AUDIO SIGNAL PROCESSING FOR TELEPRESENCE BASED ON WEARABLE ARRAY IN NOISY AND DYNAMIC SCENES

Hanan Beit-On[1], Moti Lugasi[1], Lior Madmoni[1], Anjali Menon[2], Anurag Kumar[2], Jacob Donley[2], Vladimir Tourbabin[2] and Boaz Rafaely[1]

[1] School of Electrical and Computer Engineering, Ben-Gurion University of the Negev
[2] Reality Labs Research @ Meta

ABSTRACT

Telepresence for virtual meetings has gained interest due to recent travel limitations and the new reality of working from home. However, current literature supporting real-world microphone arrays for realistic telepresence in audio is very limited. This paper investigates a scenario of a distant participant joining virtually a meeting between two dymanic participants. The audio signal processing chain (i) starts by recording using an array mounted on glasses, (ii) with initial processing providing direction-of-arrival estimation of a desired speaker using a direct-path dominance test robust to reverberation, combined with speaker separation for improved dynamic localization, (iii) followed by speech enhancement against interfering speakers and noise, (iv) and ends with applying binaural signal matching for headphone listening. This paper compares model-based processing to learning-based processing in both noisy and dynamic scenarios, and presents a novel processing using data from a real wearable array, studied by simulation and a listening test.

***Index Terms*—** Telepresence, spatial enhancement, speaker tracking, binaural signal matching

1. INTRODUCTION

Telepresence has recently gained tremendous interest due to travel limitations and the popularity of working from home. Acoustic telepresence allow a distant participant to virtually join a remote meeting. This typically requires a microphone array to capture the sound in the meeting room, and headphones to playback a binaural signal at the remote location, reproduced from the array measurement using spatial audio signal processing [1, 2]. Meeting rooms may include noise or interfering speakers, and so signal enhancement may be required. Furthermore, meeting participants and wearable arrays may change their position, imposing a challenge on the processing and playback.

Current methods for acoustic telepresence typically employ binaural arrays [1, 3] or spherical arrays to encode Ambisonics signals [4, 5], which facilitates high-quality binaural reproduction. Signal enhancement may also be incorporated [6, 7, 8] to suppress interfering speakers or noise. However, the performance of these deteriorate under conditions of multiple moving speakers and/or a moving array [9]. Moreover, binaural and spherical arrays may not be available with handheld devices or wearable arrays. In summary, current methods for acoustic telepresence have the following limitations: (i) a microphone array of a general configuration, e.g., a wearable array, cannot be incorporated, (ii) dynamic and noisy scenes lead to significant performance degradation.

This paper investigates solutions for acoustic telepresence that aim to overcome the current limitations. The proposed solution consists of three stages: (i) acoustic scene analysis, that is, the estimation of speaker directions under reverberation using a direct-path dominance (DPD) test [10], (ii) signal enhancement based on desired source and array transfer function estimation; and (iii) binaural signal reproduction using the binaural signal matching method [11].

The proposed solution is investigated by simulations based on measurements from a real glasses-mounted array [12]. Two alternative solutions are studied: the first is model-based, while the second incorporates a learning-based speech separation method [13] to support consistent enhancement under dynamic conditions. Finally, a listening test quantifies the subjective performance. In summary, this work offers two main contributions: (i) the proposal and investigation of an end-to-end acoustic telepresence system based on data from real glasses-mounted array, (ii) an investigation of the enhanced performance due to learning for dynamic scenes with interfering speakers.

