Published in final edited form as:

Ear Hear. 2018 ; 39(4): 756–769. doi:10.1097/AUD.0000000000000532.

Evaluating the Performance of a Visually Guided Hearing Aid Using a Dynamic Auditory-visual Word Congruence Task

Elin Roverud¹, Virginia Best¹, Christine R. Mason¹, Timothy Streeter², and Gerald Kidd Jr.¹

¹Department of Speech, Language & Hearing Sciences, Boston University, Boston, MA, USA

²EarLens Corporation, Menlo Park, CA, USA

Abstract

Objectives—The “visually guided hearing aid” (VGHA), consisting of a beamforming microphone array steered by eye gaze, is an experimental device being tested for effectiveness in laboratory settings. Previous studies have found that beamforming without visual steering can provide significant benefits (relative to natural binaural listening) for speech identification in spatialized speech or noise maskers when sound sources are fixed in location. The aim of the present study was to evaluate the performance of the VGHA in listening conditions in which target speech could switch locations unpredictably, requiring visual steering of the beamforming. To address this aim, the present study tested an experimental simulation of the VGHA in a newly designed dynamic auditory-visual word congruence task.

Design—Ten young NH and eleven young HI adults participated. On each trial, three simultaneous spoken words were presented from three source positions (−30, 0 and 30° azimuth). An auditory-visual word congruence task was used in which participants indicated whether there was a match between the word printed on a screen at a location corresponding to the target source and the spoken target word presented acoustically from that location. Performance was compared for a natural binaural condition (stimuli presented using impulse responses measured on KEMAR), a simulated VGHA condition (BEAM), and a hybrid condition that combined lowpass-filtered KEMAR and highpass-filtered BEAM information (BEAMAR). In some blocks, the target remained fixed at one location across trials and in other blocks the target could transition in location between one trial and the next with a fixed but low probability.

Results—Large individual variability in performance was observed. There were significant benefits for the hybrid BEAMAR condition relative to the KEMAR condition on average for both NH and HI groups when the targets were fixed. Although not apparent in the averaged data, some individuals showed BEAM benefits relative to KEMAR. Under dynamic conditions, BEAM and BEAMAR performance dropped significantly immediately following a target location transition. However, performance recovered by the second word in the sequence and was sustained until the next transition.

Corresponding author: Elin Roverud, Department of Speech, Language & Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, MA 02215, USA. erover@bu.edu.

Conflicts of Interest: No conflicts of interest are declared by any of the authors.

Conclusions—When performance was assessed using an auditory-visual word congruence task, the benefits of beamforming reported previously were generally preserved under dynamic conditions in which the target source could move unpredictably from one location to another (i.e., performance recovered rapidly following source transitions) while the observer steered the beamforming via eye gaze, for both young NH and young HI groups.

INTRODUCTION

In many social situations, we communicate with and attend to a partner in settings with multiple background competing talkers distributed in different spatial locations (e.g., at meetings, at parties, or in noisy restaurants). The process of selecting and attending to one talker among several possible talkers often involves the concerted actions of both vision and audition. At times, we look at the “target” talker – the one to which we wish to attend – to identify the source and to extract speechreading cues (Sumby and Pollack 1954) as well as other nonverbal cues such as the talker’s emotional state, while we also train our auditory attention on this talker. Adding a layer of difficulty to this already challenging task, we may also participate in and attend to a conversation with multiple communication partners in the presence of competing talkers. Usually, the participants in a conversation take turns speaking so that the target talker changes dynamically and sometimes unpredictably. This particular situation requires not only spatially selective auditory attention, but also volitional switching of selective auditory attention. The purpose of the present study was to evaluate the efficacy of an experimental device – a visually guided hearing aid – in laboratory-based multitalker listening situations in which the desired target talker location could switch unpredictably. Performance with a simulation of this device was compared to performance when listeners volitionally switched spatial auditory attention.

Spatial Auditory Attention with Stationary Sources and Hearing Impairment

For many hearing-impaired (HI) listeners, listening to speech in the presence of competing background speech is especially challenging. It has been shown that the difficulty experienced by HI listeners attending to a target talker in stationary speech-on-speech conditions may be explained in part by a reduced ability to take advantage of the spatial separation of competing sources. This point has been demonstrated in laboratory studies revealing reduced “spatial release from masking” (SRM; decrease in masked thresholds when masker sources are separated from the target in azimuth re: when sources are colocated) in HI as compared to normal-hearing (NH) listeners (e.g., Gelfand et al. 1988; Peissig and Kollmeier 1997; Marrone et al. 2008a, Glyde et al. 2013). SRM is thought to depend on the successful use of acoustic differences caused by attenuation of the sound by the head, internal computations to improve the signal to noise ratio (SNR) within auditory channels or across time and frequency (i.e., “binaural analysis”; e.g., Durlach and Colburn 1978; Wan et al. 2014) and higher-level processes involved in source segregation, selection and, for speech, message comprehension (e.g., Bronkhorst 2015).

