Viability of Eye-Tracking Glasses for Beamforming Hearing-Aids

Emily Steiner

Abstract—This study addresses the limitations of the current assistive hearing technology in isolation of specific sources within complex auditory environments, commonly known as the cocktail party problem. The proposed hardware setup incorporates a pair of glasses with a linear array of microphones along the bridge. Sound amplification can be spatially directed toward a steering angle of interest using established beamforming techniques, Delay and Sum (DS) and Minimum Variance Distortionless Response (MVDR). This paper focuses on comparing the efficacy of head movement versus eye-tracking as steering techniques to justify the associated technological complexities of the latter. A hardware and acoustic model was developed to simulate sound signal reception at each microphone. The evaluation scenarios were designed to mimic cocktail party settings, focusing on resolving sources at small angles and within the natural range gaze. Results indicate a significant Peak signal-to-noise ratio (PSNR) enhancement when fine-tuning with eye tracking and a notable qualitative improvement in output audio intelligibility. The resolution of individual signals is degraded with increasing scenario complexity. Future research should aim to improve the accuracy of the simulation by addressing assumptions made in this investigation and exploring metrics relevant to speech intelligibility. Moreover, prototyping and user studies are crucial steps to address the true practicality and intuitiveness of gaze-steering.

Index Terms—Auditory perception, Eye-tracking, Hearing Aids, Beamforming, Speech Intelligibility

1 Introduction

The cocktail party effect is a phenomenon in which humans can selectively focus on a single sound source despite being inundated with many complex background noises and voices. This ability, selective attention, results from the end-to-end human auditory system, including sensing and cognitive message selection abilities. Those who require assistive hearing technology find this task much more challenging. With hearing aids or other assistive auditory devices, key cues are lost [1].

With the recent advances and popularity in wearable technology, there is an opportunity to leverage current lightweight and low-profile designs to approach the cocktail effect problem. For this problem, we consider using regular eyeglasses, which can be fitted with a linear microphone array across the top edge. The phased microphone array will process auditory information and relay it to wearable headphones or hearing aids. Unlike the lack of spatial data in the case of a single microphone, processing algorithms can utilize small delays in received sound from the multichannel audio. This information depends on the angle of arrival and can be leveraged to assist with selective attention by algorithmically steering the sound amplification to an intended direction of audio focus.

This project explores the viability of eye-tracking as a steering technique for directing audio attention to a target source. Specifically, we explore the relative benefit of eye movement versus head movement as an attention selection mechanism to justify the added technological complexity of eye-tracking. There are two significant factors to consider. First, the effectiveness of the technology, both the hardware and algorithm, separation of one sound source from another. Without enough resolution to distinguish sources, eye-tracking, which represents a fine-tuning mechanism for head steering, would be irrelevant. Second, we discuss the steering method for ease of control from a user's perspective. Specific scenarios in which eye-steering would likely be more intuitive than head-steering are considered and sound output is compared quantitatively to address possible use cases.

2 Related Work

Beamforming is a technique in signal processing used to improve the signal-to-noise ratio of the transmission or reception of signals from a sensor array by suppressing incoming sounds from unintended directions [2] [3]. For this investigation, beamforming was selected, as it is a well-established technique that does not require training data. Alternatively, advances in related sound source separation areas, such as automatic speech recognition (ASR) [4], blind source separation (BSS) [5], and Direction of Arrival estimation [6], could be leveraged for the selective amplification portion of this application. More recent research has looked at learned approaches such as deep learning [4], which suggests replacing a beamforming algorithm with a learnt model to incorporate information from sound characteristics.

Current research also explores the effectiveness of various microphone array setups for the cocktail party problem. Nuance Hearing has developed a wearable microphone collar connected to earphones. The head pose was estimated and used as a steering mechanism. Multiple beamforming techniques were discussed and suggested [7]. Several recent studies have explored beamforming microphone arrays with gaze-directed steering. Jennings and Kidd et Al. propose a triple beamforming with eye-tracking for steering. This method uses forward, left-tilted, and right-tilted beams to produce individualized audio for the left and right ear [8], [9]. This research employs 16 microphones arranged

along the top of the patient's head. Culling et al. [10] utilized the triple beamforming approach but limited the microphone arrangement to the profile of a pair of glasses. This study compared gaze-steering beamforming against natural hearing in a user study to explore improving speech intelligibility. Anderson et al. also limit microphones to a pair of glasses [11]. Grimm et al. propose a probabilistic model of auditory attention that combines gaze direction and direction of arrival estimates for auditory sources in the environment [12].