2. SIGNAL MODEL AND SYSTEM GOAL

Consider a meeting room with K speakers. The sound field in the room is assumed to be represented by L far-field sources arriving from directions $\{\psi_l\}_{l=1}^{L}$ with source signals $\{s_l(f)\}_{l=1}^{L}$, where f denotes frequency. Source $s_l(f)$ can represent the direct-sound of an actual source, or a reflection from room boundaries. The sound field is captured by a microphone array with an arbitrary configuration, having I mi-

Thanks to Reality Labs Research @ Meta for funding.

crophones. The model of the microphone signals is given by:

$$\mathbf{x}(f) = \underbrace{\mathbf{V}_d(f)\mathbf{s}_d(f)}_{\mathbf{d}(f)} + \underbrace{\mathbf{V}_u(f)\mathbf{s}_u(f) + \mathbf{n}(f)}_{\mathbf{u}(f)}, \qquad (1)$$

where $\mathbf{x}(f) = [x_1(f), ..., x_I(f)]^T$ is the microphone signal vector, $\mathbf{s}_d(f) = [s_1(f), ..., s_{L'}(f)]^T$ and $\mathbf{s}_u(f) = [s_{L'+1}(f), ..., s_L(f)]^T$ are source signal vectors containing the direct-path signal and reflections of the desired and undesired speaker signals, respectively, $\mathbf{V}_d(f) = [\mathbf{v}(f, \psi_1), ..., \mathbf{v}(f, \psi_{L'})]$ and $\mathbf{V}_u(f) = [\mathbf{v}(f, \psi_{L'+1}), ..., \mathbf{v}(f, \psi_L)]$ are steering matrices, with their l'th column $\mathbf{v}(f, \psi_l)$ representing the array steering vector from direction ψ_l, and $\mathbf{n}(f)$ is a noise term which represents sensor noise, model error, and late reverberation.

The goal is to produce the binaural signal $p_{l,r}(f)$ that would have been received by a listener if they were in the room at the array position, while suppressing the undesired part $\mathbf{u}(f)$. The desired output can be formulated as:

$$p_{l,r}(f) = \mathbf{h}(f)^T \mathbf{s}_d(f), \qquad (2)$$

where $\mathbf{h}(f) = [h_{l,r}(\psi_1), ..., h_{l,r}(\psi_{L'})]$ contains the head-related transfer functions (HRTFs) corresponding to L' directions of the desired source signals.

3. PROPOSED SYSTEM

This section presents a three-stage system for estimating $p_{l,r}(f)$ from the microphone signals. Two versions of the system, which differ by their level of learning, are presented. The first system is primarily model-based and incorporates spectral and spatial learning, and the second incorporates a deep-learning speaker separation method in the scene analysis stage to support dynamic scenes. Both versions are presented in Fig. 1, and are described next.

3.1. System 1 - model-based processing

Scene analysis: In the first stage, speakers' DOAs are estimated using a DPD test-based method robust to reverberation. This family of methods operates in the time-frequency (TF) domain and overcomes reverberation by using only TF bins in which the direct-path is dominant for DOA estimation. In [10], the set of direct-path bins is determined by:

$$\mathcal{A}_{\text{DPD}} = \{(t, f) \,|\, \max\left(S_{t,f}\left(\psi\right)\right) > \mathcal{TH}\}, \qquad (3)$$

where t and f denote the time and frequency indices, $S_{t,f}(\psi)$ denotes a directional spectrum (e.g. MUSIC, space-domain distance (SDD)[10]) computed at bin (t, f), and $\mathcal{TH}$ is a predefined threshold. Next, a single DOA is estimated at each of the selected bins by the direction ψ that maximizes $S_{t,f}(\psi)$. Finally, the DOA estimates from the selected bins are clustered using k-means and the final DOA estimates, denoted by $\{\hat{\psi}_i\}_{i=1}^K$, are given by the mean of each cluster. Each estimated DOA is labeled as desired or undesired by the user.

Signal enhancement: In the second stage, the desired signal, $\mathbf{d}(f)$, assuming a single desired speaker, is estimated. For multiple desired speakers, the below procedure can be repeated for each of the speakers while defining $\mathbf{d}(f)$ and $\mathbf{u}(f)$ accordingly, followed by summing the contribution of each desired speaker. The DOA estimates $\{\hat{\psi}_i\}_{i=1}^K$ are used to form a minimum-variance distortionless response (MVDR) beamformer which is used for estimating the direct-sound from the desired source signal as follows:

