Towards mobile gaze-directed beamforming: a novel neuro-technology for hearing loss*

Markham H. Anderson**, Britt W. Yazel**, Matthew P. F. Stickle, Fernando D. Espinosa Iñiguez, Nathaniel-Georg S. Gutierrez, Malcolm Slaney, *Fellow, IEEE*, Sanjay S. Joshi, *Senior Member, IEEE*, and Lee M. Miller

Abstract— **Contemporary hearing aids are markedly limited in their most important role: improving speech perception in dynamic "cocktail party" environments with multiple, competing talkers. Here we describe an open-source, mobile assistive hearing platform entitled "Cochlearity" which uses eye gaze to guide an acoustic beamformer, so a listener will hear best wherever they look. Cochlearity runs on Android and its eight-channel microphone array can be worn comfortably on the head, e.g. mounted on eyeglasses. In this preliminary report, we examine the efficacy of both a static (delay-and-sum) and an adaptive (MVDR) beamformer in the task of separating an "attended" voice from an "unattended" voice in a two-talker scenario. We show that the different beamformers have the potential to complement each other to improve target speech SNR (signal to noise ratio), across the range of speech power, with tolerably low latency.**

I. Introduction

Everyday auditory environments are cluttered, noisy, and distracting. This presents a complex perceptual and computational challenge known as the "cocktail party" problem: how to extract relevant acoustic information while filtering out the background noise. Individuals with healthy hearing tend to perform well in typical multi-talker environments, as their brains are adept at discriminating sound source locations and identities [1]. However, while hearing aids can significantly boost the detection and comprehension of sounds for those with hearing loss, particularly in quiet backgrounds, and can even improve "downstream" effects on auditory cognitive function [2], they do not adequately address the issue of understanding speech in noise.

Modern digital hearing aids often use multiple features to improve perception in loud, crowded environments, such as on-board directional microphone systems and adaptive speech enhancement or noise reduction algorithms. But even with these sophisticated features, aids cannot effectively "listen" to what the user wants; they often fail in real situations, amplifying noise as much as the desired information. This shortcoming leads to listening confusion, poor real-world speech comprehension, and low rates of use for assistive devices [3].

We sought to address this issue by creating a system that can be automatically guided by a user's intentions — in this case their eye gaze direction — and thereby serve as the basis for an intelligent hearing aid [4].

Our approach builds on the seminal work of Kidd et al. (2013), Hart (2009), and Marzetta (2010) [5, 6, 7]. Similar to Kidd et al., we use gaze-directed beamforming, a method of highly-directional sound amplification and attenuation, to isolate sounds within a "beam" of auditory space. And like Kidd et al., the direction of the beam will be steered through real-time gaze tracking, as an analog to listening intention. The primary differing factors between Kidd's system and our "Cochlearity" platform are three-fold. First, Cochlearity is implemented on widely available mobile device hardware using Android as opposed to workstation-class desktop hardware. Second, this first version of Cochlearity will be entirely open-source and available under a standard, permissive license to encourage broader adoption and further improvements. And finally, we make use of both passive and adaptive beamforming algorithms, as opposed to just passive. In this report, we evaluate whether combining passive and active beamforming algorithms in parallel might improve performance substantially with little computational cost.

II. Design

Unlike a more traditional PC-based implementation of acoustic beamforming, we power Cochlearity (the software application) with a Nexus 9 tablet running Android (6.x), although this technology can feasibly work on any capable device running Android. Our implementation uses an array of eight microphones; however, with Android OS lacking support for input of more than two audio channels, it was necessary to develop our own software and hardware solution for 8-channel I/O. We used a Tascam US-16x08 audio recorder, which serves as our multi-channel Analog to Digital Converter (ADC), and for gaze input we use a Tobii Rex eye tracker and associated Tobii Gaze Android SDK and driver, both connected to the tablet via a powered USB hub and OTG (on-the-go) cable (Fig. 1).

III. Beamforming

The premise of acoustic beamforming is to combine signals from an array of multiple, precisely spaced microphones to emphasize sound energy emanating from a single region of space radially oriented about the array, all while suppressing sounds from any other regions. A passive

*Funded through the Google Faculty Research Award (Lee Miller) and ARCS Foundation: Northern California Chapter (Britt Yazel).

