ADAPTIVE BINAURAL FILTERING FOR A MULTIPLE-TALKER LISTENING SYSTEM USING REMOTE AND ON-EAR MICROPHONES

Ryan M. Corey and Andrew C. Singer

University of Illinois Urbana-Champaign

ABSTRACT

Listening devices such as hearing aids can use remote microphones placed near a distant talker to improve intelligibility in noisy environments. Most commercial remote microphone systems are designed for a single talker and they remove spatial cues that help listeners to localize sounds. Here we propose an assistive listening system that enhances speech from multiple talkers while preserving the acoustic effects of the environment and the spatial cues of each sound source. The system adaptively filters the remote microphone signals to match the signals from the earpiece microphones. Because it uses the on-ear microphones as references, the adaptive system does not need to localize or separate the sound sources and can track changes as the listener and talkers move. We compare two adaptive implementations, one that processes the microphones jointly and one that adapts each separately. The system is demonstrated experimentally with up to three remote microphones and moving talkers and listeners.

***Index Terms*—** Binaural processing, hearing aids, remote microphones, adaptive filters

1. INTRODUCTION

Listening devices such as hearing aids and cochlear implants often perform poorly in noisy environments. Remote microphones (RM), which transmit sound directly from a distant talker to the ears of a listener, have been shown to improve intelligibility in adverse environments [1–3]. The signal from a remote microphone has less noise and reverberation than the signals captured by the earpieces of a listening device, effectively bringing the talker closer.

Although they can dramatically improve intelligibility, remote microphones often sound artificial. In commercial devices, the signal from the remote microphone is generally presented diotically. This signal matches the spectral coloration of the RM rather than that of the earpieces, and it lacks interaural time and level differences that humans use to localize sounds. Researchers have proposed systems that estimate the direction of arrival of a sound and then apply filters that simulate spatial cues for that direction [4–6]. Alternatively, in listening systems that include earpieces, spatial cues can be preserved by matching the magnitude and phase of the processed signal to those of the earpiece microphones (EM). This latter approach is commonly used in binaural beamformers [7–10], which coherently combine signals from several microphones to emphasize signals from a target direction and attenuate others. Researchers have proposed binaural beamformers that incorporate external microphones into an array [11–15]. Because the external microphones in those works are not necessarily close to the talkers of interest, beamforming is necessary to achieve strong noise reduction, but such systems can be difficult to implement and are sensitive to motion. Since RMs placed near talkers already have low noise, we can use a simpler approach: Filter the low-noise RM signals to match the magnitude and phase of each on-ear microphone, ensuring that spatial cues are preserved.

Spatial cues are especially important with multiple conversation partners, as they help listeners to distinguish signals from different talkers. To preserve the spatial cues of multiple talkers, a single RM is not enough; as we will show, the system must have at least as many microphones as talkers. Here we consider two strategies for enhancing speech from multiple talkers. First, we could place one RM on or near each talker, as in a theater production or panel discussion. Each microphone would provide a reliable reference signal from its wearer, even when moving. Second, we could use a microphone array placed near a group of talkers, as in a conference room. An array enhances all nearby sounds, so it is suitable for dynamic environments where talkers may freely join or leave the conversation. We will show that these approaches have different strengths and weaknesses related to noise, motion, and crosstalk.

A key advantage of external and remote microphones is that ambient noise is weakly correlated between the remote and earpiece microphones. This property has been used to identify the acoustic channel between talkers of interest and the microphones of an array [16–18]. Here, we exploit this correlation property to match the magnitude and phase of the RM signals to those at the ears. An adaptive filter uses the RM signals as inputs and the EM signals as references for the desired outputs. If the noise is uncorrelated between the input and reference signals, then the filter will match the cues of the signals of interest. This adaptive approach never explicitly estimates the acoustic channel or attempts to separate the sources. We propose two variants of the adaptive filter: a jointly adapted multiple-input, binaural-output (MIBO) filter suitable for arrays and closely grouped talkers, and a set of independently adapted single-input, binaural-output (SIBO) filters for wearable microphones on spatially separated moving talkers.

2. MULTIPLE-TALKER ASSISTIVE LISTENING SYSTEM

2.1. System model

A binaural listening device contains two microphones, one at each ear. It is supplemented by M remote microphones placed near N talkers of interest. In this work, we assume that the RM signals are available instantaneously and synchronously to the device.