Most current attempts to solve this difficulty in speech mixtures for HI listeners first focus on ensuring that the speech information is accessible and audible to the listener. The idea here is that, through amplification (e.g., via hearing aids), the target signal may become

sufficiently audible for comprehension and for the acoustic cues for SRM to be available. However, amplification alone does not compensate fully for this problem. Normal SRM in multi-source environments is not fully achieved even with bilateral hearing aids (Marrone et al. 2008b), although the issue of whether standard amplification schemes restore sufficient speech information to support SRM is a matter of active research (Glyde et al. 2015; Best et al. 2017a). Furthermore, sensorineural hearing loss often leads to distortions/disruptions in spectral and temporal processing which cannot be compensated for via amplification (e.g., Moore 1996). On a practical level, current hearing aid amplification strategies that only provide gain and amplitude compression, or that reduce steady-state background sounds via “noise reduction” algorithms, do not restore normal function in sound fields containing speech mixtures. Survey studies show that difficulty with speech in noise is one of the top-reported reasons for discontinuing hearing aid use (Kochkin 2007; Takahashi et al. 2007).

Other current attempts to improve speech reception in noisy/multitalker settings take advantage of the fact that there is often at least some degree of spatial separation among sources/talkers in the soundfield in these situations. By selectively enhancing sounds from one direction while attenuating sounds from other directions (directional amplification/beamforming) an SNR improvement can be provided, as long as the unwanted sources are sufficiently separated from the target. Directional microphones, commercially available on most behind-the-ear and some in-the-ear hearing aids, selectively attenuate sources away from the axis of focus, which typically is oriented toward the front of the user. It has been shown in laboratory-based studies that directionality in general can lead to lower (improved) speech reception thresholds in the presence of background speech relative to those obtained with omnidirectional amplification (e.g., Amlani 2001; Ricketts and Hornsby 2006). Additionally, there have been reports of improved subjective ratings of real-world benefit for hearing aids with directional amplification relative to hearing aids with omnidirectional amplification, although equivocal results are not uncommon (e.g., Bentler 2005).

The most common type of directional hearing aid employs two microphones mounted on the aid in relatively close proximity. However, because the strength of the directional response generally increases as the number of microphones increases (Stadler and Rabinowitz 1993), the spatial selectivity that may be achieved using this conventional approach is somewhat limited. A number of researchers have explored the potential benefits of beamforming using a greater number of microphones. An array of spatially distributed microphones, commonly four, may be arranged on a headband worn across the head (Soede et al. 1993; Desloge et al. 1997). Alternatively, the two pairs of microphones on ear-level hearing aids may be used as the four microphones with which to apply beamforming (Picou et al. 2014). Although beamforming with more than two microphones is more spatially selective, the primary output of a beamformer is a single channel which inherently contains no binaural information. Thus, any benefit from the improved SNR could be offset by the loss of natural binaural cues and lack of spatial awareness (cf. Best et al., 2017c). Despite this potential disadvantage, some laboratory studies have reported a benefit of beamforming for speech recognition in background noise relative to no directional processing (Soede et al. 1993) and relative to directional configurations with two microphones (e.g., Desloge et al. 1997; Kidd, in press). In this laboratory, performance on a sentence recognition task with a beamformer (beamformer-processed stimuli presented under headphones, referred to as “BEAM” –

see Reference Note 1) has been compared to performance with stimuli containing binaural information (stimuli spatialized using impulse responses recorded through KEMAR manikin's ear canals; referred to as "KEMAR") in presence of spatialized speech maskers, with mixed results (Favrot et al. 2013, Kidd et al. 2013; Kidd et al. 2015).

To offset the loss of binaural cues, researchers have examined techniques of combining beamforming and binaural information (Desloge et al. 1997; Doclo et al. 2008; Picou et al. 2014; Best et al. 2015; Kidd et al., 2015; Neher et al., 2017; Best et al., 2017c). Benefits with these binaural cue-preserving beamformers have been demonstrated for speech recognition in noise (Desloge et al. 1997; Picou et al. 2014). In Kidd et al. (2015), a microphone processing condition called "BEAMAR" was tested which combined KEMAR-processed stimuli below 800 Hz and BEAM-processed stimuli above 800 Hz. This hybrid condition aimed to preserve binaural information in the low frequencies while preserving the SNR boost afforded by the beamformer in the high frequencies. Performance with this BEAMAR condition was shown to be better than with full-spectrum KEMAR (binaural) information or full-spectrum BEAM for both NH and HI listeners in spatialized speech maskers.

Intentional Switching of Spatial Auditory Attention in NH and HI

Adding challenges to the issues just discussed, there are also settings where it may be necessary to rapidly switch our attention from talker to talker (and therefore from location to location), all in the presence of competing background noise. There has been surprisingly little research reported on the intentional switching of auditory attention in the presence of competing talkers. The relevant studies in this domain typically ask listeners to respond in some way to information conveyed by a target talker, selected from multiple simultaneous talkers. In some studies, the target was indicated by a particular call sign (Brungart and Simpson 2007; Lin and Carlile 2015) or key word (Getzmann et al. 2015) spoken by one of the talkers, and the location or voice of the talker uttering this word changed (sometimes unexpectedly) throughout a series of trials. In other studies, the target was indicated by a visual cue specifying the intended target location (Best et al. 2008a) or target talker sex (Koch et al. 2011), and this visual cue changed throughout a series of trials. Results from these studies demonstrated a decrease in performance and/or an increase in response times following a target transition as compared to when the target was fixed for listeners with NH.