3 METHOD

The primary focus of this paper is to assess the comparative efficacy of eye-tracking versus head movement as a steering technique. A situational simulation environment was developed to achieve this, and the output of each steering option was compared. The evaluation was conducted within hardware specifications, which limited microphones to the bridge of a pair of glasses to ensure findings reflected a possible practical and lightweight design. Figure 3.1.3 outlines the overview of the pipeline process.

3.0.1 Hardware and Acoustic Model

As previously mentioned, this problem is restricted to a linear microphone array intended to be fixed to the top ridge of a pair of glasses. To simplify the number of parameters considered, the overall length of the array is limited to 14cm to represent the maximum allowable size for standard glasses, and the array consists of 8 microphones.

To simulate the microphone response to the auditory scenario, the Anechoic Acoustic Model with several assumptions was used [3]. This model assumes the signal at each microphone is a delayed version of the source sound with some additive noise. For simulation, sound sources are simplified to planar waves and no attenuation due to propagation effects was modelled. The signal response to each source was assumed independent and subsequently summed to provide the overall microphone response. The additive white noise was scaled to ensure an SNR of 30dB in the microphone signal.

Implementation and simulation of the microphone array uses the array factor, the per-channel response of a phased array to a given stimulus. The array factor accounts for the additional propagation distance of the sound wave to each microphone by multiplying each signal frequency by a phase delay in the frequency domain. This factor is a Nxm where N is the number of elements in the array and m is the number of frequency elements in the audio signal's Discrete Fourier Transform (DFT). The array factor's i, j element for a given θ is defined in Equation 1. In the case of a linear array, the time delay at the ith microphone depends on d is the distance between array elements, θ is the angular direction of arrival and $c = 346m/s$, the speed of sound. Figure 1 displays this principle visually.

$$a(f_j, \theta)_{i,j} = e^{-2\pi j(i-1)dcos(\theta)f_j/c} \tag{1}$$

This factor is multiplied by the DFT of the input signals, and additive white noise is added. Overall, the hardware and acoustic model outputs the simulated signal of each microphone element in the array.

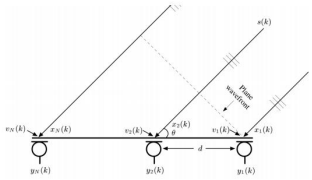

Fig. 1. Diagram of a Phased Array Response to a Plane wave [3]

3.1 Beamforming Methods

This paper utilizes beamforming for the source separation algorithm. With a specific target steering angle, beamforming algorithms suppress other incoming sounds from unintended directions. In the frequency domain, this is accomplished by combining the measured signals with a complex weighting [2].

3.1.1 Delay and Sum

The most straightforward beamforming algorithm is Delay and Sum (DS). This uses the same principle of phase shifting, or delaying, audio channels based on the additional distance a sound wave must travel from one array element to the next. Equation 2 shows the weights for the DS case are equal to the array factor at the frequency and steering angle of interest.

$$H(f) = a(f, \theta) \tag{2}$$

3.1.2 Minimum Variance Distortionless Response

Minimum Variance Distortionless Response (MVDR) is an adaptive beamforming technique incorporating a steering direction and the spatial covariance matrix of received samples to minimize any noise from outside the target direction [13]. The weights alter both the phase and scaling of individual microphone elements. Equation 3 and 4 show the weight calculation where $r(f)$ is the incoming signal in the frequency domain.

$$H(f) = \frac{R^{-1}(f)a(f, \theta)}{a^H(f, \theta)R^{-1}(f)a(f, \theta)} \tag{3}$$

$$R(f) = r(f)r^H(f) \tag{4}$$

MVDR's algorithm maintains the signal's power in the steering angle direction while minimizing the total power. The result is the cancellation or suppression of noise sources through the placement of zeros in noise source directions.