Let $\mathbf{s}[t] = [s_1[t], \dots, s_N[t]]^T$ be the sampled speech signals produced by the talkers of interest. Consider a short time interval

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 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.

during which the talkers, listener, and microphones do not move. The discrete-time signals $\mathbf{x}_\mathrm{e}[t] \in \mathbb{R}^2$ received by the EMs and $\mathbf{x}_\mathrm{r}[t] = [x_{\mathrm{r},1}[t], \ldots, x_{\mathrm{r},M}[t]]^T$ received by the RMs are given by

$$\mathbf{x}_\mathrm{e}[t] = \sum_{n=1}^{N} (\mathbf{a}_{\mathrm{e},n} \star s_n)[t] + \mathbf{z}_\mathrm{e}[t] \tag{1}$$

$$\mathbf{x}_\mathrm{r}[t] = \sum_{n=1}^{N} (\mathbf{a}_{\mathrm{r},n} \star s_n)[t] + \mathbf{z}_\mathrm{r}[t], \tag{2}$$

where $\star$ denotes linear convolution, $\mathbf{a}_{\mathrm{e},n}[t] \in \mathbb{R}^2$ and $\mathbf{a}_{\mathrm{r},n}[t] \in \mathbb{R}^M$ are equivalent discrete-time acoustic impulse responses between source n and the EMs and RMs, respectively, for $n = 1, \ldots, N$, and $\mathbf{z}_\mathrm{e}[t] \in \mathbb{R}^2$ and $\mathbf{z}_\mathrm{r}[t] \in \mathbb{R}^M$ are additive noise at the EMs and RMs, respectively.

The system produces a binaural output $\mathbf{y}[t] \in \mathbb{R}^2$ given by

$$\mathbf{y}[t] = \sum_{m=1}^{M} (\mathbf{w}_m \star x_{\mathrm{r},m})[t], \tag{3}$$

where $\mathbf{w}_m[t] \in \mathbb{R}^2$ is a discrete-time binaural filter for inputs $m = 1, \ldots, M$. Note that unlike in a binaural beamformer, the EM signals are not inputs to the filter used to generate $\mathbf{y}[t]$. However, the EM signals could be mixed with $\mathbf{y}[t]$ if desired to improve spatial awareness of ambient noise [8].

2.2. Optimization problem

The system is designed to be perceptually transparent so that the binaural output approximates the signal captured by the earpiece microphones but with less noise. Mathematically, the desired output $\mathbf{d}[t] \in \mathbb{R}^2$ is given by

$$\mathbf{d}[t] = \sum_{n=1}^{N} (g_n \star \mathbf{a}_{\mathrm{e},n} \star s_n)[t], \tag{4}$$

where $g_n[t] \in \mathbb{R}$ is the desired processing to be applied to each source n. The g_n's can be used to apply different amplification and spectral shaping to each source, for example based on distance. The binaural impulse responses $\mathbf{a}_{\mathrm{e},n}$ encode the effects of room acoustics on the spectrum of each speech signal as well as the interaural time and level differences used to localize sounds.

It will be convenient to analyze the filters in the frequency domain. Let $\mathbf{W}(\omega) \in \mathbb{C}^{2\times M}$, $\mathbf{A}_\mathrm{e}(\omega) \in \mathbb{C}^{2\times N}$, $\mathbf{A}_\mathrm{r}(\omega) \in \mathbb{C}^{M\times N}$, and $\mathbf{G}(\omega) \in \mathbb{C}^{N\times N}$ be the discrete-time Fourier transforms of their respective impulse responses, where $\mathbf{G}$ is a diagonal matrix of desired responses for the N sources. To preserve the spectral and spatial cues of the N distinct sources, the filter should satisfy

$$\mathbf{W}(\omega)\mathbf{A}_\mathrm{r}(\omega) = \mathbf{A}_\mathrm{e}(\omega)\mathbf{G}(\omega). \tag{5}$$

For arbitrary $\mathbf{A}_\mathrm{r}$, the filter can only meet this condition if $M \geq N$, that is, we have at least as many RMs as sources.

