ADAPTIVE CROSSTALK CANCELLATION AND SPATIALIZATION FOR DYNAMIC GROUP CONVERSATION ENHANCEMENT USING MOBILE AND WEARABLE DEVICES

Ryan M. Corey, Manan Mittal, Kanad Sarkar, and Andrew C. Singer

University of Illinois Urbana-Champaign

ABSTRACT

We propose a system to improve the intelligibility of group conversations in noisy environments, such as restaurants, by aggregating signals from the mobile and wearable devices of the participants. The proposed system uses a mobile device placed near each talker to capture a low-noise speech signal. Instead of muting inactive microphones, which can be distracting, adaptive crosstalk cancellation filters remove the speech of other users, including delayed auditory feedback of the listener's own speech. Next, adaptive spatialization filters process the low-noise signals to generate binaural outputs that match the spatial and spectral cues at the ears of each listener. The proposed system is demonstrated using recordings of three human subjects conversing with realistic movement.

***Index Terms*—** Acoustic sensor network, remote microphones, binaural processing, hearing aids, adaptive filters

1. INTRODUCTION

One of the most common complaints from people with hearing loss—and many normal-hearing people—is that it is difficult to follow group conversations in crowded, noisy restaurants. Conventional listening devices, such as hearing aids, work poorly in noisy environments because their microphones have the same signal-to-noise ratio (SNR) as the unaided ears. However, a network of several microphone-equipped devices spread around the group could achieve greater spatial diversity, providing better noise reduction performance than any single device. We propose a group conversation enhancement system that aggregates signals from the mobile and wearable devices of all conversation participants.

Wireless sensor networks and distributed microphone arrays have been proposed for spatial sound acquisition [1, 2]. For example, mobile phones near talkers can help fixed microphone arrays to transcribe a meeting [3]. A distributed beamforming algorithm for nonmoving hearing aid networks was proposed in [4]. Real-world human listening enhancement systems pose additional challenges: The system must operate in real time with imperceptible delay, generally several milliseconds [5]; it must preserve the spatial cues that humans use to localize and separate sounds, such as interaural time and level differences [6]; and it must contend with continuous motion of both sound sources and microphones [7].

Many modern listening devices can be paired with a wireless remote microphone (RM) accessory that transmits low-noise speech directly from a talker to the listener's ears [8], and low-latency wireless standards may soon allow smartphones to act as convient RMs. Well-placed RMs can greatly improve intelligibility of a single distant talker in noise [8–10], but current systems are unsuitable for group conversations because they support only one talker at a time and do not preserve interaural cues. Some researchers have proposed applying spatialization filters to RM signals based on the estimated direction of arrival [11, 12]. In [13], earpiece microphones are used as reference signals for an adaptive filter, eliminating the need for explicit source localization. This approach is also common in binaural beamforming systems, either using earpieces alone [14–16] or in combination with external microphones [17–20].

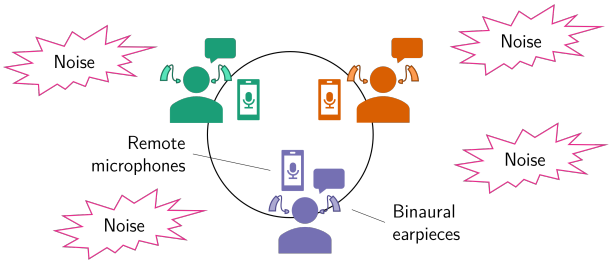

Fig. 1. A conversation enhancement system combines signals from mobile and wearable devices to reduce background noise and suppress unwanted echoes.

This work extends the adaptive spatialization system of [13] to address the challenges of close group conversations. Because the devices are closely spaced, there is significant crosstalk between microphones, which can cause distortion of spatial cues and delayed auditory feedback of the listener's own speech, which can be disturbing and impede speech production [21]. A common solution to crosstalk and own-speech echo is to disable all but one microphone at a time. However, frequent muting and unmuting of microphones can be distracting in a fast-paced group conversation and, if there is delay in the voice activity detector (VAD), can cause listeners to miss the first few syllables from a new talker. Instead, we propose using crosstalk cancellation filters to suppress echoes. This system provides a more natural listening experience in group conversations that may include frequent interruptions and double-talk.

A further challenge in group conversations is that users move constantly, causing acoustic channel parameters to change during and between utterances. The adaptive filters must therefore be updated continuously while in use [22]. In this work, we use stationary mobile devices as the remote signal sources because their acoustic channel parameters are more stable than those of wearable microphones, allowing the adaptive filters to converge more quickly as users move. Meanwhile, earpieces and other wearable devices that

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.

