf{d}_n(t)$ is the desired output image for source channel n.

In a perceptually transparent listening system, the listener should perceive $\mathbf{d}_n(t)$ as a natural sound that replaces $\mathbf{c}_n(t)$. To ensure that $\mathbf{d}_n(t)$ sounds natural, it should have a similar frequency spectrum to $\mathbf{c}_n(t)$—or a deliberately altered spectrum, for example to amplify high frequencies for hearing-impaired listeners—and it should be perceived as coming from the same direction. It should also have imperceptibly low delay, which limits the amount of frequency-selective processing that can be applied [24].

Because this work is designed to highlight interaural cue preservation, we restrict our attention to desired responses of the form

$$\mathbf{g}_n(\tau) = g_n(\tau)\begin{bmatrix}\mathbf{e}_1 & \mathbf{e}_2\end{bmatrix}^T, \tag{3}$$

where $\mathbf{e}_m$ is the unit vector with value 1 in position m and 0 elsewhere. In other words, the desired output in the left (respectively right) ear is the source signal as observed by the left (respectively right) reference microphone and processed by a diotic impulse response $g_n(\tau)$. If this desired response could be achieved exactly, the same processing would be applied to the signals in both ears and the interaural cues of all source channels would be perfectly preserved.

2.2. Space-time filtering

The output $\hat{\mathbf{y}}(t) \in \mathbb{R}^2$ of the system is produced by a linear time-invariant space-time filter $\mathbf{w}(\tau) \in \mathbb{R}^{2\times M}$ such that

$$\hat{\mathbf{y}}(t) = \int_{-\infty}^{\infty} \mathbf{w}(\tau)\mathbf{x}(t-\tau)\,\mathrm{d}\tau \tag{4}$$

$$= \sum_{n=1}^{N} \underbrace{\int_{-\infty}^{\infty} \mathbf{w}(\tau)\mathbf{c}_n(t-\tau)\,\mathrm{d}\tau}_{\hat{\mathbf{d}}_n(t)}. \tag{5}$$

The signal $\hat{\mathbf{d}}_n(t) \in \mathbb{R}^2$ is the output image for source channel n; it is the processed version of the original source image $\mathbf{c}_n(t)$ perceived by the listener.

In an AL system, sound signals must be processed in real time with a total delay of no more than a few milliseconds to avoid disturbing distortion. Thus, $\mathbf{w}(\tau)$ should be a causal filter that predicts a possibly delayed version of $\mathbf{y}(t)$. It has been shown that such a constraint limits the performance of a space-time filter, especially for small arrays [24]. Because space-time filters are most conveniently studied in the frequency domain, $\mathbf{w}(\tau)$ is allowed to be noncausal in our mathematical analysis, but the discrete-time filters implemented in Sec. 4 are causal with realistic delay constraints.

2.3. Weighted remixing filter

We would like to choose $\mathbf{w}(\tau)$ so that $\hat{\mathbf{d}}_n(t) \approx \mathbf{d}_n(t)$ for all n. To make the space-time filter as flexible as possible, we will use the multiple-speech-distortion-weighted multichannel Wiener filter (MSDW-MWF) [25]. The MSDW-MWF is a generalization of the well-known speech-distortion-weighted multichannel Wiener filter (SDW-MWF) [26], which allows the system designer to trade noise for spectral distortion of a single target source. The MSDW-MWF minimizes the following weighted squared-error cost function:

$$\text{Cost} = \sum_{n=1}^{N} \lambda_n \mathbb{E}\left[\left\|\hat{\mathbf{d}}_n(t) - \mathbf{d}_n(t)\right\|^2\right], \tag{6}$$

where $\mathbb{E}$ denotes statistical expectation and λ_n is a *distortion weight* that controls the relative importance of each sound source.

If the source images are statistically uncorrelated with each other, then the noncausal MSDW-MWF is given by the frequency-domain expression

$$\mathbf{W}(\Omega) = \left(\sum_{n=1}^{N} \lambda_n \mathbf{G}_n(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega)\right) \underbrace{\left(\sum_{n=1}^{N} \lambda_n \mathbf{R}_{\mathbf{c}_n}(\Omega)\right)^{-1}}_{\bar{\mathbf{R}}_{\mathbf{x}}(\Omega)}, \tag{7}$$

where $\mathbf{G}_n(\Omega)$ is the Fourier transform of $\mathbf{g}_n(\tau)$ and $\mathbf{R}_{\mathbf{c}_n}(\Omega)$ is the power spectral density of source image $\mathbf{c}_n(t)$ for $n = 1, \ldots, N$, which must be measured or estimated. It is assumed that $\bar{\mathbf{R}}_{\mathbf{x}}(\Omega)$ is invertible, which can be achieved by including a diffuse noise channel with full-rank spectral density. Many commonly used space-time filters can be derived from the MSDW-MWF: the multichannel Wiener filter, which minimizes mean squared error, is a special case with $\lambda_n = 1$ for all n. Linearly constrained filters such as the minimum variance distortionless response beamformer are limiting cases as the distortion weights approach infinity [2].