Adaptive filters are often designed to minimize the mean square error (MSE) between the output and desired signals. If the speech sources and noise were wide-sense stationary random processes with known second-order statistics and if the acoustic impulse responses were known, we could directly minimize the MSE loss

$$\mathrm{MSE}[t] = \mathbb{E}\left[|\mathbf{y}[t] - \mathbf{d}[t]|^2\right], \tag{6}$$

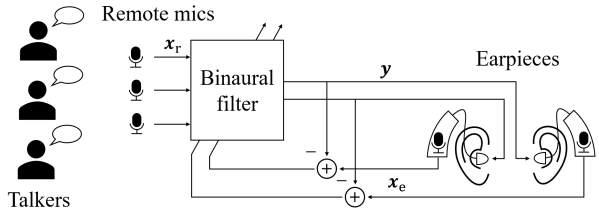

Figure 1: The binaural adaptive filter system uses the earpiece microphones as reference signals.

where $\mathbb{E}$ denotes statistical expectation. If the filters are allowed to be noncausal and to have infinite length, then the linear minimum-mean-square-error (MMSE) filter can be readily computed in the frequency domain. We assume that all signals have zero mean and that the speech signals are uncorrelated with the noise signals. Let $\mathbf{R}_\mathbf{s}(\omega) \in \mathbb{C}^{N\times N}$, $\mathbf{R}_{\mathbf{z}_\mathrm{e}}(\omega) \in \mathbb{C}^{2\times 2}$, and $\mathbf{R}_{\mathbf{z}_\mathrm{r}}(\omega) \in \mathbb{C}^{M\times M}$ be the power spectral density matrices for $\mathbf{s}[t]$, $\mathbf{z}_\mathrm{e}[t]$, and $\mathbf{z}_\mathrm{r}[t]$, respectively, and let $\mathbf{R}_{\mathbf{z}_\mathrm{e}\mathbf{z}_\mathrm{r}}(\omega) \in \mathbb{C}^{2\times M}$ be the cross-power spectral density between $\mathbf{z}_\mathrm{e}[t]$ and $\mathbf{z}_\mathrm{r}[t]$. Then the MMSE filter is given by

$$\mathbf{W}_\mathrm{MMSE}(\omega) = \mathbf{A}_\mathrm{e}(\omega)\mathbf{G}(\omega)\mathbf{R}_\mathbf{s}(\omega)\mathbf{A}_\mathrm{r}^H(\omega) \cdot [\mathbf{A}_\mathrm{r}(\omega)\mathbf{R}_\mathbf{s}(\omega)\mathbf{A}_\mathrm{r}^H(\omega) + \mathbf{R}_{\mathbf{z}_\mathrm{r}}(\omega)]^{-1}. \tag{7}$$

If $\mathbf{A}_\mathrm{r}$ has full column rank, then the Woodbury identity can be used to show that the MMSE filter satisfies (5) in the high-SNR limit.

In the remainder of the paper we omit the frequency variable ω for brevity.

3. ADAPTIVE FILTERING USING EARPIECES

The MMSE filter relies on the signal statistics and the transfer functions between the sources and microphones, which can be difficult to estimate. Fortunately, when RMs are close to the sources, they provide high-quality reference signals that eliminate the need for complex source separation algorithms. Gößling and Doclo [17] noted that ambient noise signals are often mostly uncorrelated between on-ear and remote microphones. They used this property to efficiently estimate the relative transfer function between the source and earpieces using the noisy mixture. We can apply the same principle to the adaptive filtering problem, replacing the desired signal $\mathbf{d}[t]$ with the noisy EM signal, as shown in Fig. 1.

In this section, we consider two adaptive implementations of the binaural enhancement system: a multiple-input, binaural-output (MIBO) filter that processes all microphones jointly, and a set of single-input, binaural-output (SIBO) filters computed separately for each microphone.

3.1. Multiple-input filter

Suppose that the desired response is the same for all talkers, that is, $g_n[t] = g[t]$ for all n and $\mathbf{G}(\omega) = G(\omega)\mathbf{I}$ for all ω. Instead of minimizing the true MSE, we can minimize the loss function

$$\mathcal{L}[t] = \mathbb{E}\left[|\mathbf{y}[t] - (g \star \mathbf{x}_\mathrm{e})[t]|^2\right]. \tag{8}$$