3.1.3 Broadband Beamforming

In broadband beamforming, the principles of narrowband beamforming discussed above are applied independently to frequency bins. For this implementation, 12 frequency bands were used. The center of the bands was guided by the frequencies most important to speech intelligibility [14], as well as looking at the spectrum of input signals. Each frequency is masked, and in the case of MVDR, the spatial covariance matrix is calculated based on the masked signal. Following the calculation and application of the weights, the independent frequency bins are summed and converted back to the domain. The result is a single audio signal. The Figure 3.1.3 diagram includes a visualization of broadband beamforming [13].

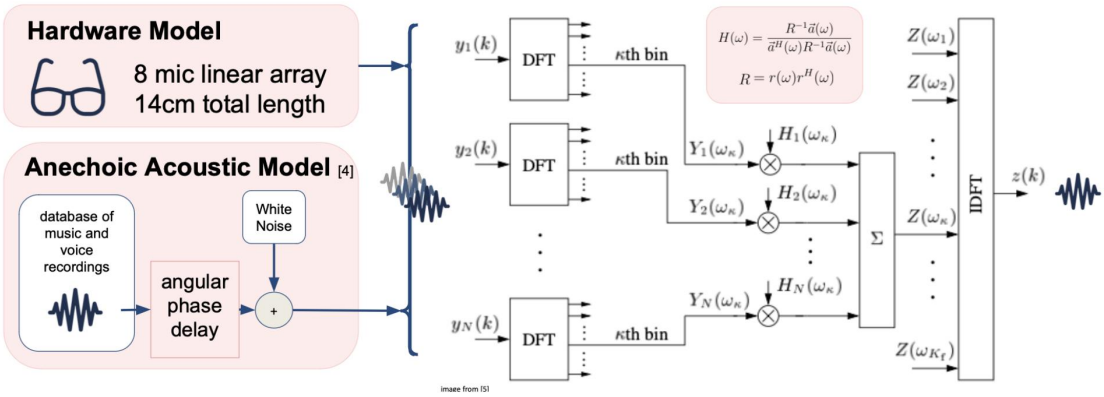

Fig. 2. **Pipeline:** The pipeline overview for simulation of audio environments and the processing of the beamforming algorithm. For all input audio signals, the hardware and acoustic model simulate the signal received at each microphone. For the same simulation, the broadband beamforming algorithm [13] evaluates the resulting audio of any number of desired steering vectors.

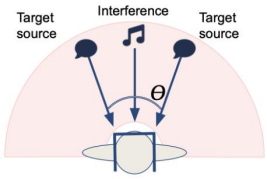

Fig. 3. **Scenario 1:** considers a conversation with two human speakers or 'target sources' arranged equidistant on either side of the 'listener.' The angular distance between these far-field sources is defined as θ.

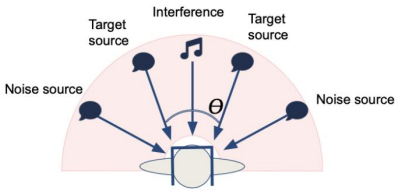

Fig. 4. **Scenario 2:** adds two speakers coined 'noise sources' on either side of the existing setup in Scenario 1. Each voice source is arranged θ degrees apart from the next voice source.

3.2 Acoustic Scenarios

The following simulation scenarios propose scenarios where eye steering is likely more intuitive than head or body steering. Eye-steering can be considered the fine-tuning mechanism of the body and head alignment. This would benefit environments where attention is shifted dynamically between sources within the comfortable angular gaze range. In particular, a fine-tuning mechanism would be most helpful in small-angle resolution scenarios. Two 'cocktail party' setups are considered; for each, we assume the gaze is aligned with the intended direction of auditory focus and can be aligned directly with the source direction. The listener's head will always face forward towards a music interference directly in front, but the source of interest will vary among speakers. Figure 3 and Figure 4 show a visual representation of the two scenarios considered, defined as Scenario 1 and 2, respectively.

The speech and music audio used will be sourced from the experimental setup of a University of Bern dataset intended for processing for the cocktail party problem [15]. Speech audio was selected by researchers at Bern from the LibriTTS corpus [16] with the criteria of having a signal-to-noise ratio (SNR) of at least 20 dB. Music was sourced from the Musan "popular" corpus [17] and sliced to avoid fade-in and fade-out. This project uses a microphone setup that is different from the Bern dataset. The audio used from the above datasets was not altered in power or SNR.