The remainder of the analysis will be performed in the frequency domain and we omit the frequency variable Ω for brevity.

2.4. Spectral distortion of the remixing filter

The following identity will be useful in analyzing the performance of the MSDW-MWF:

$$\mathbf{W} = \sum_{m=1}^{N} \lambda_m \mathbf{G}_m \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \tag{8}$$

$$= \mathbf{G}_n + \sum_{m=1}^{N} \lambda_m (\mathbf{G}_m - \mathbf{G}_n) \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1}. \tag{9}$$

If the filter is computed using the true source statistics, then the weighted error spectral density of the MSDW-MWF is given by

$$\bar{\mathbf{R}}_{\text{err}} = \sum_{n=1}^{N} \lambda_n \left(\mathbf{G}_n - \mathbf{W}\right) \mathbf{R}_{\mathbf{c}_n} \left(\mathbf{G}_n - \mathbf{W}\right)^H \tag{10}$$

$$= \sum_{n=1}^{N} \sum_{m=1}^{M} \lambda_n \lambda_m \left(\mathbf{G}_n - \mathbf{G}_m\right) \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \left(\mathbf{G}_n - \mathbf{W}\right)^H \tag{11}$$

$$= \sum_{n=1}^{N} \sum_{m=1}^{M} \lambda_n \lambda_m \left(\mathbf{G}_n - \mathbf{G}_m\right) \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \mathbf{G}_n^H, \tag{12}$$

where the last step follows from the orthogonality principle. Thus, the performance of the remixing filter can be expressed in terms of pairs of source channels. It is clear from this expression that if $\mathbf{G}_n = \mathbf{G}_m$, then the (m, n) source pair contributes nothing to the error of the system: if we wish to apply the same processing to both sources, then we need not separate them.

If each desired response has the form of (3) and if the Fourier transform $G_n(\Omega)$ of $g_n(\tau)$ is real-valued for all $n = 1, \ldots, N$, then by manipulating (12) the weighted error spectra in the left and right ears can be written

$$\bar{R}_{\text{err}}^{\text{left}} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \lambda_n \lambda_m \left|G_n - G_m\right|^2 \mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \mathbf{e}_1 \tag{13}$$

$$\bar{R}_{\text{err}}^{\text{right}} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \lambda_n \lambda_m \left|G_n - G_m\right|^2 \mathbf{e}_2^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \mathbf{e}_2. \tag{14}$$

Thus, the performance of the remixing filter depends on the differences between desired responses and on the spatial and spectral separability of the signal pairs.

3. DISTORTION OF INTERAURAL CUES

The interaural level difference (ILD) and interaural phase difference (IPD) can both be derived from the interaural transfer function (ITF). If the source image $\mathbf{c}_n(t)$ for channel n has a Fourier transform $\mathbf{C}_n(\Omega)$, then the input and output ITFs for source channel n are

$$\text{ITF}_n^{\text{in}} = \frac{\mathbf{e}_2^T \mathbf{C}_n}{\mathbf{e}_1^T \mathbf{C}_n} \tag{15}$$

$$\text{ITF}_n^{\text{out}} = \frac{\mathbf{e}_2^T \hat{\mathbf{D}}_n}{\mathbf{e}_1^T \hat{\mathbf{D}}_n} = \frac{\mathbf{e}_2^T \mathbf{W} \mathbf{C}_n}{\mathbf{e}_1^T \mathbf{W} \mathbf{C}_n}. \tag{16}$$

The ILD and IPD are the magnitude and phase of the ITF:

$$\text{ILD}_n = 20 \log_{10} \left|\text{ITF}_n\right| \quad \text{and} \quad \text{IPD}_n = \angle \text{ITF}_n. \tag{17}$$

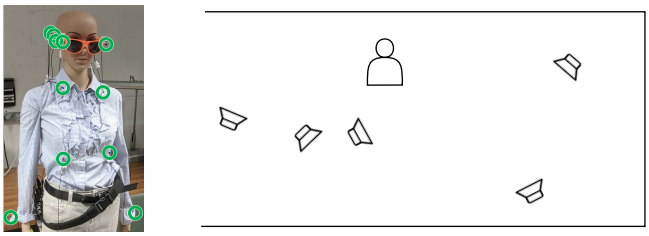

Fig. 2: Five loudspeakers were placed around an acoustically treated laboratory. A wearable array includes up to 16 microphones.

