EE367 Project Proposal

Emily Steiner

1 Project Description

The cocktail party effect is a well-discussed phenomenon in which humans can selectively focus on a single sound source despite being inundated with a number of complex background noises and voices. This ability, selective attention, is the result of the end-to-end human auditory system: sensing and cognitive message selection. 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 the use of 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 information in the case of a single microphone, processing algorithms can utilize small delays in received sound from the multichannel audio to estimate the Direction-Of-Arrival (DOA). The DOA information can be used to assist with selective attention by algorithmically steering the sound amplification to an intended direction of audio focus.

This project aims to explore the potential 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. There are two major factors to consider. First, the effectiveness of the technology, both the hardware and algorithm, to be able to separate one sound source from another. Without enough resolution to distinguish sources, the steering method used would be irrelevant. Second, we consider the focus selection method, for ease of control, from a user's perspective. In this case, specific scenarios in which eye-steering would likely be more intuitive versus head steering are considered and sound output is quantitatively compared.

2 Related Work

For the source separation algorithm, the investigation will primarily consider beamforming. Beamforming is well-established in signal processing as a method to improve the signal-to-noise ratio of the transmission or reception of signals from a sensor array [2]. With a specific target steering angle, beamforming algorithms suppress other incoming sounds from unintended directions. The simplest algorithm is Delay and Sum (DS). This uses the principle of delaying (phase shifting) audio channels with respect to the additional distance a sound wave must travel from one array element to the next. Figure 1 displays this principle visually.

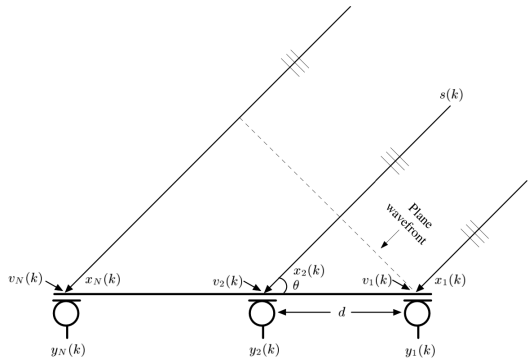

Fig. 1. Delay and Sum Beamforming [3]

Minimum Variance Distortionless Response (MVDR) is an adaptive beamforming technique that incorporates a steering direction and the spatial covariance matrix of received samples to minimize any noise not coming from the target direction [4]. Both the phase and scaling of individual sensors are modified for MVDR through array weights.

Addressing only the first sub-problem, sound source separation, there are several related research venues including automatic speech recognition (ASR) [5], blind source separation (BSS) [6], and DOA estimation [7]. More recent research has looked at learned approaches such as deep learning [5].

Nuance Hearing approached the combined problem (spatially directed assistive hearing) using a wearable microphone collar. The collar would sit on the user's chest around their neck and be connected to earphones. The head pose was estimated and used as a steering angle. Multiple beamforming techniques were suggested and MVDR was specifically derived [8].

3 Proposed Method

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. The overall length of the array is limited to 14cm to represent the maximum allowable size of the array for standard glasses. The number of microphones will not be limited for this investigation. Implementation and simulation of the microphone array uses the array factor, the per-channel response of a phased array to a given stimulus.

If time permits, noise should be injected into the simulation of the sensors.

First, the limitations of the hardware setup will be explored by comparing the maximum phase change that would occur when varying the DOA. This will evaluate the DOA resolution which can be inherently expected from the constraint in total array length.

For the following scenario evaluation, we assume the gaze is aligned with the intended direction of auditory focus. The following planned simulations attempt to propose scenarios in which eye steering may be considered more intuitive than head or body steering.

Scenarios:

- **Group discussion:** N simultaneous speakers arranged as a circle where the microphone array is worn by one speaker. The listener's head will always face forward but the speaker of interest will vary among speakers.
- **Distracted viewer:** A neighbour speaker will be offset from the center while a loud sound source (eg. voice presentation, musical performance etc.) will be placed directly in front of the listener.
- **Scanning eavesdropper:** N sources (voice/music) arranged randomly where the sound source of interest will vary.

For simulation, sound sources are simplified to planar waves (far field) and their direction of arrival and relative power (loudness) will be varied. The speech and music audio used will be sourced from the experimental setup of a University of Bern dataset intended for processing for cocktail party [9]. Speech audio was selected by researchers at Bern from the LibriTTS corpus [10] with the criteria of having a signal-to-noise ratio (SNR) of at least 20 dB. Music was sourced from the Musan "popular" corpus [11] and sliced to avoid fade-in and fade-out. This project uses a different microphone setup than the mentioned Bern dataset.

Simulation parameters to consider include the number of sensors, source locations, source types (audio variation), source loudness, the direction of steering, and the algorithm used. Metrics for evaluation include the SNR of target audio and the half power bandwidth at key frequencies for speech intelligibility [12].

Given the short timeline, this project's scope will only include the evaluation of beam-forming algorithms. The final paper will discuss the limitations of the beamforming algorithm approach. Future investigation should include learning-based source separation approaches such as a neural network. A learnt model's goals would be to jointly separate multiple sources and estimate the DOAs from multi-channel audio data.

In the future, the ease of this steering method should be evaluated and compared to head-only steering with user studies. However, there are inevitably cases where the gaze (and head steering) does not align with the direction of audio attention. For example, looking at a presentation but listening to a speaker. Another avenue of investigation could include the use of head direction, gaze, and DOA to estimate the intended attention direction without the assumption any one method is ground truth.

4 Milestones & Timeline

Week 7: Using the existing simulation framework, systematically vary system parameters for the first proposed simulation scenario (group discussion).

Week 8: Extend evaluation to the second and third scenarios. Prepare an extended set of scenarios (with both forward and offset steering directions) with sufficient randomization for possible extension to a learning-based approach.

Week 9: Paper writing and Poster session. Possible Implementation of a Neural Network for sound source separation and comparison to performance to beamforming techniques using the specified scenarios.

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] *Adaptive Beamformers*. John Wiley Sons, Ltd, 2002, ch. 7, pp. 710–916. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471221104.ch7

[5] 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.

[6] 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.

[7] 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.

[8] N. Hearing, "Nuance hearing 2020 patent."

[9] 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.

[10] 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.

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

[12] 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.