*The project described was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number UL1 TR001860. The Content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

**Authors B. Yazel and M. Anderson are co-first authors

M. Slaney is Research Scientist in the Machine Hearing Group at Google Research

S. Joshi is a Professor in the Department of Mechanical and Aerospace Engineering at U.C. Davis (e-mail: maejoshi@ucdavis.edu)

L. Miller is an Associate Professor in the Center for Mind and Brain and Department of Neurobiology, Physiology, & Behavior at U.C. Davis (email: leemiller@ucdavis.edu)

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

beamformer uses only the array geometry and speed of sound to combine the signals mathematically; as a result, it will tend to be simple with low latency. An adaptive beamformer additionally uses statistical learning about noise sources in the environment, which can improve performance but may be practically limited in real-time applications by the additional computational cost. In both cases, the beamformer outputs a single, spatially-sensitive audio signal that contains proportionally more information from one region of space than from any others [8]. Cochlearity currently implements two distinct beamforming algorithms, 'delay-and-sum' and 'Minimum Variance Distortionless Response' (MVDR).

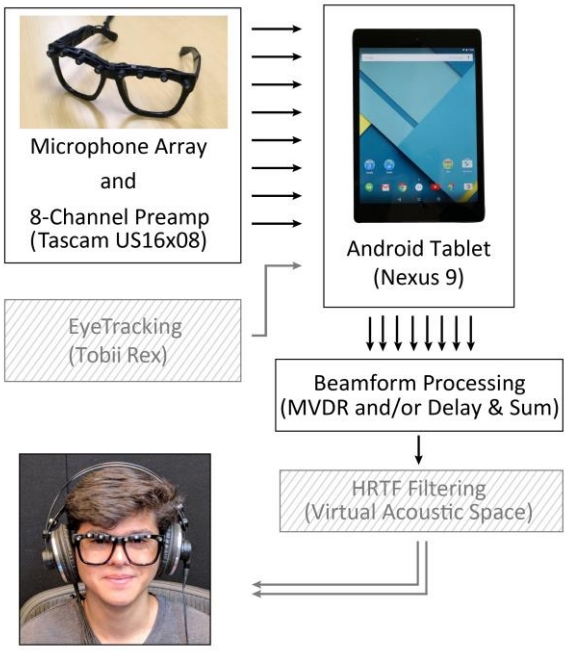

Figure 1. "Cochlearity": a mobile, gaze-directed beamforming platform for assistive listening. Elements shown in hatching were not used for the current experiments.

Cochlearity first buffers the 8-channel USB audio inputs into 1024 sample (21 millisecond) frames, which are then passed on to the filtering or beamforming operations. Shorter frames would have enabled lower latency but would diminish granularity for the discrete Fourier transform used by our adaptive beamformer and would eventually have presented an I/O bottleneck. Thus, the samples being processed are always necessarily (at least) 21 milliseconds in the past, though with processing and I/O operations this latency is considerably longer, detailed later.

A. *Delay-and-sum Beamforming (D&S)*

Delay-and-sum beamforming is a passive algorithm that leverages the propagation time of sound, which manifests as signal delays from one microphone to the next. This delay varies with the angle of the target audio relative to the microphone array. By offsetting the signal in each channel by the delay for a given "steering" angle and then summing the resulting signals across channels, it delivers a signal that contains a constructively reinforced component coming from the desired angle, with all other angles destructively attenuated [9].

B. *Minimum Variance Distortionless Response (MVDR) Beamforming*

The MVDR beamformer is an adaptive algorithm, as opposed to the delay-and-sum beamformer. In addition to compensating for the time delays due to steering angle, it uses an adaptive filter to null interference from other angles [10]. Time and frequency are divided into bins of fixed size, and for each time-frequency bin, an NxN noise correlation matrix is computed (N being the number of microphones). From the noise correlation matrices, a linear transformation is computed to minimize the noise (anything other than the desired signal) on current and future inputs, with the constraint being that the signal originating on the target angle be preserved [11].

To reduce computational expense, our implementation of the MVDR beamformer used only the two end microphones, bringing the noise correlation matrices down to a 2x2 dimension instead of 8x8. The delay-and-sum beamformer, however, used all eight microphones to improve the resolution of its constructive/destructive interference operation, with scarcely higher cost than a 2-microphone delay-and-sum beamformer. When used in isolation, each beamforming algorithm is given the full audio bandwidth as input.