3.1. Spatial distortion of the MSDW-MWF

The output ITF for the MSDW-MWF (9) with binaurally matched responses (3) is

$$\text{ITF}_n^{\text{out}} = \frac{G_n \mathbf{e}_2^T \mathbf{C}_n + \sum_{m=1}^{N} \lambda_m \left(G_m - G_n\right) \mathbf{e}_2^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{C}_n}{G_n \mathbf{e}_1^T \mathbf{C}_n + \sum_{m=1}^{N} \lambda_m \left(G_m - G_n\right) \mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{C}_n} \tag{18}$$

for $n = 1, \ldots, N$. Notice that if the second terms in the numerator and denominator were removed, the output ITF would be identical to the input ITF. This would be the case if the same processing were applied to every source channel so that all $G_m - G_n = 0$ or if the sources were fully separable so that $\mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{C}_n = \mathbf{0}$ for $m \neq n$.

The error in the ILD and IPD are the real and imaginary parts, respectively, of the logarithm of $\text{ITF}_n^{\text{out}} / \text{ITF}_n^{\text{in}}$. If G_n and ITF_n^{in} are nonzero, then the ITF error for source channel n can be written

$$\Delta\text{ITF}_n = \ln \frac{1 + \sum_{m=1}^{N} \lambda_m \frac{G_m - G_n}{G_n} \frac{\mathbf{e}_2^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{C}_n}{\mathbf{e}_2^T \mathbf{C}_n}}{1 + \sum_{m=1}^{N} \lambda_m \frac{G_m - G_n}{G_n} \frac{\mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{C}_n}{\mathbf{e}_1^T \mathbf{C}_n}}. \tag{19}$$

3.2. First-order approximation

Using the first-order approximation $\ln(1 + u) \approx u$ for both the numerator and denominator of (19), the logarithmic ITF error is

$$\Delta\text{ITF}_n \approx \sum_{m=1}^{N} \lambda_m \frac{G_m - G_n}{G_n} \left(\frac{\mathbf{e}_2^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{C}_n}{\mathbf{e}_2^T \mathbf{C}_n} - \frac{\mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{C}_n}{\mathbf{e}_1^T \mathbf{C}_n} \right). \tag{20}$$

Furthermore, if diffuse noise were negligible and every source channel were well modeled by a rank-1 spectral density matrix $\mathbf{R}_{\mathbf{c}_n} \approx R_{s_n} \mathbf{A}_n \mathbf{A}_n^H$ with $\mathbf{C}_n$ parallel to the acoustic transfer function $\mathbf{A}_n$ for $n = 1, \ldots, N$, then the ITF error would be

$$\Delta\text{ITF}_n \approx \sum_{m=1}^{N} \lambda_m R_{s_m} \frac{G_m - G_n}{G_n} \mathbf{A}_m^H \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n \left(\frac{\mathbf{e}_2^T \mathbf{A}_m}{\mathbf{e}_2^T \mathbf{A}_n} - \frac{\mathbf{e}_1^T \mathbf{A}_m}{\mathbf{e}_1^T \mathbf{A}_n} \right) \tag{21}$$

Thus, spatial distortion depends on the power and distortion weight of each interfering source, the relative difference in desired responses between sources, the spatial separability of the sources, and the difference in interaural cues between source channels. Distant sound sources that have different acoustic transfer functions will be easier to separate (small $\mathbf{A}_m^H \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n$) but also have more different interaural cues (large $\frac{\mathbf{e}_2^T \mathbf{A}_m}{\mathbf{e}_2^T \mathbf{A}_n} - \frac{\mathbf{e}_1^T \mathbf{A}_m}{\mathbf{e}_1^T \mathbf{A}_n}$). Meanwhile, if $\mathbf{A}_m$ is parallel to $\mathbf{A}_n$, the source channels have the same interaural cues and therefore the source pair does not introduce spatial distortion.