The angular distance between voices is varied between 30 and near zero degrees to investigate the ability to resolve different voices at small angles. The experiment was repeated 30 times at each angular distance to limit the influence of the select combination of audio sources. The combination of audio inputs is randomized for each iteration, using 12 voice recordings and four music recordings.

4 Experimental Results & Discussion

Simulation parameters considered include the number of sources, the angular distance between voice sources, source types (audio variation), the direction of steering, and the algorithm used. The metric used for evaluation is the peak signal-to-noise ratio (PSNR) of target audio with respect to the original audio of interest.

4.1 Scenario 1

The result of Scenario 1 was quantified using the PSNR metric. The steering direction is varied to explore the effect of fine-tuning the steering. The plot in Figure 5 and 6 represents when the voice sources are space 20 ° apart using DS and MVDR, respectively. The intersections with dotted vertical line represents the PNSR when the steering angle is aligned perpendicular to the microphone array or the head steering case.

For DS, the steering angle has little effect on the PSNR of the output signal. DS results are not reported due to poor behaviour for the remainder of the discussion. However, for

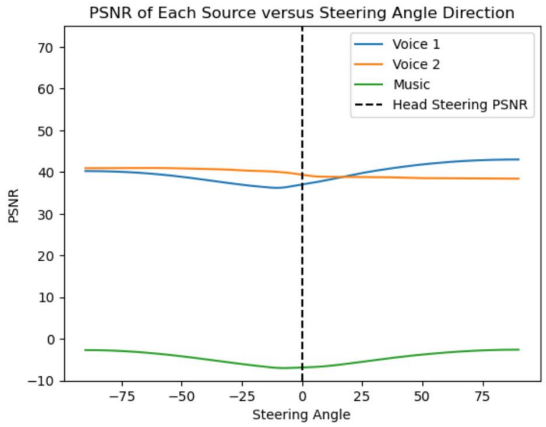

Fig. 5. Example Experiment: Scenario 1, DS algorithm, Voice sources spaced 20° apart

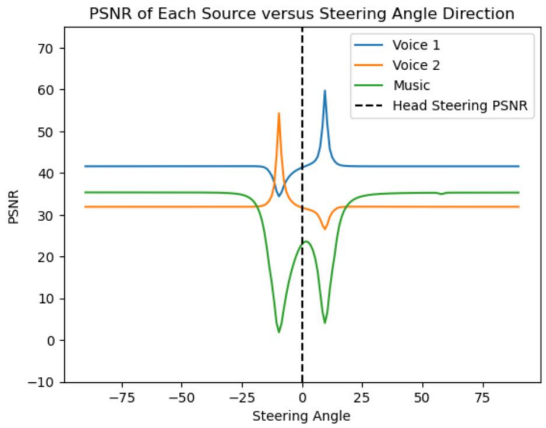

Fig. 6. Example Experiment: Scenario 1, MVDR algorithm, Voice sources spaced 20° apart

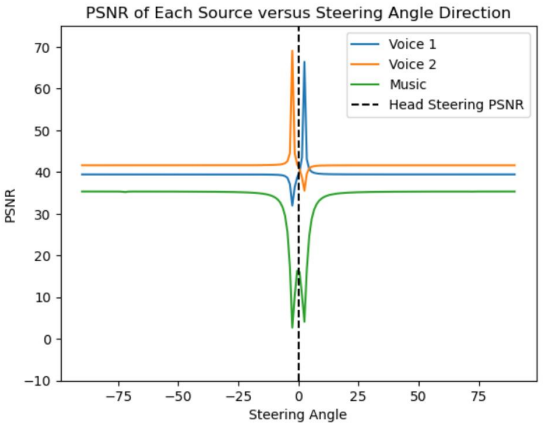

Fig. 7. Example Experiment: Scenario 1, MVDR algorithm, Voice sources spaced 5° apart

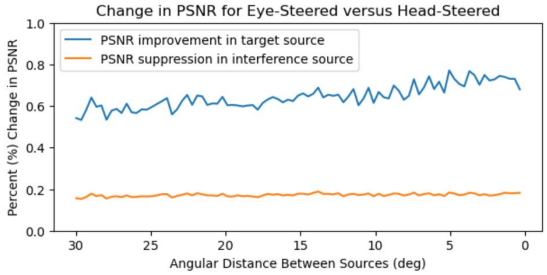

Fig. 8. **Scenario 1:** Percent improvement in PSNR of target sources and percent suppression in PSNR of interference sources varying the angular distance between sources.