IV. Methods

All testing was conducted in a sound treated room with dimensions of 3.5 x 2.5 meters. Eight Audio-Technica AT8537 phantom powered microphones with an 80Hz high-pass pre-amp filter were mounted linearly upon a set of eyeglasses with 1.86cm spacing and a total length of 13cm. The glasses were set upon on an anatomically accurate dummy head positioned facing forward (defined as 0º) on a table, with two Tannoy Precision 6 speakers positioned 140cm away at -50º and +50º pointing directly at the array. Speaker outputs were balanced using a digital sound level meter to within 1 dB SPL using Gaussian white noise.

During each performance test, the beamformer steering direction was manually set by the researcher. Two audio tracks were played simultaneously, each through one of the speakers at a comfortably loud listening level. The speaker positioned at -50º played "20,000 Leagues Under the Sea" while the speaker at +50º played "Journey to the Center of the Earth", both by author Jules Verne. Both stories were read by the same male reader at a constant pacing and were equalized for power, but the voice at -50º was pitch-shifted up by 7%, and the voice at +50º was pitch-shifted down by 7% using Adobe Audition.

For all recordings, beamformer output was captured directly from the tablet through the headphone jack, which was then passed into a Sound Devices X-3 headphone amplifier, amplifying the audio before it was sent to the USB audio capture card, an Edirol (Roland) UA-25, and then onto the PC using Audacity.

The recordings were made in sets of two for each beamforming paradigm: beam steered to -50° azimuth (left), and +50° azimuth (right). Lastly, two reference recordings were made with all beamformer processing turned off, and the speech was played separately out of the -50° speaker or the +50°. These references were necessary as a point of comparison for the aforementioned recordings, as they

captured the same signal filtering imposed by the room, speaker placement, and microphones, as well as the generic I/O overhead in Android. Thus, any differences between them and the beamformed recordings should be due entirely to the processing imposed by Cochlearity's beamforming.

To compare our two different beamformers, we use spectral coherence as a measure of the relatedness between our beamformed audio recordings and the reference recordings. Each recording was compared against the left or right talker reference for spectral coherence. Specifically, in each beamforming paradigm:

"Attended Voice" is the congruent coherence between the left speaker reference recording and the beamformed recording when steered to the left, and coherence for the right speaker reference recording with the beamformed recording when steered to the right, averaged together.

"Unattended Voice" is the incongruent coherence between the left speaker reference recording and the right-steered beamformed recording, and coherence for the right speaker reference recording to the left-steered, beamformed recording, averaged together.

Likewise, whereas coherence shows beamformer performance as a function of frequency, overall performance can be summarized as SNR (dB) between attended and unattended voices. We calculated SNR as $10*\log_{10}$ of the average ratio in coherence between the two conditions, weighted by the speech power across frequencies.

A. Coherence Difference Index (CDI) and Latency

To quantify the effectiveness of a beamformer, we computed a "Coherence Difference Index (CDI)" for each paradigm (Table 1). We calculated this by taking the difference between the "attended" and "unattended" coherence for a given beamformer, and then performed a weighted average between 0 and 5000Hz (which captures the majority of speech power), weighted by the spectral density estimate of both voices combined (using Matlab's *pwelch* function). We then multiplied this number by a factor of 100. This provided a rough global approximation of how well each paradigm emphasized the voice from the steering direction *and* suppressed the interfering voice.

A "CDI Efficiency" was determined as the CDI per millisecond of latency (CDI ÷ Latency) (Table 1).

Latency was computed using a series of clicks played through the speakers and recorded both in a reference microphone (not-connected to Cochlearity) and through Cochlearity for each beamforming paradigm (Table 1).

B. Spatial Analysis

To characterize the spatial effectiveness of our two beamformers, we performed an analysis in which we placed a speaker at 0 ° and kept the beamformer steered to this angle. The story played by this speaker, "20,000 Leagues Under the Sea", is referred to as the attended voice. Next, we moved a second speaker playing a masking voice, "Journey to the Center of the Earth", in 10° increments from -50° to +50°, recording 1 minute of speech at each location. We then assessed the performance of the beamformer at each masking angle relative to the 0° fixation using the CDI as our metric, illustrating the effectiveness of the beamformer in extracting the attended voice from the masking voice.

Lastly, while real-time gaze tracking and virtual 3-d audio rendering using head-related transfer functions or HRTFs are integral and fully realized parts of Cochlearity, the data reported in this study are only meant to characterize the efficacy of Cochlearity's beamforming implementation, and as such there is no gaze tracking component to the tests (Fig 1). A future study will explore the effects of how Cochlearity performs with human subjects.