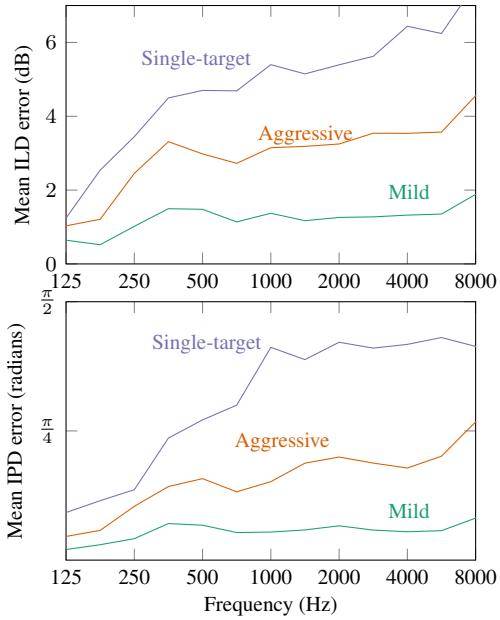

Fig. 3: Interaural cue preservation for a binaural source-remixing filter with 4 microphones and different desired responses.

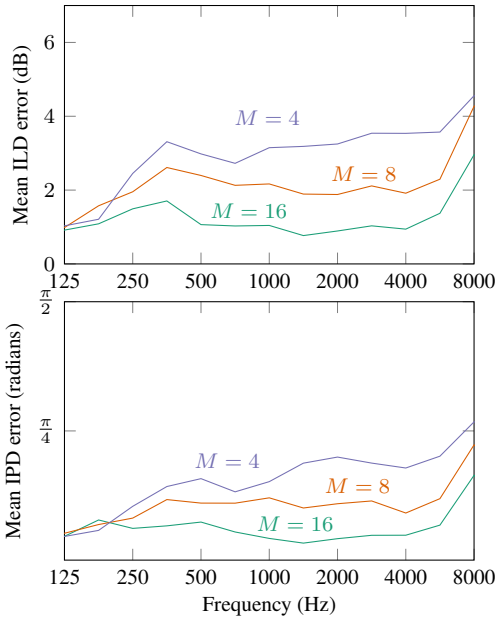

Fig. 4: Interaural cue preservation for a binaural source-remixing filter with different array configurations.

4. EXPERIMENTS WITH A WEARABLE ARRAY

To evaluate the performance of the source-remixing system for augmented listening applications, it was applied to a wearable microphone array in a challenging noisy environment. Sixteen microphones were spread across the body of a mannequin, as shown in Fig. 2. One microphone was placed just outside each ear canal. Although this mannequin does not have realistic head-related transfer functions, it is suitable for this experiment because its head has contralateral attenuation that is only slightly weaker than that of a human [27]. The sound sources were five speech clips derived from the CSTR VCTK corpus [28] and played through loudspeakers as well as the diffuse mechanical and ventilation noise in the room.

Impulse response measurements and long-term average speech and noise spectra were used to design causal discrete-time MSDW-MWFs with a delay of 16 ms and a unit pulse response length of 256 ms. The experimental ITFs of the five speech sources were measured in the STFT domain using their sample cross-correlations [15]:

$$\mathrm{ITF}_n^{\mathrm{in}}[f] = \frac{\sum_k \mathbf{e}_1^T \mathbf{C}_{\mathrm{stft},n}[k,f] \mathbf{C}_{\mathrm{stft},n}^H[k,f] \mathbf{e}_2}{\sum_k \mathbf{e}_1^T \mathbf{C}_{\mathrm{stft},n}[k,f] \mathbf{C}_{\mathrm{stft},n}^H[k,f] \mathbf{e}_1} \tag{22}$$

$$\mathrm{ITF}_n^{\mathrm{out}}[f] = \frac{\sum_k \mathbf{e}_1^T \hat{\mathbf{D}}_{\mathrm{stft},n}[k,f] \hat{\mathbf{D}}_{\mathrm{stft},n}^H[k,f] \mathbf{e}_2}{\sum_k \mathbf{e}_1^T \hat{\mathbf{D}}_{\mathrm{stft},n}[k,f] \hat{\mathbf{D}}_{\mathrm{stft},n}^H[k,f] \mathbf{e}_1}, \tag{23}$$

for $n = 1, \ldots, 5$. The ILD and IPD errors were computed from the ITFs and averaged over the five directional sources.