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

move with the users are helpful for VAD and as references for tracking interaural cues.

2. ENHANCEMENT SYSTEM

2.1. Signal model

Consider a group of $N \geq 2$ talkers and N remote microphones, as shown in Fig. 1, numbered such that RM n is placed near talker n for $n = 1, \ldots, N$. Let $s_n[t]$ be the discrete-time speech signal from talker n as captured by RM n. Consider a short time interval during which the acoustic channels from talkers to microphones can be considered time-invariant. Let $a_{\mathrm{r},m,n}[\tau]$ be the relative impulse response (RIR) describing the acoustic channel from talker n to RM m relative to RM n and let $z_{\mathrm{r},m}[t]$ be the ambient noise at RM m. Then the mixture $x_{\mathrm{r},m}[t]$ captured by RM m is given by

$$x_{\mathrm{r},m}[t] = \sum_{n=1}^{N} (a_{\mathrm{r},m,n} \star s_n)[t] + z_{\mathrm{r},m}[t], \quad m = 1, \ldots, N, \quad (1)$$

where $\star$ denotes linear convolution. Note that because each s_n is defined with respect to RM n, each $a_{\mathrm{r},n,n}[\tau]$ is the unit impulse $\delta[\tau]$. If each RM is placed close to its corresponding talker, then the RIRs of the other microphones should be well modeled by causal filters.

In addition to the remote microphones, each user wears a binaural listening device containing a left and a right microphone. Let $\mathbf{a}_{\mathrm{e},m,n}[\tau] = [a^{\mathrm{left}}_{\mathrm{e},m,n}[\tau], a^{\mathrm{right}}_{\mathrm{e},m,n}[\tau]]^T$ be the vector of RIRs from talker n to the left and right ears of listener m for $m, n = 1, \ldots, N$ and let $\mathbf{z}_{\mathrm{e},m}[t] \in \mathbb{R}^2$ be the ambient noise at those earpiece microphones. Then the mixture $\mathbf{x}_{\mathrm{e},m}[t] = [x^{\mathrm{left}}_{\mathrm{e},m}[t], x^{\mathrm{right}}_{\mathrm{e},m}[t]]^T$ captured by the earpieces of listener m is given by

$$\mathbf{x}_{\mathrm{e},m}[t] = \sum_{n=1}^{N} (\mathbf{a}_{\mathrm{e},m,n} \star s_n)[t] + \mathbf{z}_{\mathrm{e},m}[t], \quad m = 1, \ldots, N. \quad (2)$$

2.2. Binaural remixing

The objective of the conversation enhancement system is to remove ambient noise and own-speech echoes while preserving the speech of other talkers with correct spatial cues [23]. The desired output $\mathbf{y}_m[t] = [y^{\mathrm{left}}_m[t], y^{\mathrm{right}}_m[t]]^T$ for listener m is given by

$$\mathbf{y}_m[t] = \sum_{n \neq m} (\mathbf{a}_{\mathrm{e},m,n} \star s_n)[t], \quad m = 1, \ldots, N. \quad (3)$$

This binaural output may be amplified, equalized, compressed, or otherwise processed before it is presented to the listener. In some situations, the enhanced signals $\mathbf{y}_m$ may be mixed with the earpiece signals $\mathbf{x}_{\mathrm{e},m}$ to better preserve situational awareness [14]. Because the spatialized signals will be mixed with live signals—either electronically within the device or acoustically in the ear—each output must have near-zero delay relative to the live signal at the corresponding ear. If the listening device is strongly occluding, then it may be desirable to mix the listener's own speech into the output. This sidetone channel generally has different processing requirements than the speech of conversation partners, so for simplicity it is excluded from our analysis here.

The proposed processing system is shown in Fig. 2. It consists of two main stages: crosstalk cancellation (Sec. 3) to improve separation and suppress echoes of the listener's own speech, and spatialization (Sec. 4) to preserve realistic spatial and acoustic cues. It also uses voice activity detection (not pictured) to control adaptation.