V. Results

The delay-and-sum beamformer did not perform well at frequencies <300Hz, with little difference between the attended voice and the unattended voice coherence. However, at frequencies above 300Hz, and most notably >1000Hz the beamformer was able to effectively separate the attended from the unattended voices. The low-frequency performance reflects, in part, the relatively small array size and confirms the literature that delay-and-sum beamforming works best at moderate to higher frequencies [5] (Fig. 2).

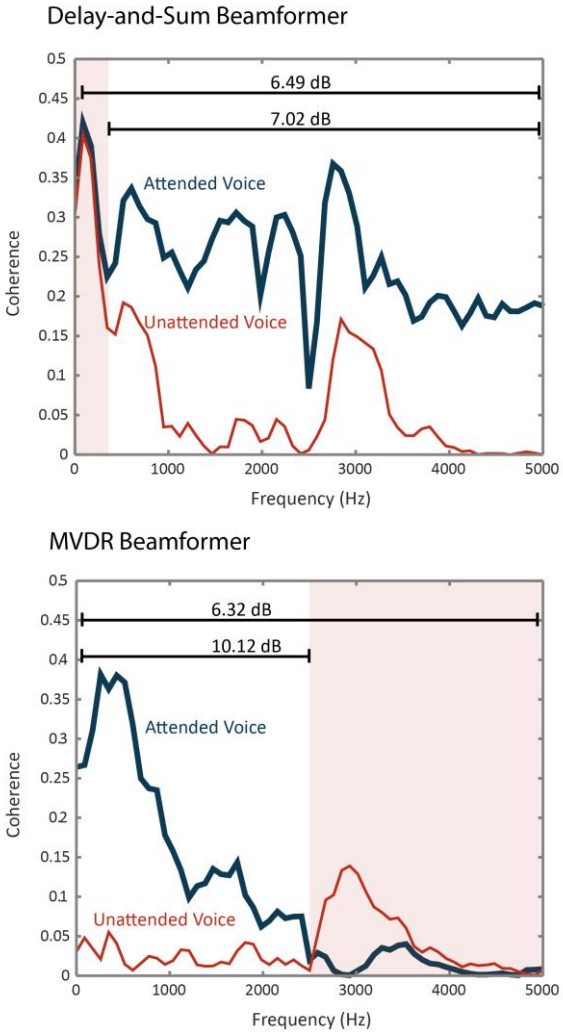

Figure 2. Delay-and-Sum and MVDR beamformer performance (red shaded frequencies indicate poor performance). Decibel (dB) labels indicate average SNR between attended and unattended voices.

The MVDR beamformer performed well, most notably at frequencies <2500Hz. Conversely, considering the performance of delay-and-sum, the MVDR beamformer

performed the worst at middle to higher frequencies, leading to little or no improvement in signal coherence between the attended and unattended voices (Fig. 2).

Thus, the two beamformers complement one another in performing across the crucial frequency range where speech has high power.

Spatially, both beamformers performed well, showing a clear trend in increasing CDI values the farther away the masking voice was from 0°. This is expected given that maximal spatial overlap between attended and unattended voices occurs at 0°. With regard to the magnitude of the CDI, at 0° both beamformers were equivalent, each with a CDI of ~0, but within +/-10-20° the MVDR beamformer outpaced the delay-and-sum, ending at -50° and +50° with more than double the CDI of the delay-and-sum (Fig. 3). This indicates that the MVDR has better performance at much smaller masking angles than that of the delay-and-sum. Nevertheless, Table 1 shows that at 100° of separation between the attended and masking voice there is near equal CDI.

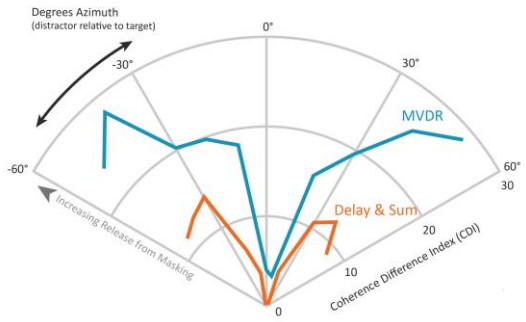

Figure 3. Spatial release from masking: how well each beamformer can reject interference.