For listeners with hearing impairment, these listening situations just described (attending to one talker or following a conversation among multiple communication partners in the presence of competing speech) pose a particular challenge. HI listeners have provided subjective ratings of their own capacity in these domains as near complete inability (Noble and Gatehouse 2006). There are few objective data available regarding how well HI participants follow transitions in sources in multi-talker listening environments as compared to NH listeners. Best et al. (2008b) found poorer performance in HI than in NH individuals for the task of identifying strings of digits randomly transitioning in spatial location in the presence of spatially separated reversed-speech maskers, but a reference fixed condition was

REFERENCE NOTES 1: Throughout the paper, "BEAM" refers to a listening condition involving the beamforming microphone array, which is distinct and unrelated to the trademarked BEAM® technology from Cochlear Corporation.

not also tested. Overall, our understanding of the ability of NH listeners to follow transitions in sources in multi-talker listening situations is currently quite limited, and even less is known objectively about the capabilities of HI listeners in these situations.

The Visually Guided Hearing Aid

Following transitions or turn-taking in conversations with beamforming technology requires re-focusing the acoustic look direction (ALD). Automatic signal-based steering algorithms can re-focus directivity based on the detection of speech in the environment (e.g., Chalupper et al. 2011; Adiloglu et al. 2015). However, this technique is not expected to be as effective in speech mixtures because the algorithm depends on spectro-temporal differences between the target and the background. In dynamic multi-talker situations, user-directed steering may be more effective than automatic systems because only the user can determine the speech signal of interest. If the ALD is fixed with respect to the head (e.g., aimed towards the front of the wearer), the listener must turn their head or torso to redirect the ALD. Given the frequent and sometimes unpredictable turn-taking involved in some multi-participant conversations, regularly reorienting the ALD in this way in an attempt to follow these transitions may be impractical and socially unacceptable. In this laboratory, we are testing the concept of a “visually guided hearing aid” (VGHA; Best et al., 2017b; Favrot et al. 2013; Kidd et al. 2013; Kidd et al. 2015; Kidd, In Press). The VGHA employs a beamforming microphone array mounted on a band worn across the top of the head. The outputs of the array microphones are combined to optimize directionality for a specified ALD which is determined by sensing eye gaze using an eye tracker. In its wearable form, the user adjusts the beamforming focus with head, torso – and, importantly – eye movements, which may lead to more rapid and intuitive ALD reorientation.

Only two previous studies have investigated the visual-guidance component of the VGHA (Kidd et al. 2013; Best et al., 2017b). In Kidd et al. (2013), NH listeners identified a target digit spoken in a mixture of five spoken digits. The target was indicated by the location of a marker visually displayed one second in advance on the monitor screen. Kidd et al. reported that performance was approximately equal in KEMAR and BEAM conditions for the NH listeners tested. In that experiment HI listeners were not tested nor was the hybrid BEAMAR microphone condition described or tested. In Best et al. (2017b), NH and HI listeners were tested in KEMAR, BEAM, and BEAMAR conditions in a question and answer task where participants indicated whether or not the answer was a correct response to the preceding question. Target talkers spoke questions and answers with simultaneous visual cues indicating the spatial location of each item in the presence of background masking speech. Generally, performance was better for the BEAM and BEAMAR conditions relative to the KEMAR condition for both NH and HI groups for fixed-target location conditions, but these performance benefits were reduced or eliminated for dynamic conditions. However, because the target transitioned between every question and answer in the dynamic conditions, performance “recovery” could not be analyzed following a location transition.

The aim of the present study was to determine whether the visual guidance system used in the VGHA could preserve the benefits of acoustic beamforming under conditions where the target source transitioned in location abruptly and unpredictably. Furthermore, we aimed to

examine recovery of performance following a source transition. This was assessed using an experimental simulation of the VGHA using a new type of auditory-visual task. In this task, concurrent stimuli (words) were presented both auditorily (via earphones) and visually (printed on a display monitor). The observer was required to follow the visual target with their eyes and indicate by a simple button press whether the auditory and visual stimuli were the same (“congruent”) or not, a comparison process that presumably involves accessing the mental lexicon via both modalities in parallel (e.g., Rollins and Hendricks 1980; Hanson, 1981; Dijkstra et al. 1989; Repp et al. 1992). This simple, eyes-free response was used so as not to interfere with the visual guidance. In this AV congruence task, three auditory words were presented from three different source locations while a single printed target word was displayed visually at only one of the three locations. Because the printed word designated the target location, it was the means for directing both auditory and visual attention to the correct source and it was the means by which the target location transitions were indicated to the observer. Because the transitions occurred unpredictably (i.e., were randomly determined according to a specified probability) and responses were obtained on a word-by-word basis, an analysis of performance following target location transitions was undertaken.

MATERIALS AND METHODS

Participants

Ten individuals with NH (all female) and 11 individuals with HI (six female) who were all native speakers of English participated in the study. An eleventh NH subject was initially run, but was excluded from the analysis due to an inability to perform the task above chance performance. The NH listeners ranged in age from 19 to 24 years, and had audiometric thresholds <20 dB HL from 250 to 8000 Hz. The listeners with hearing loss ranged in age from 19 to 41 years (mean: 24.3; median: 22) and all had long-standing, stable, bilateral, symmetrical sensorineural hearing loss. The audiometric thresholds for each HI listener, and the averaged audiometric thresholds for the NH listeners are shown in Fig. 1. All participants provided their signed consent to participate in the study and all procedures were approved by the Boston University Institutional Review Board.