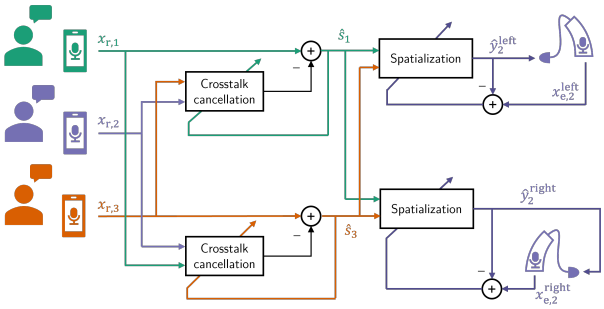

Fig. 2. The proposed enhancement system, shown from the perspective of user 2, uses adaptive crosstalk cancellation and spatialization filters to combine signals from mobile devices.

2.3. Voice activity detection

Both crosstalk suppression and spatialization rely on accurate VAD to determine which users are speaking. Wearable devices are attractive for VAD because they are physically attached to users. Earpieces can use hardware features such as bone-conduction microphones to perform reliable VAD even in strong noise [24, 25]. The design of such VADs is beyond the scope of this work. In our experiments, we compared two wearable VAD implementations: a more-reliable VAD using headset microphones and a less-reliable VAD using lapel microphones. Speech was detected using a multivariate Gaussian likelihood ratio test in the short-time Fourier transform domain. Second-order statistics were estimated using training data and time-frequency log-likelihood ratios were averaged from 0 to 1 kHz in half-second time windows. The resulting statistics were compared against a manually tuned threshold. The headset and lapel VADs were 90% and 82% accurate, respectively, in a one-at-a-time conversation with moving talkers.

3. CROSSTALK CANCELLATION

In a group conversation, the talkers are close together so that each microphone captures speech from all users. Instead of muting the microphones of users who are not speaking, which could be distracting and cause listeners to miss parts of the conversation, the proposed system keeps all microphones on at all times, but uses adaptive cancellation filters to remove crosstalk. The processed microphone signals $\hat{s}_n[t]$ are given by

$$\hat{s}_n[t] = \begin{cases} x_{\mathrm{r},n}[t], & \text{if user } n \text{ is talking,} \\ x_{\mathrm{r},n}[t] - \sum_{m \neq n} (u_{n,m} \star x_{\mathrm{r},m})[t], & \text{otherwise,} \end{cases} \quad (4)$$

for $n = 1, \ldots, N$, where each $u_{n,m}$ is a finite-impulse-response filter. In a low-noise environment, each $u_{n,m}$ models the corresponding RIR $a_{\mathrm{r},n,m}$. The filter must be disabled when user n is speaking to prevent target signal cancellation; merely pausing adaptation was found to be ineffective, presumably due to motion. Note that the filter cancelling source m at microphone n remains active even when user m is quiet in order to avoid echoes in case of false negatives from the VAD. When user m is quiet, the filter will help to suppress noise from the direction of that user.

3.1. Filter adaptation

Because human talkers move frequently, the echo cancellation filters $u_{n,m}$ are updated continuously when they are active. When user n is quiet, $\hat{s}_n[t]$ is a linear prediction error signal with $x_{\mathrm{r},n}[t]$ as the reference signal. The filters are adapted to solve the echo cancellation optimization problem

$$\min_{\{u_{n,m}\}_{m\neq n}} \mathbb{E}\left[|\hat{s}_n[t]|^2\right], \tag{5}$$

where $\mathbb{E}$ denotes statistical expectation. In our experiments, we solve (5) iteratively using the normalized least-mean-squares (NLMS) algorithm with first-order prewhitening [26].

3.2. Distortion effects

It is instructive to compare the behavior of the cancellation system to that of a muting system with an imperfect VAD. Consider $N = 2$ users and zero ambient noise. When user 1 is speaking and user 2 is quiet, the cancellation filter converges to the Wiener solution $u_{2,1}[\tau] = a_{2,1}[\tau]$ so that user 1 is perfectly cancelled and $\hat{s}_2[t] = 0$, just as in a muting system. Suppose that user 2 interrupts and the VAD does not immediately detect the interruption. In the muting system, user 2's speech would be inaudible. In the proposed system, the output immediately following the interruption is

$$\hat{s}_2[t] = x_{\mathrm{r},2}[t] - (u_{2,1} \star x_{\mathrm{r},1})[t] \tag{6}$$

$$= ((a_{2,1} - a_{2,1}) \star s_1)[t] + ((\delta - a_{2,1} \star a_{1,2}) \star s_2)[t] \tag{7}$$

$$= ((\delta - a_{2,1} \star a_{1,2}) \star s_2)[t]. \tag{8}$$

The speech from user 1 is still cancelled correctly and the speech from talker 2 is audible but distorted. The severity of the distortion depends on the crosstalk channels between microphones. With well-positioned directional microphones, the RIRs $a_{2,1}$ and $a_{1,2}$ should both have magnitude responses much smaller than 1 so that the distortion has little effect on s_2. In a system with strong crosstalk, such as a compact microphone array, the proposed system may cause strong distortion; in that case, a linearly constrained beamformer may be more appropriate [27].