MVDR, there is a significant improvement in the PSNR of a target voice source when the steering angle is aligned with the true angle of arrival. Additionally, there's a suppression in the interference sources at that steering angle.

As the angle between voice sources decreases, the magnitude of the PSNR peak slightly increases; Figure 7 shows the results for voices spaced 5° apart. Here, the fine-tuning provides only a 2.5° shift in steering from the head steering case; despite this, there's a significant peak in PSNR of the target sources.

The qualitative results of this particular example scenario can be listened to through demo audio submitted alongside this paper. An audio file named "scenario1_mvdr.wav" contains output audio from the simulated scenario corresponding to Figure 7. The audio is split into three different steering angle options in the same scenario. The first roughly 3.5 seconds of audio represent the head steering case, where the MVDR algorithm is focused directly in front of the linear array. Next, the audio scenario is repeated, and the MVDR algorithm is instead eye-steered directly on one of the target voices and then the other. In the first clip, you can hear a mix of audio and a little background music; in the focused, fine-tuned clips, you hear precise amplification of a particular voice.

To explore the limit of the MVDR algorithm's ability to resolve small angular distances between sources, the experiment was repeated for source spacing ranging between [30, 0) ° . Output results varied significantly with the randomized selection of input audio. The percent improvement in PSNR at the true signal's incoming angle and the 0° case was calculated to compare these randomized experiments properly. The same premise was applied to observe the percent suppression for interference sources. The experiment was repeated 30 times at each angle, and the results were averaged. The results of this evaluation for scenario one are plotted in Figure 8.

Overall, the percent improvement and PSNR remained approximately constant for Scenario 1 even as the angular distance approached zero degrees. Despite the experiment averaging, the variation between data points was high; thus, precise trends are not discernible.

4.2 Scenario 2

For the second scenario, example experiments are plotted for voice sources spaced 20° and 5° apart in Figures 9 and 10, respectively. For the angular spacing of 20°, there are more modest improvements in the PSNR than Scenario

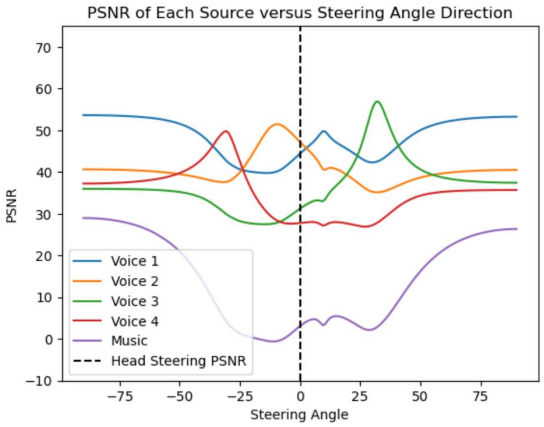

Fig. 9. Example Experiment: Scenario 2, MVDR algorithm, Voice sources spaced 20° apart

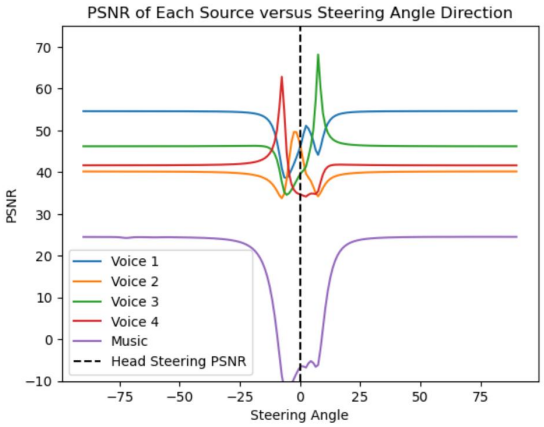

Fig. 10. Example Experiment: Scenario 2, MVDR algorithm, Voice sources spaced 5° apart

1. The PSNR peaks are much less steep when comparing eye-tracking and head-steering; this indicates limitations in angular resolution. From Figure 10, the 5° spacing experiment shows better isolation when focused on the outer sources, coined noise sources, versus the innermost, or target, sources. This is expected as there is a limit on how narrow the MVDR main beam can focus, given the microphone spacing, which is limited by the hardware requirements. This limit is tested when the interfering audio sources are spaced equally on either side of the target source, as in Scenario 2.