$$\hat{s}(f) = \left(\frac{\mathbf{P}(f)^{-1}\mathbf{v}(f, \hat{\psi}_d)}{\mathbf{v}(f, \hat{\psi}_d)^H \mathbf{P}(f)^{-1}\mathbf{v}(f, \hat{\psi}_d)} \right)^H \mathbf{x}(f), \qquad (4)$$

where $\mathbf{v}(f, \hat{\psi}_d)$ is the array's free-field steering vector corresponding to the estimated DOA of the desired speaker $\hat{\psi}_d$, and $\mathbf{P}(f) = E[\mathbf{u}(f)\mathbf{u}(f)^H]$, with $E[\cdot]$ and H denoting the statistical expectation and the conjugate transpose, respectively. In this paper, we assume that the matrix $\mathbf{P}(f)$ is known. In practice, this may not be the case and an estimation of $\mathbf{P}(f)$ may be required. Then, the acoustic transfer function (ATF) between the desired speaker and the array $\mathbf{a}(f)$ is estimated using an H1 estimator as follows [14]:

$$\hat{\mathbf{a}}(f) = \frac{S_{\hat{s}\mathbf{x}}(f)}{S_{\hat{s}\hat{s}}(f)}, \qquad (5)$$

where $S_{\hat{s}\mathbf{x}}(f) = E[\hat{s}(f)\mathbf{x}(f)^*]$ and $S_{\hat{s}\hat{s}}(f) = E[\hat{s}(f)\hat{s}(f)^*]$. Practically, $S_{\hat{s}\mathbf{x}}(f)$ and $S_{\hat{s}\hat{s}}(f)$ are estimated in the STFT domain by replacing the expectation operation by averaging over time frames. Finally, $\mathbf{d}(f)$ is estimated by:

$$\hat{\mathbf{d}}(f) = \hat{s}(f) \cdot \hat{\mathbf{a}}(f). \qquad (6)$$

Binaural reproduction: In the final stage, $p_{l,r}(f)$ is estimated using the binaural signal matching (BSM) method presented in [11]. With this method, $p_{l,r}(f)$ are estimated from the desired microphone signal $\hat{\mathbf{d}}(f)$ as follows:

$$\hat{p}_{l,r}\left(f\right) = \mathbf{w}_{l,r}^H\left(f\right) \hat{\mathbf{d}}\left(f\right), \qquad (7)$$

where $\mathbf{w}_{l,r}(f)$ are left and right filters applied to $\hat{\mathbf{d}}(f)$. $\mathbf{w}_{l,r}(f)$ are computed as follows:

$$\mathbf{w}_{l,r}(f) = \arg\min_{\mathbf{w}} E\left[\left|\mathbf{w}^H \mathbf{d}(f) - p_{l,r}(f)\right|^2\right], \qquad (8)$$

where $p_{l,r}(f)$ is the desired binaural signal presented in (2). Since $p_{l,r}(f)$ is unknown, the minimization problem in (8) is solved for a binaural signal $p_{l,r}(f)$ and a microphone signals $\mathbf{d}(f)$ generated by Q independent and uniformly distributed sources (see [11] for further details).

The system presented above is most suitable for static scenarios. To support dynamic conditions, multiple-speaker tracking algorithms, such as [15, 16] are typically required.

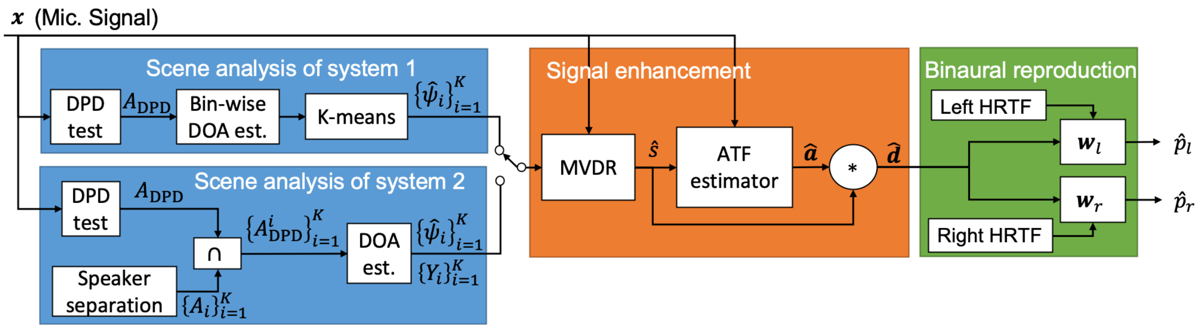

Fig. 1. Block diagram of the proposed acoustic telepresence systems.