4. BINAURAL SPATIALIZATION

The spatialization filters process the low-noise source estimates $\hat{s}_1[t], \ldots, \hat{s}_N[t]$ to match the spatial and acoustic cues at the ears of each listener, including interaural time and level differences, spectral shaping, and early reflections. The binaural output mixture for listener m is given by

$$\hat{\mathbf{y}}_m[t] = \sum_{n\neq m} (\mathbf{w}_{m,n} \star \hat{s}_n)[t], \tag{9}$$

where each $\mathbf{w}_{m,n}[\tau] \in \mathbb{R}^2$ is a casual finite-impulse-response filter. The filters for each listener m are updated to solve

$$\min_{\{\mathbf{w}_{m,n}\}_{n\neq m}} \mathbb{E}\left[\|\mathbf{x}_{\mathrm{e},m}[t] - \hat{\mathbf{y}}_m[t]\|^2\right]. \tag{10}$$

In our experiments, this cost function is minimized iteratively using the NLMS algorithm with first-order prewhitening.

Unlike the crosstalk cancellation filters, the spatialization filters are always active, even when their respective users are not speaking. However, each $\mathbf{w}_{m,n}$ is updated only while user n is speaking. If the filters were updated continuously, then they would amplify nearby noise sources during speech pauses.

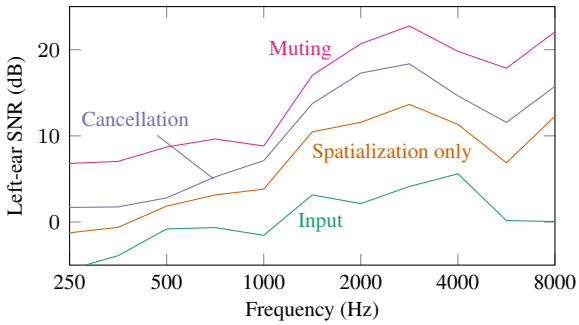

Fig. 3. Noise reduction performance at the listener's left earpiece (higher is better).

4.1. Spatializing multiple talkers

When multiple users are speaking simultaneously, the spatialization filter coefficients are updated jointly. They therefore act as a multiple-input, binaural-output (MIBO) filter that maps from input mixtures to output mixtures. It was shown in [13] that an N-input MIBO filter can preserve the spatial cues of up to N sources. MIBO filters do not require that the sources be separated and are unaffected by residual crosstalk, making them well suited for closely spaced talkers. However, they do rely on accurate VAD: False negatives would cause them to blend the cues of multiple active talkers, while false positives would cause them to amplify a nearby noise source in place of the missing talker. The crosstalk cancellation stage therefore helps to mitigate spatial-cue distortion with an unreliable VAD.

5. EXPERIMENTS

The proposed conversation enhancement system was evaluated with three live human subjects seated around a table in an acoustically treated laboratory ($T_{60} \approx 150$ ms). Each subject wore an omnidirectional lavalier microphone behind each ear to simulate behind-the-ear hearing aids. Another such microphone was affixed to the table in front of each subject to simulate a mobile phone. Each subject also wore lapel and headset microphones, which were used only for VAD. Noise was produced by a set of six loudspeakers playing clips from the VCTK speech corpus [28].

To simulate a group conversation, the subjects took turns reading from a script for 60 seconds. In one recording, the subjects looked straight ahead and tried not to move. In another, they turned to look at each other and gestured while speaking. A third recording with moderate motion was used for VAD training. To quantify the input and output SNR of the system, the noise was recorded separately and added to the live speech recordings; the noise was therefore recorded with a different motion pattern than the live speech. Likewise, double- and triple-talk mixtures were simulated by combining separate recordings. All microphones were sampled synchronously at 48 kHz and processed at 16 kHz. The results shown here are for the left-ear output of one user's listening device.