Figure 3 shows the performance of earpieces with $M = 4$ total microphones for three sets of target responses:

1. Mild remixing with gains 1.0, 0.8, 0.7, 0.6, and 0.5 on the speech channels and 20 dB attenuation on the noise channel,
2. Aggressive remixing with gains 1.0, 0.4, 0.3, 0.2, and 0.1 on the speech channels and 20 dB attenuation on noise, and
3. Single-target beamforming with $G_1 = 1$ and $G_2 = \cdots = G_6 = 0$.

The mild filter has low interaural cue distortion, the single-target beamformer severely distorts the non-target sources, and the aggressive filter falls in between. Distortion is mild below a few hundred hertz; these wavelengths are much larger than a human head and so the ILD and IPD of all sources are close to zero.

A larger wearable array should be able to apply complex remixing to more sources than a small earpiece-based array can. Figure 4 shows the ILD and ITD distortion for the "aggressive" remixing responses with arrays of different sizes. The four-microphone earpiece array does not have enough degrees of freedom to preserve the interaural cues of all five directional sources. The 8-microphone head-mounted array does better, and the 16-microphone upper-body array produces little distortion in any of the source channels.

5. CONCLUSIONS

The design of source-remixing filters requires a tradeoff between audio enhancement—removing and altering different sound sources to improve intelligibility—and perceptual transparency. Filters that alter the signal less, that is, that apply similar desired responses to the different source channels, cause less spectral distortion and less interaural cue distortion. They also sound more immersive and natural to the listener. However, they do not provide as much benefit in complex noise. Space-time remixing filters with large wearable microphone arrays could provide the advantages of both approaches: they have enough spatial resolution to meaningfully suppress strong background noise, but they have enough degrees of freedom to ensure that those attenuated noise sources sound natural.

A full understanding of source-remixing filter design tradeoffs will require new clinical research. The choice of desired responses will likely depend on the nature of the sources and environment and on the preferences of the individual. Compared to conventional beamforming and source separation, audio source remixing provides a more versatile approach to human sensory augmentation that could dramatically—but seamlessly—change how we perceive our world.

6. REFERENCES

[1] V. Valimaki, A. Franck, J. Ramo, H. Gamper, and L. Savioja, "Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 92–99, 2015.

[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] W. Soede, F. A. Bilsen, and A. J. Berkhout, "Assessment of a directional microphone array for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 94, no. 2, pp. 799–808, 1993.

[4] J. E. Greenberg and P. M. Zurek, *Microphone-Array Hearing Aids*, pp. 229–253. Berlin: Springer Berlin Heidelberg, 2001.

[5] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks* (S. Haykin and K. R. Liu, eds.), pp. 269–302, Wiley, 2008.

[6] 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.

[7] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[8] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica United with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[9] 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.

[10] T. Van den Bogaert, *Preserving Binaural Cues in Noise Reduction Algorithms for Hearing Aids*. PhD thesis, KU Leuven, June 2008.

[11] D. Marquardt, *Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques*. PhD thesis, Carl von Ossietzky University of Oldenburg, 2016.

[12] A. Koutrouvelis, *Multi-Microphone Noise Reduction for Hearing Assistive Devices*. PhD thesis, Delft University of Technology, 2018.

[13] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1579–1585, 2007.

[14] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

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

[16] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel Wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2384–2397, 2015.

[17] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 543–558, 2016.

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

[19] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, "Improved multi-microphone noise reduction preserving binaural cues," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 460–464, 2016.

[20] J.-M. Jot, B. Smith, and J. Thompson, "Dialog control and enhancement in object-based audio systems," in *Audio Engineering Society Convention*, 2015.

[21] B. Shirley, M. Meadows, F. Malak, J. Woodcock, and A. Tidball, "Personalized object-based audio for hearing impaired TV viewers," *Journal of the Audio Engineering Society*, vol. 65, pp. 293–303, 2017.

[22] H. Wierstorf, D. Ward, R. Mason, E. M. Grais, C. Hummersone, and M. D. Plumbley, "Perceptual evaluation of source separation for remixing music," in *Audio Engineering Society Convention*, 2017.

[23] R. M. Corey and A. C. Singer, "Dynamic range compression for noisy mixtures using source separation and beamforming," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[24] R. M. Corey, N. Tsuda, and A. C. Singer, "Delay-performance tradeoffs in causal microphone array processing," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

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

[26] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel Wiener filtering techniques for noise reduction," in *Speech Enhancement* (J. Benesty, S. Makino, and J. Chen, eds.), pp. 199–228, Springer, 2005.

[27] 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.

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