However, since these methods rely on DOA estimates alone, their performance may drop following inactive speech periods (due to speakers moving while silent) when DOA estimates are not available [9].

3.2. System 2 - extended speaker learning

To overcome the limitation of current tracking methods, a speech separation network [13] that separates speakers using masking in the STFT domain is incorporated in the first stage. Speakers TF masks, estimated by the network, are employed to associate each TF bin (DOA estimate) with a specific speaker. Unlike the estimate-speakers association provided by current speaker tracking methods [15, 16], the proposed association is based on spectral information; thus, it may not be affected by inactive speech periods. The input of the speaker separation network is the spectrogram of a single microphone signal, and the outputs are TF masks $\mathbf{M}_i(t,f)$, $i = 1, ..., K$, indicating the dominance of the i'th speaker at each bin. Each entry of $\mathbf{M}_i(t,f)$ takes values in the range $[0,1]$, where higher values signify higher dominance of a speaker. The set of TF bins in which the i'th speaker is dominant is determined as follows:

$$\mathcal{A}_i = \{(t,f)|\mathbf{M}_i(t,f) > \mathcal{TH}_{sep}\}, \tag{9}$$

where $\mathcal{TH}_{sep}$ is a selected threshold. Combining (3) and (9), the set of direct-path bins of a specific speaker is formed by:

$$\mathcal{A}_{\text{DPD}}^i = \mathcal{A}_i \cap \mathcal{A}_{\text{DPD}}. \tag{10}$$

Given the sets $\{\mathcal{A}_{\text{DPD}}^i\}_{i=1}^K$, the DOA of the i'th speaker is estimated by the most frequent DOA estimate in the i'th set. The obtained DOA is labeled accordingly with label $\{Y_i\}_{i=1}^K$, i.e. desired or undesired.

4. SIMULATION STUDY

A simulation study was conducted to evaluate the performance of each of the proposed system stages in a scenario that includes a wearable array, desired and undesired speakers, and noise. In particular, the quality of the reproduced binaural signal and the effect of estimate-speaker association on signal enhancement were investigated.

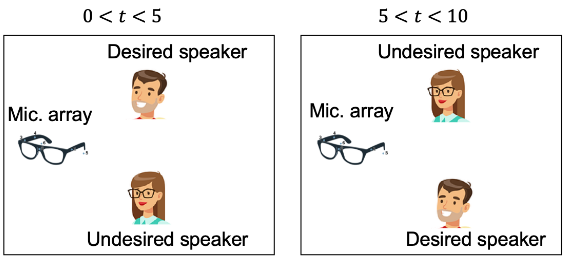

Fig. 2. Simulated scenario of speakers switching positions.

Setup: A meeting scene that included two speakers, desired and undesired, in a reverberant room and a microphone array mounted on glasses was simulated. To examine the proposed association, we simulated speakers position exchange at time instance $t = 5$. This positions exchange challenges the system to adapt to new and instant speaker positions, which may occur following inactive speech segments. The simulated scenario is presented in Fig. 2. The room size was $10 \times 15 \times 3$ m with an approximate reverberation time of 0.5 seconds. The array, illustrated in Fig. 3, was used for recording in the recent EasyCom dataset [12]. In this study, only microphones 1-4 were used for processing, where the signal at microphones 5 and 6 served as a reference binaural signal, with the steering vectors to microphone 5 and 6 serving as HRTFs. The array steering vectors, provided as part of the dataset from 1020 directions, were downsampled from 48 to 16 kHz. The array was positioned at $(x, y, z) = [3, 5, 1.7]$ m and the speakers were placed 2.5 m away from the array, at the same height, and at an azimuth angle of $\pm 45°$ relative to the array coordinate system. The impulse responses from the speakers' positions to the array were simulated using the image method [17]. 10 second male and female speech segments, from the wsj0-2mix dataset [18], sampled at 16 kHz, were convolved with the room impulse responses corresponding to the positions of the desired and undesired speakers. Speakers position exchange was simulated by switching the impulse responses.