Equipment

Participants completed the experiment within a double-walled IAC booth. Auditory stimuli were processed through MATLAB 2011b (MathWorks Inc., Natick, MA) software at a sampling rate of 44.1 kHz, delivered to a RME HDSP 9632 24-bit soundcard (ASIO) and presented to the listener through Sennheiser HD280 Pro headphones. Listeners were seated at a distance of 21 inches in front of a 34" UltraWide, LG 34UM64-P monitor (32.5-inches wide from screen edges) on which visual stimuli were displayed. A *TOBII* eyex eye tracker affixed to the bottom of the monitor tracked the participant's eye movements, and was calibrated for each individual participant using the *TOBII* calibration software. During the experiment, the eye tracker readings were probed by the Matlab software every 98 ms and rounded to the nearest 2° increment (within the range of -40° to +40°). All auditory conditions were simulated under headphones. A chair-mounted head and neck rest comfortably restrained the participant's head, preventing overt head movements so as to minimize the influence of visual and auditory misalignment and to require participants to

follow visual stimuli with their eyes. The subject made responses with a hand-held USB number keypad.

Stimuli and Processing

On each test trial, listeners were presented with three simultaneous spoken words – one target word and two masker words, time-aligned at their onsets. The words were selected from a corpus of 44 monosyllabic words spoken with neutral inflection by 8 young-adult females, developed and recorded for this laboratory (Kidd et al. 2008). The corpus consists of *name*, *verb*, *number*, *adjective*, and *object* categories, each with eight exemplars, and a *conjunction* category with four exemplars. The average word duration was 691 ms (range: 286–1244 ms). On each trial, the target word was randomly selected from the full corpus. The two masker words were then randomly selected from the same category as the target word, but were different from the target word and from each other. The voices speaking each word were different and were randomly selected on each trial. All words were bandpass filtered between 80 and 8000 Hz.

Each of the words was convolved with the microphone condition-specific impulse responses. The impulse responses differed in each condition in the manner by which the recordings were made and in whether eye gaze could influence the processing. All impulse responses were recorded in a mildly reverberant sound booth (see BARE room condition in Kidd et al. 2005a) using loudspeakers positioned at a distance of 5 feet at 0°, and $\pm 30^\circ$ azimuth. The target and two maskers each occurred at one of the three source positions, with no two words in the same location.

As in Kidd et al. (2015) and Best et al. (2017b), three microphone conditions were tested: KEMAR, BEAM, and BEAMAR. In the KEMAR condition, impulse responses were recorded from the microphones located in the ear canals of a KEMAR manikin. Each of the words was convolved with the desired source azimuth impulse response for each ear. The goal of this processing was to provide the stimuli with binaural information so that each word would be perceived as originating from a distinct spatial location. This condition served as a reference for how each listener used spatial auditory attention in this experiment. The eye tracking data was irrelevant in this condition.

In the BEAM condition, the impulse responses for each source position were recorded through the VGHA microphone array worn on the KEMAR manikin's head. This microphone array is very similar to that described in Favrot et al. (2013), Kidd et al. (2013), and Kidd et al. (2015) except that the array now uses microelectro-mechanical systems (MEMS) microphones for increased durability. The array consists of four sets of four omnidirectional microphones arranged on a flexible headband (see Fig. 2 for schematic). The outputs of the microphones are weighted and combined to optimize directional processing for a particular ALD – details of the algorithm used are described in Stadler and Rabinowitz (1993). Each of the three words in a trial was convolved with the microphone array impulse responses for the relevant source position. The directional focus (or ALD) of the BEAM was controlled live – updated every 23 ms due to processing delays – based on the eye tracker reading. As noted previously, the eye tracker was independently queried and the reading updated every 98 ms. Because BEAM processing delays were independent from

the delays in eye tracker querying, the total delay (between change in eye gaze and a signal change) ranged from 23 to 121 ms. The broadband spatial tuning characteristics of the BEAM as a function of look position (for -30° to $+30^\circ$ ALDs) for each of the three source positions used in this study are shown in Fig. 3A. It is apparent from this figure that looking 30° away from a source can lead to up to 10 dB of (broadband) attenuation of that source (e.g., solid line at $\pm 30^\circ$ focus re: 0° focus, or dashed line at 0° focus re: 30° focus). Fig. 3B takes a cross-section at three VGHA focus positions from panel A and plots these data as a function of source position. These attenuation patterns were measured in 7.5° increments, but the symbols show the source positions used in the current study. For reference, KEMAR right and left ear broadband level responses are shown for the same source positions. For each VGHA focus angle, the on-focus source was between 6 and 10 dB higher in level than the other two off-focus sources. However, note that there is not more than 10 dB of attenuation, even for sources 60° away from the ALD. Fig. 3C deconstructs the broadband attenuation characteristics of the BEAM into five 1/3-octave band frequencies for a focus of 0° . Note that the greatest attenuation of off-look sources occurs for higher frequencies. The single-channel BEAM-processed stimuli were presented diotically and, as such, contained no relevant spatial information; instead, sources outside of the BEAM focus were attenuated.

In a third hybrid condition (“BEAMAR”), each of the three words on each trial were low-pass and high-pass filtered. The low- and high-pass filters were created by applying a Hann window to ideal frequency-domain filters, and the crossover point of these filters was 689 Hz. The low-pass filtered stimuli were processed with KEMAR impulse responses and the high-pass filtered stimuli were processed with the VGHA BEAM for the relevant source positions, and these frequency portions were then recombined. This condition provides binaural information in the low frequencies (where the BEAM directionality is poorer, see Fig. 3C) and the BEAM SNR benefit in the higher frequencies.