5.1. Noise reduction

The SNR improvement of the proposed conversation enhancement system is shown in Fig. 3. Because the system does not perform beamforming or other noise reduction processing, the SNR improvement depends strongly on the placement of the remote microphones. The smartphone-like tabletop microphones had higher input SNR

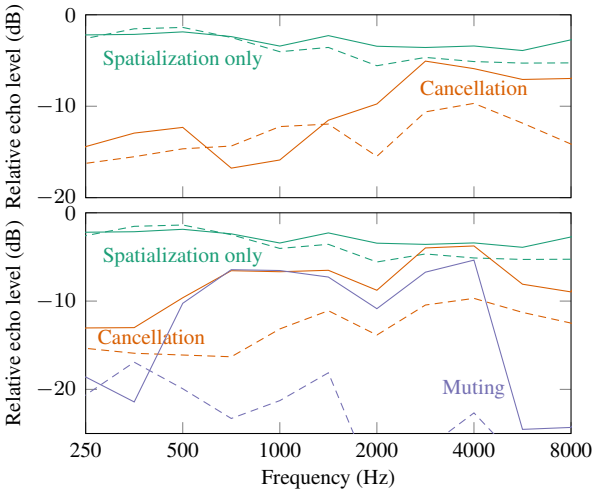

Fig. 4. Own-speech echo suppression performance at the listener's left earpiece (lower is better). Dashed curves show the nonmoving experiment and solid curves show the experiment with moving subjects. Top: Headset mic VAD. Bottom: Lapel mic VAD.

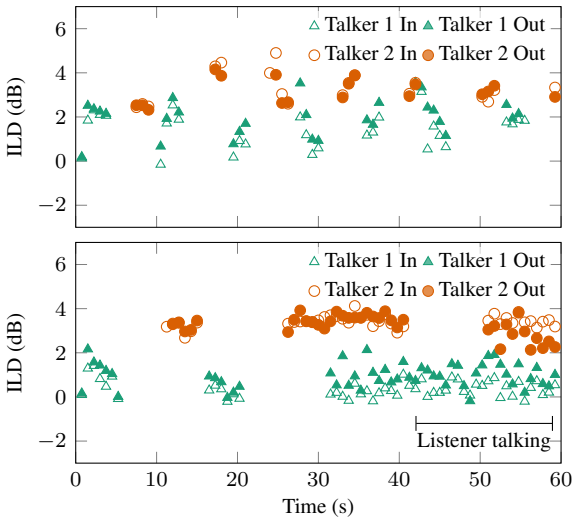

Fig. 5. High-frequency interaural level differences of other talkers at the listener's ears. Top: Subjects take turns speaking while moving to face each other. Bottom: Simulated double- and triple-talk with subjects facing forward.

and lower crosstalk than the earpiece and lapel microphones, especially at high frequencies. Using the MIBO spatialization filters of [13] without crosstalk cancellation improves the high-frequency SNR at the left ear by up to 10 dB. The crosstalk filter helps to further suppress noise when nearby users are not speaking, providing another 2–5 dB benefit to SNR. A conventional remote microphone system that mutes all but one microphone achieves the best average output SNR, but is too distracting to be practical.

The plot shows performance using the headset-based VAD for nonmoving subjects. The results for other experimental conditions were similar and so are not reported; VAD accuracy and user motion appear to have little effect on ambient noise reduction.

5.2. Echo suppression

Figure 4 shows the echo reduction performance of the system for the listener's own speech. The curves show the echo level relative to the direct acoustic path to the earpiece. Because the users are seated close together, the own-speech echo in the baseline spatialization-only system is just 2–5 dB weaker than the direct path. The crosstalk cancellation filters were able to suppress own-speech echoes by up to 15 dB more than the baseline system, but their performance depends on talker motion and on VAD accuracy. The residual echo levels for moving talkers (solid curves) are higher than those for stationary talkers (dashed curves), especially at high frequencies for which source positions may suddenly change by multiple acoustic wavelengths. Echo suppression was also worse for the less-reliable lapel-based VAD (bottom plot) compared to the more-reliable headset-based VAD (top plot). The performance of the muting system depends entirely upon VAD performance: With the reliable VAD, it removed virtually all echoes; with the unreliable VAD, it performed little better than the cancellation system at most frequencies in the motion experiment.