The audio output corresponding to Scenario 2 with 5° spacing is included as a demo in an audio file named "scenario2_mvdr.wav". The audio is split into five different steering angle options for the same scenario, as in the other included demo, where the first roughly 3.5 seconds of audio represent the head steering case. Next, the audio scenario is repeated, and the MVDR algorithm is instead eye-steered directly on one of the inner target voices and then the other. Finally, the audio scenario is repeated for each outer voice. In the first clip, again, the layer mix of

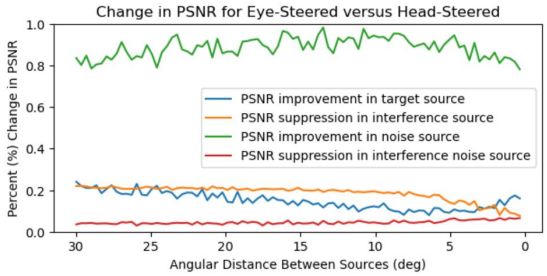

Fig. 11. **Scenario 2:** Percent improvement in PSNR of target sources and percent suppression in PSNR of interference sources varying the angular distance between sources.

voices and a little background music are apparent. Interference voices can be heard in the focused, fine-tuned clips of the inner target audio. In this randomized example, one input voice clip includes a forceful use of the word 'not,' which made isolation and intelligibility of different target sources more challenging. This section of the audio clip was likely more disruptive because of the subject's clear and forceful pronunciation of a word that often has significant emotional weight. PSNR is limited in its ability to quantify more complex intelligibility effects. The outer targets can be mostly isolated from other interference.

As in Scenario 1, the experiments were repeated for voice source spacing ranging between [30, 0) ° . The results are plotted in Figure 11. Since the target and noise voices (inner and outer sources) vary in PSNR behaviour, they are plotted individually. A more significant PSNR improvement in eye-tracking to steer to the outer voices matches the observation in the individual example experiments. Regardless of source angular distance, the inner target source improvement is much smaller than In Scenario 1, from over 80% PSNR improvement to roughly 20%. There is a trend of a decrease in PSNR improvement until about 5° between the sources and a slight uptick, but all improvements remain below 30%. However, similar to Scenario 1, there are significant variations between data points due to the dependence on the audio clips used as input.

5 Conclusion, Limitations, & Future Work

Overall, the number of sources and the complexity of the auditory environment had a more significant effect than the angular distance between sources on the maximum PSNR of a target. Despite variation in the amount of improvement, fine-tuning the steering angle always improved the PSNR of the target source. For this study, the ability to steer off-axis is assumed to be contingent on eye-tracking. Thus, the increase in PSNR demonstrates the promising benefit of eye-steering versus head-steering.

The PSNR metric is limited in its ability to characterize the complexity of how similar two audio signals sound. Although the quality of the isolation of the output audio generally matched what was expected given the PSNR, there were some discrepancies. In the future, using an out-of-the-box speech recognition program on the output and calculating the word error rate may provide a better metric for speech intelligibility.

It is also important to note that these findings are contingent upon the assumptions made in the simulation model and likely represent an idealized ability to isolate and resolve target audio sources. Room reverberations, echoing, non-ideal far-field sources, and sound attenuation are examples of additional factors which could be considered to improve the accuracy of future simulations. Possible acoustic models to consider are the Multiple-Source Reverberant Model [3] and the Image Source Model [18]. Additionally, the number of audio clips used for testing should increase to improve the variation in evaluation. Alternatively, the investigation could consider the creation of a prototype to record audio in natural acoustic environments. The latency of the algorithm would be of concern for a real-time implementation for a prototype; time-windowing and algorithmic complexity may become limiting factors in the system's performance.