The stimulus levels were normalized so that a word presented at 0° azimuth in BEAM and KEMAR conditions (and with a look direction of 0° in the BEAM condition) had an earphone output level of 55 dB SPL for the NH listeners. As the normalization was done based on overall level, the natural variations in spectra across microphone conditions were unchanged. The BEAMAR condition was a combination of BEAM and KEMAR at the normalized levels, and so potentially could have a different overall level. However, the resulting level of the BEAMAR condition was very near to the BEAM and KEMAR conditions at 53 dB SPL. For the HI listeners, individualized gain was also applied to the stimuli using the NAL-RP gain formula (Byrne et al. 1991) following the stimulus convolution with the appropriate impulse responses. The nominal 55 dB SPL level was used to avoid stimulus peak clipping and listener loudness discomfort following the application of linear gain for the HI listeners.

Testing Procedures

On each test trial, synchronous with the onset of the three spoken words, one word was printed on the monitor at one of three screen positions (left, center, and right). These screen positions corresponded to -30 , 0 , and $+30^\circ$ look angles, respectively, and intentionally matched the angles/azimuths of the auditory stimuli in conditions where relevant spatial

information was available (i.e., KEMAR and BEAMAR). Subjects were instructed to maintain focus on the printed word with their eyes but to always keep their head facing forward. The auditory target (the word to attend to) was indicated by the location of the text printed on the screen. In the KEMAR and BEAMAR conditions, the location of the text indicated the auditory spatial location of the target word (e.g., if the word was printed on the left side of the screen, the target word was the word that sounded like it was coming from the left). In the BEAM condition, the location of the text did not correspond to the auditory spatial percept, but looking at the printed word attenuated the non-target words and thus the target word was “boosted” relative to the masker words. Subjects were instructed to attend to the word coming from the direction of the text (KEMAR condition), to the “boosted” (higher level/louder) word (BEAM condition), or both (BEAMAR condition). Schematics of a single trial in these three microphone conditions are shown in Fig. 4. Subjects were familiarized with each microphone condition prior to testing, and then were provided with training trials with feedback in order to acquaint them with the cues necessary to identify the target in each of these conditions.

Participants responded as to whether the word printed on the screen matched the auditory target word – an auditory-visual (AV) word congruence – by pressing a button on the keypad corresponding to “yes” or “no”. They were instructed to respond as quickly and as accurately as possible. Fig. 4 gives an example of a congruent trial in each of the three microphone conditions, where the correct response is “yes” (a “hit”). An AV congruence occurred with a probability of .5 on each trial. In instances where a target AV congruence did not occur, the text on the screen did not match the acoustic target word but *did* match one of the masker words at the other source positions. Thus, the text always matched one of the spoken words, but the participant’s task was to indicate whether the text matched the acoustic target word.

The printed word was displayed until the subject responded. During test trials, once the subject responded, the subsequent trial was presented 250 ms later. During training trials, “correct/incorrect” feedback was printed on the screen for 500 ms after the participant’s response, and the subsequent trial began 250 ms later. Trials were concatenated to form “runs” of a given condition, each consisting of 54 trials.

The effect of a target location transition was examined by comparing performance when the target was fixed in location (“Fixed”) to performance when the target could move unpredictably from trial to trial (“Moving”). In the Fixed condition, the word printed on the screen (and, therefore, the auditory target) remained in the same location across trials. The run of 54 trials was divided into 3 sub-blocks of 18-trial runs, where each of the locations (“Left”, “Center”, “Right”) was tested in random order. The subject was informed of the location prior to each sub-block. In the Moving condition, the run was an uninterrupted series of 54 trials. The target/printed word location of the first word was randomly selected. On each subsequent trial the target could switch with a probability of .2 to either of the two other locations each with equal probability. This probability was chosen based on preliminary analysis which showed sufficient trial numbers in which the target remained at a fixed location in order to examine recovery of performance after a location switch.

Additionally, preliminary listening with this transition probability revealed target locations switches that were perceived as unpredictable.

Prior to each run, the subject was informed about relevant aspects of the upcoming condition – what the microphone condition was (KEMAR, BEAM, or BEAMAR), and whether the text could move from trial to trial or was fixed (Fixed or Moving). This yielded 6 condition combinations. The subjects were informed of the microphone conditions because the auditory cue for distinguishing the target from the maskers in order to perform the task was different in each condition (see Fig. 4) – listeners had access to auditory spatial information in the KEMAR condition, but had access to a level cue (the target word was “boosted” when looking in the correct location) in the BEAM condition. A combination of these cues were available in different frequency regions in the BEAMAR condition. For testing, the set of all conditions was run in randomized order 12 times across separate sessions, yielding a total of (12×54) 648 scored trials per condition.