5.3. Spatial cue preservation

We can evaluate spatialization performance by comparing the interaural cues of the system output to the cues of the noise-free speech signals at the ears. Figure 5 shows the input and output interaural level differences (ILD) of speech from the two other talkers at the ears of the listener. The ILDs are averaged over 1.5 sec windows from 1–8 kHz and color-coded to show the active talker(s). When the talkers take turns (top plot), only one spatialization filter adapts at a time. The output cues closely match the input cues, even as the listener turns their head. When both other talkers speak simultaneously (bottom plot, 30–42 s), two filters adapt jointly, preserving the spatial cues of both sources despite residual crosstalk. When the listener and talker(s) speak simultaneously (bottom plot, 42–60 s), crosstalk cancellation is disabled and the spatialization filters are unable to distinguish the listener's own speech from that of the other talkers, so their spatial cues are blended. Thus, a user will have trouble localizing conversation partners while interrupting them.

6. CONCLUSIONS

The proposed conversation enhancement system can reduce ambient noise and delayed auditory feedback without distracting the listener by frequently switching between microphones. Because the system does not perform beamforming or other noise reduction, and because the crosstalk cancellation filters use the remote microphones as references for adaptation, its performance depends strongly on the noise and crosstalk levels at the remote microphones. The experiments presented here showed meaningful enhancement using simple omnidirectional microphones placed on a table. In larger groups or with stronger noise and reverberation, it may be necessary to use directional microphones, arrays, or active noise reduction techniques to improve RM input SNR. Furthermore, the adaptive filters rely strongly on VAD to prevent spectral distortion and target signal leakage. Further research is required to develop reliable VAD systems for noisy group conversations, for example using special hardware features of wearable devices. With well-performing wearable and mobile devices and low-latency wireless connections, the proposed system can substantially reduce noise in the most challenging listening environments using devices that users already have with them.

7. REFERENCES

[1] Alexander Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, 2011, pp. 1–6.

[2] Maja Taseska and Emanuël AP Habets, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1195–1207, 2014.

[3] Shoko Araki, Nobutaka Ono, Keisuke Kinoshita, and Marc Delcroix, "Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based MVDR beamformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5694–5698.

[4] Alexander Bertrand and Marc Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 530435, 2009.

[5] Jeremy Agnew and Jeffrey M Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[6] Jens Blauert, *Spatial hearing: The psychophysics of human sound localization*, MIT press, 1997.

[7] Ryan M Corey and Andrew C Singer, "Motion-tolerant beamforming with deformable microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

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

[9] Linda 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.

[10] Jace Wolfe, Mila Morais Duke, Erin Schafer, Christine Jones, Hans E Mülder, Andrew John, and Mary 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.

[11] James M Kates, Kathryn H Arehart, Ramesh Kumar Muralimanohar, and Kristin 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.

[12] James M Kates, Kathryn H Arehart, and Lewis O Harvey, "Integrating a remote microphone with hearing-aid processing," *The Journal of the Acoustical Society of America*, vol. 145, no. 6, pp. 3551–3566, 2019.

[13] Ryan M. Corey and Andrew C. Singer, "Adaptive binaural filtering for a multiple-talker listening system using remote and on-ear microphones," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

[14] Bram Cornelis, Simon Doclo, Tim Van dan Bogaert, Marc Moonen, and Jan 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.

[15] Elior Hadad, Daniel Marquardt, Simon Doclo, and Sharon 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.

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

[17] Joseph Szurley, Alexander Bertrand, Bas Van Dijk, and Marc 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.

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

[19] Randall Ali, Giuliano Bernardi, Toon van Waterschoot, and Marc 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.

[20] Nico Gößling and Simon 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.

[21] Michael A Stone and Brian CJ Moore, "Tolerable hearing aid delays. II. Estimation of limits imposed during speech production," *Ear and Hearing*, vol. 23, no. 4, pp. 325–338, 2002.

[22] Simon Haykin, *Adaptive Filter Theory*, Prentice Hall, 2002.

[23] Ryan M. Corey and Andrew C. Singer, "Binaural audio source remixing with microphone array listening devices," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[24] Mingzhe Zhu, Hongbing Ji, Falong Luo, and Wei Chen, "A robust speech enhancement scheme on the basis of bone-conductive microphones," in *International Workshop on Signal Design and Its Applications in Communications (IWSDA)*, 2007.

[25] Heming Wang, Xueliang Zhang, and DeLiang Wang, "Attention-based fusion for bone-conducted and air-conducted speech enhancement in the complex domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[26] Eberhard Hänsler and Gerhard Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Wiley, 2005.

[27] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, *Audio Source Separation and Speech Enhancement*, Wiley, 2018.

[28] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017.