Methodology: The proposed systems were implemented with the following STFT parameters: Hann window of 256 samples, 192 samples overlap and FFT length of 256 samples. The local SDD spectrum [10] was calculated for a grid of directions in the frontal-horizontal plane with 31 azimuth angles ranging between $-90°$ and $90°$. The set of direct path bins was obtained according to (3), with a frequency dependent threshold, which was designed to select the top 5% of the bins. For system 1, a k-means algorithm with 2 means was employed for clustering. For system 2, an Asteroid [19] implementation of the speech separation method presented in [13] was employed. The network was trained with the clean speech mixtures from the wsj0-2mix dataset [18]. The training data did not include speech samples or speakers that were used for the evaluation. The sets $\{\mathcal{A}_i\}_{i=1}^{2}$ were generated according to (9) with $\mathcal{TH}_{sep}$ of 0.8. The matrix $\mathbf{P}(f)$ in (4) was computed in advance for each of the allowed speaker positions from a 110 seconds length microphone signal and using a 1-second STFT window. The undesired speaker DOA estimate, $\hat{\psi}_u$, was used for selecting the matrix $\mathbf{P}(f)$ to be substituted in (4). Note that, for a general setup, calculating $\mathbf{P}(f)$ in advance may not be practical, and an alternative way for estimating $\mathbf{P}(f)$ may be required. $S_{\hat{s}\mathbf{x}}(f)$ and $S_{\hat{s}\hat{s}}(f)$ were also computed from an appropriate 110 seconds length microphone signal using a 1-second STFT window. The BSM filters $\mathbf{w}_{l,r}(f)$ in (8) were computed assuming that the sound field is composed of 1020 independent sources from the 1020 directions, and magnitude least squares (MagLS) processing was applied to the HRTFs, as suggested in [20].

Results: Table 1 presents azimuth estimates and normalized mean square error (NMSE) in estimating $\mathbf{d}$. Table (1) shows that both systems perform well for $0 < t < 5$ and that only system 2, which employed speaker separation, manages to track the position exchange, resulting in low NMSE for $t > 5$. This result suggests that the proposed method can handle dynamic scenes with inactive moving speakers.

5. LISTENING TESTS

Two listening tests were conducted to study the system's enhancement and binaural reproduction stages following the multiple stimuli with hidden reference and anchor (MUSHRA) protocol. 2 female and 7 male subjects with previous experience participated in both tests. Participants rated the similarity to the reference with respect to the overall quality.

The first test examines the reproduction stage by comparing the performance of the BSM method with the Ambisonics reproduction. The microphone signal processed by the BSM and the Ambisonics signals were generated using the simulation scenario described above and included only the desired signal part before the position exchange. A high order ambisonics (HOA) signal of order N = 9, was used to render the reference signal. Two other test signals were rendered using first order Ambisonics (FOA) signals, with and without MagLS. The upper plot in Fig. 4, which depicts the results, shows that the BSM was rated close to the FOA with MagLS and the reference, and much better than FOA, indicating that the proposed method with the 4-microphone array achieves good-quality binaural signals. The second test examines the enhancement stage. All four test signals were generated using the simulation described above after position exchange. The enhanced binaural signals obtained with the proposed systems were compared with a reference and anchor signals that were generated by applying the BSM to the clean desired and the unprocessed microphone signals, respectively. The test result presented in the lower plot of Fig. 4 shows that system 2 operates well while system 1 fails. This is due to wrong association, which resulted in enhancing the undesired speaker.