Training Procedures and Order of Testing

Listeners were tested in these conditions over the course of two days (sessions). The beginning half of the first session was designated as training. Listeners were first familiarized with each of the microphone conditions in a passive listening task (where no responses were made). For each of these familiarization runs, subjects were first provided with a written description of the upcoming microphone condition and then listened to a string of 18 words with the auditory word printed at the corresponding location. Subjects were exposed to each of the three microphone conditions in quiet and also with three simultaneous words, two repetitions of each (12 runs total). Listeners were next introduced to the experimental AV Congruence task with three 54-trial runs (one of each microphone condition) in quiet in an active task (“yes/no” responses were required). Finally, participants completed 12 runs of training trials (each of the six condition combinations repeated twice). The training trials were identical to the testing trials, except that participants received correct/incorrect feedback after every response. Testing blocks commenced in the second half of this first session. The beginning of the second session, the subjects were reminded of the task with three runs (each microphone condition) of training trials, and then commenced testing trials without feedback.

Analyses

Data were analyzed across Microphone and Transition conditions, and by Group. Data were further subdivided by Location due to inherent differences in “Center” and “Sides” target locations. Specifically, targets in the center (0° azimuth) were flanked on both sides by maskers, whereas targets on the left (-30°) or right ($+30^\circ$) were flanked only on one side by maskers. Furthermore, whereas conditions with targets flanked on both sides by maskers have been assessed in previous VGHA studies (Favrot et al. 2013; Kidd et al. 2013; Kidd et al. 2015), conditions with targets on the sides have only just recently been used (Best et al., 2017b). For the analysis, a performance index of sensitivity (d') was used in lieu of percent correct to minimize the influence of response bias in observed results. Response bias was calculated as $c = -0.5[z(H) + z(FA)]$ (Macmillan and Creelman), where $z(H)$ is the z-score

of the proportion of hits using a normal Gaussian probability distribution z transformation and $z(\text{FA})$ is the z -score of the proportion of false alarms.

RESULTS

Averaged Results – Bias (c) and Sensitivity (d')

Virtually all of the bias ' c ' values were negative (Mean: -1.104 , SD: 0.435), which indicates that there was a bias towards responding “yes”. This bias likely reflects the fact that there was always a match of the text on the screen with one of the words, either target or masker. When subjects erred, they appeared to be responding (erroneously) “yes” to the masker match rather than missing a target match. There is evidence from previous studies that there can be facilitatory priming effects between the two modalities in the lexical access stage that is fairly automatic (Frost et al. 1988; Repp et al. 1992). The printed word may have primed this word in the lexicon, facilitating detection of a match and making the instance of a masker match difficult to ignore.

Fig. 5 shows sensitivity (d') across conditions for NH and HI groups for data broken down into Center and Sides target positions. The data were analyzed using a repeated measures, four-way, mixed model ANOVA. Mauchly's test indicated that the assumption of sphericity had been violated in some cases; Greenhouse-Geisser corrections were applied, where relevant. The ANOVA results are shown in Table 1. All main effects were significant: Transition condition (Fixed, Moving), Group (NH, HI), Location (Center, Sides), and Microphone type (KEMAR, BEAM, BEAMAR). The two-way interactions of Transition $\times$ Group, Location $\times$ Microphone, Location $\times$ Transition, and Microphone $\times$ Transition also were significant. The three-way interactions of Location $\times$ Microphone $\times$ Group, Location $\times$ Transition $\times$ Group, and Location $\times$ Microphone $\times$ Transition were significant. The four-way interaction was not significant. These significant results will be examined further in the following sections. Given the goals of the study, special focus will be paid to effects of Transition, Group, and Microphone. As the effect of Location was significant (performance for Sides targets was overall better than for Center targets) and interacted significantly with each factor, these other effects will be explored for Center and Sides locations separately.

Averaged Results – Analysis of Target Transitions

It was apparent in Fig. 5 that there were decreases in performance in some cases for the Moving condition relative to the Fixed condition. Statistical analysis revealed that the effect of Transition and the interactions involving Transition were significant, as well (Table 1). Fig. 6 shows the averaged results from Fig. 5, but the Moving runs are broken down into number of trials after an unexpected target location transition to examine the recovery of performance after this move. Performance in the 1st, 2nd, and 3rd trials after a transition are shown. Additionally, performance is shown for the 4th and all following trials after a transition (4+) and for the 2nd and all subsequent trials following a transition (2+). Performance in the Fixed runs is shown for comparison as the unconnected right-most symbols in each panel. Within each panel, it is evident that performance was generally poorest the 1st trial after an unexpected target location move. By the 2nd trial after this move, performance improved and was similar to that in the Fixed runs. To test these observations,

post hoc Sidak-corrected one-sided t tests were performed for the 1st trial vs. Fixed as well as 2nd trial vs. Fixed for each location, microphone condition, and group. Asterisks in Fig. 6 indicate performance significantly lower than the Fixed performance. In the BEAM and BEAMAR conditions, performance the 1st trial after a transition was significantly different (lower) than performance in the Fixed run. In the KEMAR condition, performance the 1st trial after a transition was not significantly lower than performance in the Fixed run, with the exception of the NH Sides condition. Thus, overall, the 1st trial “switching cost” was significant for BEAM and BEAMAR conditions but not for KEMAR. However, performance by the 2nd trial after a transition was not significantly different from performance in the Fixed run of that condition (with the exception of the NH BEAMAR Center condition). This indicates that the decreases observed in the Moving condition (Fig. 5) were almost entirely a result of performance decrements the 1st trial after a target move in the BEAM and BEAMAR conditions, and that performance generally recovered to Fixed levels by the 2nd trial in the run. To examine performance without these 1st trial deficits, in subsequent analyses, the 2+ and Fixed data were combined (averaged) for each subject.