If the signals are wide-sense stationary, then the linear MMSE filter that minimizes $\mathcal{L}$ is given in the frequency domain by

$$\mathbf{W}_\mathrm{MIBO} = G[\mathbf{A}_\mathrm{e}\mathbf{R}_\mathbf{s}\mathbf{A}_\mathrm{r}^H + \mathbf{R}_{\mathbf{z}_\mathrm{e}\mathbf{z}_\mathrm{r}}][\mathbf{A}_\mathrm{r}\mathbf{R}_\mathbf{s}\mathbf{A}_\mathrm{r}^H + \mathbf{R}_{\mathbf{z}_\mathrm{r}}]^{-1}. \tag{9}$$

This filter attempts to replicate both the desired speech and the unwanted noise at the ears. However, if the noise is uncorrelated between the EMs and RMs, then $\mathbf{R}_{\mathbf{z}_\mathrm{e}\mathbf{z}_\mathrm{r}}(\omega) = \mathbf{0}$ and the adaptive filter (9) is identical to the MMSE filter (7). That is, the filter cannot use the RM inputs to predict the noise, only the talkers of interest.

With correlated noise, the spatial cues of the target are distorted by those of the noise, as can be readily seen in the special case where $M = N = 1$:

$$\mathbf{W}_{\mathrm{MIBO}} A_\mathrm{r} = G \frac{\mathbf{A}_\mathrm{e} |A_\mathrm{r}|^2 R_s + \mathbf{R}_{\mathbf{z}_\mathrm{e} z_\mathrm{r}} A_r}{|A_\mathrm{r}|^2 R_s + R_{z_\mathrm{r}}}. \tag{10}$$

In the numerator, the noise at the EMs distorts interaural cues to the extent that it is correlated with the noise at the RM. In the denominator, the magnitude of the noise at the RM alters the magnitude of the output, just as it would for the MMSE filter. Thus, system performance depends strongly on RM placement.

A key property of the MIBO adaptive filter is that it does not separate the sources of interest, nor does it explicitly model their acoustic transfer functions. Since the inputs to the filter can be combinations of the speech signals of interest, the MIBO filter is suitable for systems with significant crosstalk, such as wearable microphones on nearby talkers or a microphone array placed near a group of talkers. It can also adapt easily as talkers move around the area near the microphones or as they enter and leave a conversation, as long as no more than M talkers participate at a time.

This adaptability of the MIBO filter comes with drawbacks. If $M > N$, the MMSE filter (7) uses its additional degrees of freedom to reduce noise, but the adaptive MIBO filter (9) does not. Because it does not attempt to localize or separate the sources, it cannot distinguish between desired and undesired sounds. It will enhance up to M sound sources that have strong correlation between the on-ear and remote microphones, including unwanted noise. Furthermore, because it coherently combines signals from multiple microphones, it is sensitive to unmodeled relative motion between microphones.

3.2. Single-input filters

If we wish to restrict the listening system to only certain talkers or to apply different amplification to different talkers, or if the microphones move so that coherent combining is difficult, then the MIBO system is not suitable. Instead, we can place one RM on or near each talker of interest. Each SIBO filter $\mathbf{w}_m$ is designed to reproduce the speech of talker m, so that $(\mathbf{w}_m \star x_{\mathrm{r},m})[t] \approx (g_m \star \mathbf{a}_{\mathrm{e},m} \star s_m)[t]$ for $m = 1, \ldots, M = N$. Each filter is computed separately to minimize its own loss function

$$\mathcal{L}_m[t] = \mathbb{E}\left[|(\mathbf{w}_m \star x_{\mathrm{r},m})[t] - (g_m \star \mathbf{x}_\mathrm{e})[t]|^2\right]. \tag{11}$$

The solution is given in the frequency domain by the 2×1 filter

$$\mathbf{W}_m = G_m [\mathbf{A}_\mathrm{e} \mathbf{R}_\mathbf{s} \mathbf{A}_{\mathrm{r},m}^H + \mathbf{R}_{\mathbf{z}_\mathrm{e}\mathbf{z}_{\mathrm{r},m}}][\mathbf{A}_{\mathrm{r},m} \mathbf{R}_\mathbf{s} \mathbf{A}_{\mathrm{r},m}^H + R_{z_{\mathrm{r},m}}]^{-1}, \tag{12}$$

where $\mathbf{A}_{\mathrm{r},m}$ is the row of $\mathbf{A}_\mathrm{r}$ corresponding to microphone m. If the speech sources are uncorrelated, then the SIBO filter is

$$\mathbf{W}_m = G_m \frac{\mathbf{A}_{\mathrm{e},1} R_{s_1} A_{\mathrm{r},m,1}^* + \cdots + \mathbf{A}_{\mathrm{e},M} R_{s_M} A_{\mathrm{r},m,M}^* + \mathbf{R}_{\mathbf{z}_\mathrm{e} z_{\mathrm{r},m}}}{|A_{\mathrm{r},m,1}|^2 R_{s_1} + \cdots + |A_{\mathrm{r},m,M}|^2 R_{s_M} + R_{z_{\mathrm{r},m}}}. \tag{13}$$