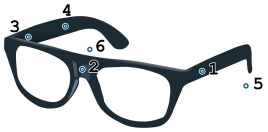

Fig. 3. Illustration of the microphone array mounted on glasses with the approximate microphone positions.

Azimuth	$0 < t < 5$		$5 < t < 10$	
	Desired speaker	Undesired speaker	Desired speaker	Undesired speaker
True	$45°$	$-45°$	$-45°$	$45°$
System 1	$43.1°$	$-45.2°$	$44°$	$-44.3°$
System 2	$42°$	$-42°$	$-42°$	$42°$

NMSE	$0 < t < 5$	$5 < t < 10$
Unprocessed	-6.1 dB	-2 dB
System 1	-13.4 dB	1.6 dB
System 2	-13.4 dB	-15 dB

Table 1. Azimuth estimates (upper table) and NMSE in estimating $\mathbf{d}(f)$ (lower table) with each system.

6. CONCLUSIONS

A practical acoutic telepresence system may be required to support wearable arrays, suppress interfering speakers and noise, and handle speakers and array movements. In this work, a three-stage system that addresses these challenges is proposed and investigated. The simulation study and listening tests demonstrate that the proposed approach applied to the glasses array may be useful under realistic conditions.

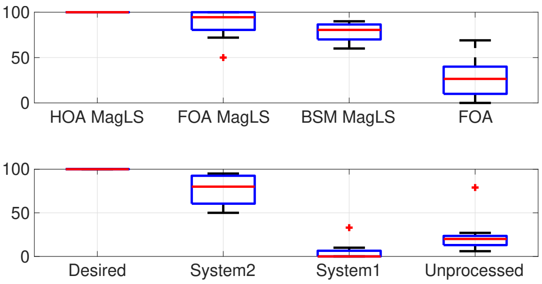

Fig. 4. Listening tests results.

7. REFERENCES

[1] F. Keyrouz and K. Diepold, "Binaural source localization and spatial audio reproduction for telepresence applications," *PRESENCE: Teleoperators and Virtual Environments*, vol. 16, no. 5, pp. 509–522, 2007.

[2] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich, "A 3d ambisonic based binaural sound reproduction system," in *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, Audio Engineering Society, 2003.

[3] D. Hammershøi and H. Møller, "Binaural technique—basic methods for recording, synthesis, and reproduction," *Communication acoustics*, pp. 223–254, 2005.

[4] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. 11, pp. 1004–1025, 2005.

[5] D. L. Alon, J. Sheaffer, and B. Rafaely, "Robust plane-wave decomposition of spherical microphone array recordings for binaural sound reproduction," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1925–1926, 2015.

[6] M. Lugasi and B. Rafaely, "Speech enhancement using masking for binaural reproduction of ambisonics signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1767–1777, 2020.

[7] N. R. Shabtai and B. Rafaely, "Spherical array beamforming for binaural sound reproduction," in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, pp. 1–5, IEEE, 2012.

[8] C. Borrelli, A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "A denoising methodology for higher order ambisonics recordings," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 451–455, IEEE, 2018.

[9] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The locata challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[10] V. Tourbabin, D. L. Alon, and R. Mehra, "Space domain-based selection of direct-sound bins in the context of improved robustness to reverberation in direction of arrival estimation," in *Proc. 11th European Congress and Exposition on Noise Control Engineering (EURONOISE18)*, pp. 2589–2596, 2018.

[11] L. Madmoni, J. Donley, V. Tourbabin, and B. Rafaely, "Beamforming-based binaural reproduction by matching of binaural signals," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society, 2020.

[12] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," 2021.

[13] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 61–65, IEEE, 2017.

[14] K. J. Keesman, *System identification: an introduction*. Springer Science & Business Media, 2011.

[15] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Online localization and tracking of multiple moving speakers in reverberant environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 88–103, 2019.

[16] Y. Ban, X. Alameda-Pineda, C. Evers, and R. Horaud, "Tracking multiple audio sources with the von mises distribution and variational em," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 798–802, 2019.

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[18] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[19] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[20] F. Zotter and M. Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature, 2019.