Individual Results – Analysis of Effect of Hearing Loss

The significant effect of Group generally reflects the poorer overall performance for the HI group than for the NH group as seen in Fig. 5, despite the fairly large variability (error bars representing standard deviations) within each group. Large performance variability across subjects has been observed in previous studies of the VGHA. Kidd et al. (2015) demonstrated that some of the variability in KEMAR and BEAM performance for targets fixed at 0° could be accounted for by quiet speech-reception thresholds in the HI listeners. Thus, it is feasible that at least some of the variability observed here may be explained by differences in hearing level. The relationship between performance in each microphone condition (KEMAR, BEAMAR, and BEAM) and 3-frequency pure-tone average (PTA; average of thresholds at 0.5, 1 and 2 kHz) for NH and HI listeners for the average of the Fixed and 2+ data (see Fig. 6) are plotted in Fig. 7. Best-fit regression lines, slopes, and Pearson correlation coefficients are also shown. The correlation coefficients were slightly weaker for the KEMAR conditions, but all correlations were significant. However, it appears that these significant correlations were driven primarily by the overall group difference between NH and HI. Separate analyses of the HI group data revealed significant correlations in all conditions except for BEAM Center. This is consistent with Kidd et al. (2015) who found a significant relationship for the HI group with the Center targets tested in that study for KEMAR but not for BEAM. The present findings indicate that a significant relationship holds for BEAMAR, as well. None of the correlations were significant for the NH data independently.

Individual Results – Analysis of Microphone Condition and VGHA Benefits

One goal of the present study was to determine whether VGHA beamforming processing offered any benefit relative to spatially selective auditory attention using binaural information in this task. The significant effect of Microphone condition indicates differences among KEMAR, BEAMAR, and BEAM. The significant interaction of Location $\times$ Microphone indicates that the differences among the microphone conditions depended on whether the target was in the Center or on the Sides. And finally, the significant interaction

of Location $\times$ Microphone $\times$ Group indicates that this interaction differed for the NH and HI group. This information can also be gleaned by comparing averaged performance among microphone conditions (Figs 5 and 6), however, it is not clear if any benefits observed here were driven by large benefits in a few listeners or by benefits in most/all listeners. Each individual's KEMAR performance was subtracted from his/her BEAM or BEAMAR performance, yielding relative BEAM or BEAMAR performance. The averaged Fixed and 2+ trial after a transition in the Moving data were again used. Fig. 8 shows individual relative BEAM or BEAMAR performance as a function of PTA. Averaged relative performance within each group also is shown in the separate left column in each panel. Error bars are standard errors and asterisks depict the performance differences re: KEMAR that were significantly different from 0 (Student's *t* tests). Values greater than 0 depict BEAM or BEAMAR benefits re: KEMAR, whereas values less than 0 depict better KEMAR performance.

Regarding the BEAM conditions, the Mean data indicates no BEAM benefits relative to KEMAR. In fact, for targets on the Sides, KEMAR performance was significantly better than BEAM for NH. The individual data show that the range of BEAM performance re: KEMAR was quite large for both NH and HI. Even though there was no significant BEAM benefit in the Mean, a few individuals did obtain a BEAM benefit. There was no significant correlation between relative BEAM performance and PTA for Center targets overall. The relationship was significant for HI data alone ($r=.777$; $p=0.005$), but not for NH. There was a significant correlation between relative BEAM performance and PTA for Sides targets that was not significant for either NH or HI groups independently. Overall, the relative BEAM performance variation among listeners was partly explained by variation in hearing level such that individuals with more severe hearing loss were more likely to obtain a BEAM benefit than those with milder hearing loss. Results in Fig. 7 indicate why there was a relationship. Because relative BEAM performance involves a subtraction of KEMAR and BEAM performance, the regression line slopes of absolute performance as a function of PTA in each condition are relevant. The slope of KEMAR performance as a function of PTA is steeper than the slope of BEAM performance as a function of PTA. Thus, the change in relative BEAM performance appears to be driven more by variation in KEMAR performance than by variation in BEAM performance.

Regarding the BEAMAR conditions, the Mean data indicate significant BEAMAR benefits for NH for targets in the Center, and for HI for targets on the Sides. In Figs 5 and 6 there was evidence that average BEAMAR performance was better than average KEMAR performance for both NH and HI groups, at least 2+ trials after a location move, and primarily when the target was in the Center. The upper right panel of Fig. 8 indicates that these averaged BEAMAR benefits were driven by benefits in virtually all subjects rather than large benefits in only a few subjects. It is interesting to note, however, that more NH listeners demonstrated at least some degree of BEAMAR benefit (10 out of 10) than did HI listeners (7 out of 11). The individual data indicate no significant relationship between relative BEAMAR performance and PTA for NH, HI, or combined data. This may be explained by the similarity in regression line slopes for KEMAR and BEAMAR performance vs. PTA as shown in Fig. 7.

GENERAL DISCUSSION

Summary of Findings

The present study aimed to evaluate the VGHA under conditions where the target source could transition in location abruptly and unpredictably. To accomplish this aim, performance in an AV word congruence task was compared for headphone simulations of KEMAR, BEAM, and BEAMAR conditions in fixed and dynamic trials for NH and HI listeners. On average, in fixed conditions, BEAMAR performance was superior to KEMAR performance in NH listeners (for Center targets) and in HI listeners (for Sides targets). Conditions with visually guided beamforming (BEAM and BEAMAR) showed a particular disadvantage in the initial trial after an unexpected target location switch. However, performance in these conditions recovered to fixed levels rapidly – by the 2nd word in the trial.