It can be seen from (13) that the interaural cues are distorted by crosstalk among the RMs as well as by correlated noise. Crosstalk can also produce unintended interference effects, such as comb-filtering distortion, when the SIBO filter outputs are summed.

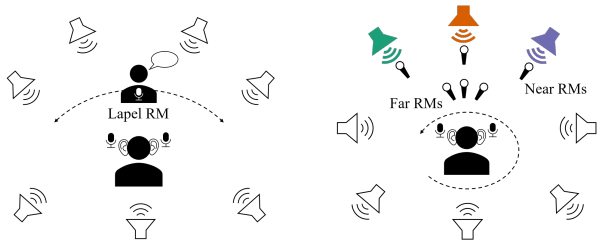

Figure 2: Experimental setup. Left: Single moving human talker and nonmoving listener. Right: Three loudspeaker talkers of interest and a moving listener. Additional loudspeakers produce unwanted noise.

4. DYNAMIC CAUSAL FILTER

The frequency-domain analysis above assumes that the filters can be noncausal and can have infinite length. In a real listening system, the filters must be causal and have finite length. Fortunately, because the RMs are placed near the talkers, the binaural filters should closely resemble the acoustic impulse responses between the talkers and listener. As long as the group delay of the desired responses (g_n) plus any transmission delay between the RMs and earpieces is smaller than the acoustic time of flight between talkers and listener, it should be possible to design causal binaural filters.

The analysis of the previous section also assumes that the acoustic system is stationary. In reality, human talkers and listeners move constantly. To adapt to changing conditions, we must use a time-varying filter. Let $\mathbf{w}_m[\tau; t] \in \mathbb{R}^2$ be the filter coefficients at time t for $m = 1, \ldots, M$ and $\tau = 0, \ldots, L-1$, where L is the length of each filter. The filter output is given by

$$\mathbf{y}[t] = \sum_{m=1}^{M} \sum_{\tau=0}^{L-1} \mathbf{w}_m[\tau; t] x_{\mathrm{r},m}[t - \tau]. \tag{14}$$

We can write (14) as a matrix-vector multiplication,

$$\mathbf{y}[t] = \bar{\mathbf{w}}[t] \bar{\mathbf{x}}_\mathrm{r}[t], \tag{15}$$

where $\bar{\mathbf{x}}_\mathrm{r}^T[t] = [\mathbf{x}_\mathrm{r}^T[t], \mathbf{x}_\mathrm{r}^T[t-1], \ldots, \mathbf{x}_\mathrm{r}^T[t-L+1]]$ and $\bar{\mathbf{w}} \in \mathbb{R}^{2 \times LM}$.

In the experiments in this work, we update the filter coefficients with the least mean squares (LMS) algorithm [19]. The MIBO update is given by

$$\bar{\mathbf{w}}[t+1] \leftarrow \bar{\mathbf{w}}[t] + \mu((g \star \mathbf{x}_\mathrm{e})[t] - \mathbf{y}[t]) \bar{\mathbf{x}}_\mathrm{r}^T[t], \tag{16}$$

where μ is a tunable step size parameter.

The SIBO updates have the same form except that each RM filter is adapted independently:

$$\bar{\mathbf{w}}_m[t+1] \leftarrow \bar{\mathbf{w}}_m[t] + \mu((g_m \star \mathbf{x}_\mathrm{e})[t] - \bar{\mathbf{w}}_m[t] \bar{\mathbf{x}}_{\mathrm{r},m}[t]) \bar{\mathbf{x}}_{\mathrm{r},m}^T[t]. \tag{17}$$

5. EXPERIMENTS

The proposed adaptive filtering system was evaluated experimentally using a binaural dummy head in an acoustically treated laboratory ($T_{60} \approx 250$ ms). Speech signals were either produced by a human talker or derived from the VCTK dataset [20] and played back over loudspeakers. Each talker was recorded separately and the

Sources	Listener	RMs	Input	Remote	MIBO	SIBO
1 moving	Still	Lapel	−4.3	9.3	—	7.2
3 still	Moving	Near	0.8	21.3	18.8	18.2
3 still	Moving	Far	0.7	12.0	9.8	11.7

Table 1: Wideband signal-to-noise ratio (dB) for acoustic experiments. Input and filter output SNRs are measured at the left ear.