The latency of our entire system (the time from sound production to playback) when running the two beamformers differed by approximately 17.6 milliseconds, with the delay-and-sum taking a total of 127.60 and the MVDR taking a total of 145.24 milliseconds. We should note that part of this latency is due to the necessary framing or buffering of the real-time audio (presently 21ms frame size); however, part of it is due to basic device I/O (input and output) operations on Android. This is encouraging given that great strides have been made in the time since Android 6.x was released to decrease audio pass-through latency. Even in this preliminary form, the overall output latency of our system still falls into a range that would allow sound to be naturally combined with visual cues such as mouth movements as audiovisual integration supports a synchrony window up to ~200ms [12]. The delay-and-sum had somewhat lower latency, as expected, but similar CDI performance as compared with MVDR (13.71 v 13.64 respectively), and the delay-and-sum beamformer had a CDI Efficiency marginally better than the MVDR. We note, by restricting the MVDR to two channels, performance is preserved – in a complementary frequency range – and latency is reduced, to be comparable to the simpler delay-and-sum algorithm.

VI. Conclusion

The results from this project show the potential for wearable, gaze-directed beamforming to improve speech perception in realistic environments. To our knowledge this is the first time multiple beamformers have been implemented successfully on a mobile, assistive listening platform, to capture the range of important speech frequencies with reasonable latency and computational cost.

	D&S	*MVDR*
Latency (ms)	127.60ms	145.24ms
Coherence Difference Index (CDI)	13.71	13.64
CDI Efficiency (CDI/ms)	**0.11**	**0.09**

Table 1. Latency, CDI, and CDI Efficiency at 100° separation of attended voice and masking voice

Given that the MVDR beamformer worked most effectively on lower frequencies and the delay-and-sum worked best on high frequencies, our current work aims to filter the input to restrict each algorithm to its best range and combine them to yield improved results. Future work will also demonstrate how Cochlearity performs with its real-time eye tracking and virtual 3-d rendered audio, with both hearing impaired and healthy listeners in a laboratory setting as well as in real-life, social scenarios.

References

[1] C. Alain, S. R. Arnott, S. Hevenor, S. Graham, and C. L. Grady, "What and Where the Human Auditory System," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98. National Academy of Sciences, pp. 12301–1230, 2001.

[2] B. Acar, M. F. Yurekli, M. A. Babademez, H. Karabulut, and R. M. Karasen, "Effects of hearing aids on cognitive functions and depressive signs in elderly people," *Arch. Gerontol. Geriatr.*, vol. 52, no. 3, pp. 250–252, May 2011.

[3] D. G. Blazer, S. Domnitz, and C. T. Liverman, "Hearing Loss: Extent, Impact, and Research Needs," *in Healthcare for Adults, Priorities for Improving Access and Affordability: Committee on Accessible and Affordable Hearing Health Care for Adults*. *Washington,D.C., USA,* National Academies Press (US), Sep. 2016.

[4] A. Favre-Felix, C. Graversen, T. Dau, and T. Lunner, "Real-time estimation of eye gaze by in-ear electrodes," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 4086–4089.

[5] G. Kidd, S. Favrot, J. G. Desloge, T. M. Streeter, and C. R. Mason, "Design and preliminary testing of a visually guided hearing aid," *J. Acoust. Soc. Am.*, vol. 133, no. 3, p. EL202-EL207, Mar. 2013.

[6] J. Hart, D. Onceanu, C. Sohn, D. Wightman, and R. Vertegaal, "The Attentive Hearing Aid: Eye Selection of Auditory Sources for Hearing Impaired Users," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* vol. 5726, pp. 19-35, 2009.

[7] T. L. Marzetta, "Self-Steering Directional Hearing Aid and Method of Operation Thereof," U.S. Patent 2010/0074460 A1, March 25, 2010.

[8] G. Kidd, C. R. Mason, V. Best, and J. Swaminathan, "Benefits of Acoustic Beamforming for Solving the Cocktail Party Problem," *Trends Hear.*, vol. 19, no. 0, p. 23312165155933 8, 2015.

[9] N.-V. Vu, H. Ye, J. Whittington, J. Devlin, and M. Mason, "Small footprint implementation of dual-microphone delay-and-sum beamforming for in-car speech enhancement," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 1482–1485.

[10] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[11] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New Insights Into the MVDR Beamformer in Room Acoustics," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no.1, pp. 158-170, 2010.

[12] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Temporal window of integration in auditory-visual speech perception," *Neuropsychologia*, vol. 45, no. 3, pp. 598–607, Jan. 2007.