An additional benefit of a prototype is the ability to evaluate the most prominent assumption in this study: eye-tracking represents the ground truth of the user's intended audio focus. Inevitably, there are cases where neither gaze nor head tracking aligns with the direction of audio attention—for example, looking at a presentation but listening to a speaker. Audio information is processed through a complicated cognitive process, and as a result, there is no one body language signal that maps directly one-to-one to auditory focus.

First, an investigation with a prototype and a user study should understand the frequency at which eye-tracking and head-steering deviate from one another and the intended auditory focus in complex acoustic environments. Eye-tracking will provide little additional fine-tuning benefit if these two steering techniques are well aligned regardless of accuracy with ground truth. Next, the study should investigate if eye-tracking is a steering technique that is easy, intuitive and convenient. If this is the case, regardless of a direct mapping to intended auditory attention, eye-tracking could still be a useful hands-free selection tool between multiple audio streams.

If neither eye-tracking nor head-steering provides sufficient alignment with focus, another avenue of investigation could be implementing a probabilistic attention model. The use of head direction, gaze, sound characteristics, and outward-facing cameras could estimate the intended attention direction without the assumption any one method is ground truth. Future investigations should include learning-based source separation approaches rather than beamforming, such as a neural network. A learnt model's goal would be to jointly separate multiple sources and estimate the direction of arrival from multichannel audio data. With an estimation of attention, the source separation model could be jointly optimized.

This investigation shows promise in developing a fine-tuning mechanism for spatially directed audio augmentation. However, the author believes it is unlikely that eye-tracking alone will be able to sufficiently provide this fine-tuning mechanism. If its limitations are properly addressed and considered in future investigations, eye-tracking may still provide important information about auditory focus.

Acknowledgments

The author would like to thank Professor Gordon Wetzstein, Qingqing Zhao, and Axel Levy for a great course!

References

[1] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O society*, vol. 12, no. 7, pp. 35–50, 1992.

[2] *Synthesis of Linear Arrays and Apertures*. John Wiley Sons, Ltd, 2002, ch. 3, pp. 90–230. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471221104.ch3

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[4] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[5] L. Drude and R. Haeb-Umbach, "Integration of neural networks and probabilistic spatial models for acoustic blind source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 815–826, 2019.

[6] K. Ahuja, A. Kong, M. Goel, and C. Harrison, "Direction-of-voice (dov) estimation for intuitive speech interaction with smart devices ecosystems," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 1121–1131.

[7] Y. H. et al., "Directional hearing aid," U.S. Patent 0 567 888B2, Feb. 18, 2020. [Online]. Available: https://patents.google.com/patent/US10567888B2/en?oq=10567888

[8] T. R. Jennings and G. Kidd, "A visually guided beamformer to aid listening in complex acoustic environments," in *Proceedings of Meetings on Acoustics*, vol. 33, no. 1. AIP Publishing, 2018.

[9] G. Kidd, T. R. Jennings, and A. J. Byrne, "Enhancing the perceptual segregation and localization of sound sources with a triple beamformer," *The Journal of the Acoustical Society of America*, vol. 148, no. 6, pp. 3598–3611, 2020.

[10] J. F. Culling, E. F. D'Olne, B. D. Davies, N. Powell, and P. A. Naylor, "Practical utility of a head-mounted gaze-directed beamforming system," *The Journal of the Acoustical Society of America*, vol. 154, no. 6, pp. 3760–3768, 2023.

[11] M. H. Anderson, B. W. Yazel, M. P. Stickle, N.-G. S. Gutierrez, M. Slaney, S. S. Joshi, L. M. Miller *et al.*, "Towards mobile gaze-directed beamforming: A novel neuro-technology for hearing loss," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 5806–5809.

[12] G. Grimm, H. Kayser, M. Hendrikse, and V. Hohmann, "A gaze-based attention model for spatially-aware hearing aids," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[13] H. V. Trees, *Optimum Array Processing*. John Wiley Sons, Ltd, 2002. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471221104.ch7

[14] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *The journal of the Acoustical society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[15] T. Fischer, M. Caversaccio, and W. Wimmer, "Multichannel acoustic source and image dataset for the cocktail party effect in hearing aid and implant users," *Scientific data*, vol. 7, no. 1, p. 440, 2020.

[16] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[17] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[18] J. E. Summers, "Auralization: Fundamentals of acoustics, modelling, simulation, algorithms, and acoustic virtual reality," 2008.