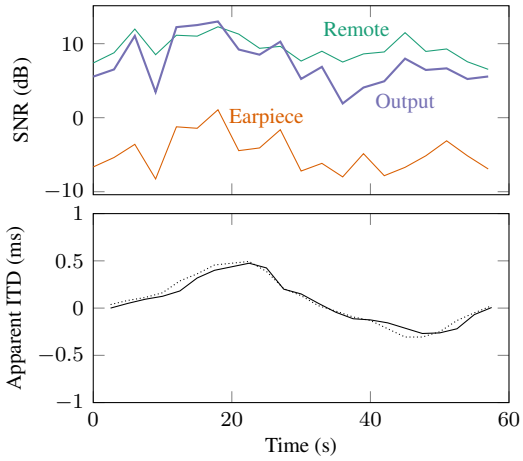

Figure 3: Filter performance for a single moving talker. Top: SNR at the left ear. Bottom: Apparent ITD of the target source in the filter output. The dotted curve shows the true ITD.

recordings were mixed to simulate simultaneous speech. For each experiment, the adaptive filter coefficients were computed based on the mixture but applied separately to each source recording in order to track the effect of the system on each component signal. The filters were about 20 ms in length and were designed to be transparent for the source(s) of interest ($g_n[t] = \delta[t]$). The step size μ was tuned manually. For each experiment, the wideband signal-to-noise ratio was computed after highpass filtering at 200 Hz to exclude mechanical noise in the laboratory. The apparent interaural time delays (ITD) were computed by finding the peak of the cross-correlation within overlapping 5 second windows. The experiments are summarized in Fig. 2 and Table 1.

5.1. Single moving talker and nonmoving listener

In the first experiment, which simulates the typical use case for remote microphone systems today, a lapel microphone was worn by a moving human talker. Noise was produced by seven loudspeakers placed around the room. The human subject followed the same route during each source recording so that sound and motion are roughly synchronized. The top plot of Fig. 3 shows the wideband input and output SNR at the left ear and the input SNR at the RM. The SNR varied as the talker moved among the interfering loudspeakers. The output SNR closely tracks the RM input SNR, as expected. The bottom plot shows the apparent ITD of the target speech at the output of the binaural filter compared to that of the clean signal at the ears. The adaptive filter is able to track the spatial cues as the talker moves from center to left to right and back again. Thus, the filter output matches the SNR of the remote microphone and the spatial cues of the earpieces.

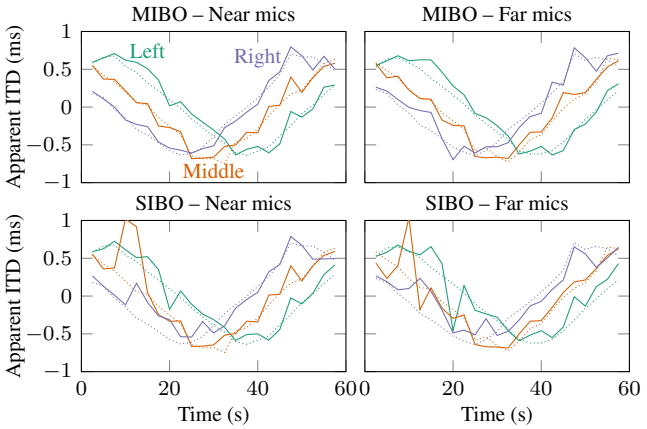

Figure 4: Apparent ITDs of processed speech sources at the ears of the rotating dummy head. The dotted curves show the true ITDs of the three loudspeaker sources illustrated in Fig. 2.

5.2. Multiple nonmoving talkers and moving listener

A second experiment simulated a multiple-talker application with a moving listener. The dummy head was placed on a motorized turntable, which made one rotation during the one minute recording, starting from the right in Fig. 2. Loudspeakers simulated three talkers of interest and five unwanted speech sources. The RMs were three end-address cardioid vocal microphones. First, to simulate personal RMs, each microphone was placed about 30 cm in front of its corresponding speaker. Second, to simulate an array, the three microphones were grouped together about 60 cm from the talkers.

The SNR results are shown in Table 1 and the apparent ITDs are shown in Fig. 4 for the four combinations of filter type and microphone placement. When the RMs were close to the talkers, the SIBO filters and MIBO filter both performed well, with the MIBO filter achieving a slightly higher SNR and better preserving interaural cues. When the RMs were farther from the talkers, the MIBO filter still preserved interaural cues but also reproduced more unwanted noise. The SIBO filters were better at rejecting noise, but crosstalk between sources caused distortion of the interaural cues.

6. CONCLUSIONS

The two adaptive binaural filtering methods proposed here have different strengths and weaknesses. The MIBO filter, which jointly processes all microphones, does not suffer from crosstalk and can enhance up to M sound sources without performing source separation, but it cannot apply different processing to different sound sources. It is most appropriate when the talkers of interest are close to one another. For example, a microphone array placed on a table can enhance speech from everyone around that table, even as talkers join and leave the conversation. The SIBO filters, which separately process each remote microphone, are attached to each talker, so they can track talkers as they move and apply different processing to each. However, they can suffer from crosstalk among nearby talkers. Both systems can adapt to motion without explicitly separating the sources or tracking their locations. The adaptive multitalker listening system combines the spatial cues of the earpiece microphones with the SNRs of the remote microphones, improving audibility while providing an immersive listening experience.

7. REFERENCES

[1] A. Boothroyd, "Hearing aid accessories for adults: The remote FM microphone," *Ear and Hearing*, vol. 25, no. 1, pp. 22–33, 2004.

[2] L. Thibodeau, "Comparison of speech recognition with adaptive digital and FM remote microphone hearing assistance technology by listeners who use hearing aids," *American Journal of Audiology*, vol. 23, no. 2, pp. 201–210, 2014.

[3] J. Wolfe, M. Morais Duke, E. Schafer, C. Jones, H. E. Mülder, A. John, and M. Hudson, "Evaluation of performance with an adaptive digital remote microphone system and a digital remote microphone audio-streaming accessory system," *American Journal of Audiology*, vol. 24, no. 3, pp. 440–450, 2015.

[4] G. A. Courtois, "Spatial hearing rendering in wireless microphone systems for binaural hearing aids," Ph.D. dissertation, EPFL, 2016.

[5] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, 2017.

[6] J. M. Kates, K. H. Arehart, R. K. Muralimanohar, and K. Sommerfeldt, "Externalization of remote microphone signals using a structural binaural model of the head and pinna," *The Journal of the Acoustical Society of America*, vol. 143, no. 5, pp. 2666–2677, 2018.

[7] A. Bertrand and M. Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 530435, 2009.

[8] 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.

[9] 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.

[10] D. Marquardt, "Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques," Ph.D. dissertation, Carl von Ossietzky University of Oldenburg, 2016.

[11] J. Szurley, A. Bertrand, B. Van Dijk, and M. Moonen, "Binaural noise cue preservation in a binaural noise reduction system with a remote microphone signal," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 952–966, 2016.

[12] N. Gößling and S. Doclo, "RTF-based binaural MVDR beamformer exploiting an external microphone in a diffuse noise field," in *ITG Symposium on Speech Communication*, 2018.

[13] R. Ali, G. Bernardi, T. van Waterschoot, and M. Moonen, "Methods of extending a generalized sidelobe canceller with external microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1349–1364, 2019.

[14] N. Gößling and S. Doclo, "RTF-steered binaural MVDR beamforming incorporating an external microphone for dynamic acoustic scenarios," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 416–420.

[15] N. Gößling, D. Marquardt, and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 462–476, 2020.

[16] R. Ali, T. Van Waterschoot, and M. Moonen, "Completing the RTF vector for an MVDR beamformer as applied to a local microphone array and an external microphone," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 211–215.

[17] N. Gößling and S. Doclo, "Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 146–150.

[18] N. Gößling, W. Middelberg, and S. Doclo, "RTF-steered binaural MVDR beamforming incorporating multiple external microphones," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 373–377.

[19] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2002.

[20] 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