Effects with a Fixed Target – Comparison to Previous Results

Although the task was different, the Fixed-target, Center-location condition in the present study resembled that used in previous studies of the VGHA from this lab. To the extent that they are comparable, the results reported here generally were consistent with the findings from these previous studies with exceptions noted below. When the target was fixed, there was no significant difference in performance with BEAM relative to KEMAR, on average, for either NH or HI groups with Center targets (see Fig. 8). This is generally consistent with Kidd et al. (2015), although the results of that study indicate that better BEAM performance relative to KEMAR can be obtained for HI listeners when maskers are more widely spatially separated from the target. It should be noted that the present results are inconsistent with Favrot et al. (2013), who did find better performance with BEAM (called “Microphone array” in that study) than with KEMAR in NH. This discrepancy may be attributed to the difficulty of the task in the present study. Recall that there was a match of the text on the screen to one of the auditory words (target or maskers) on every trial. Thus, in order to be successful in the task of detecting a *target* match, the listener needed to use the cue distinguishing the target from the maskers. In the BEAM condition, this was a level cue (listen for the “boosted” word), whereas in the KEMAR condition this was a spatial location cue (e.g., listen for the word on the left). It may have been that a level cue was less salient than the spatial cue, or that the level cue was less perceptually consistent than a spatial cue when both the word and talker changed on every trial.

The present study found better performance on average with BEAMAR relative to KEMAR for Center targets in NH and for Sides targets in HI (Fig. 8). This is consistent with Kidd et al. (2015) who reported better BEAMAR performance relative to KEMAR for Center targets in NH and HI, and with Best et al. (2017b) who reported better BEAMAR performance relative to KEMAR for Center and Sides targets in NH and HI. The superior performance in this hybrid BEAMAR condition across studies was by no means guaranteed, as this condition does not contain full-spectrum binaural information or a full-spectrum SNR boost. Rather, the signal contains binaural information (contributing to a spatialized image) in the low frequencies and an SNR boost (lacking any relevant spatial information) in the high frequencies. However, the fact that BEAMAR performance is better than either KEMAR or BEAM performance across these various studies and tasks indicates that listeners are able to make

use of the two types of information and merge them successfully. These studies examining the BEAMAR condition have used cutoff frequencies of approximately 700–800 Hz separating the two types of information. However, the relationship between performance and cutoff frequency has not yet been systematically explored. It is possible that improved BEAMAR performance may be found with a lower or higher cutoff frequency, or an altogether different method for combining these two types of information.

There was large individual performance variability across conditions in this study. Inspection of the individual data in the fixed conditions revealed that not all individuals obtained a BEAMAR benefit for targets in the Center (Fig. 8). Interestingly, fewer HI listeners obtained a BEAMAR benefit for Center targets than NH listeners. Some listeners (including the majority of HI listeners) obtained a BEAMAR benefit on the Sides. Although not apparent in the averaged relative data, some individuals showed BEAM benefits for Center and Sides target positions (Fig. 8). There was a significant relationship between PTA and BEAM benefit for Sides targets and, for the HI listeners alone, there was a significant relationship between PTA and BEAM benefit for Center targets. This suggests that listeners with more severe hearing loss might be more likely to benefit from beamforming than listeners with milder hearing loss. Hearing loss severity (represented as PTA here) may not be the only factor related to performance or benefit with a beamformer. Neher et al. (2017) reported that performance with a beamformer or a hybrid beamformer (much like the BEAMAR condition in this study) was significantly correlated with one's ability to benefit from spatial separation of speech and noise (i.e., binaural intelligibility level difference).

The present study found a significant correlation between threshold (PTA in this case) and KEMAR performance for the data from the HI listeners and when all of the data were combined (NH and HI). This significant relationship was expected given the significant relationship reported in Kidd et al. (2015) between speech reception thresholds and KEMAR performance and given the results of previous studies examining SRM in HI listeners. HI listeners have been shown to benefit less from spatial separation of maskers from a target than NH listeners (e.g., Marrone et al. 2008a), and SRM for speech maskers has been shown to be negatively correlated with audiometric thresholds (Glyde et al. 2013). What the present study adds is that BEAMAR performance was also significantly correlated with hearing loss. Overall, the consistencies across studies using fixed target locations reported here suggest that beamforming benefits remain regardless of the task used (whether AV word congruence detection, closed-set sentence recognition, or question-answer veracity judgments).

Apart from the AV congruence task itself, novel aspects of the present study included assessing performance for unexpected dynamic target location transitions (discussed more below), specifying performance in terms of sensitivity (d'), and the use of lateral targets with maskers positioned only to one side. Performance with these Sides targets was improved relative to Center targets. This is generally consistent with previous studies showing better speech reception thresholds for a target in the presence of asymmetrical maskers compared to symmetrical maskers (e.g., Bronkhorst and Plomp 1992; and see Kidd et al. 2005b).

Effects with an Unexpectedly Moving Target

A primary motivation of the present study was to examine the effects of dynamic switching on performance using the VGHA. Results showed that performance in the simulated VGHA conditions (BEAM and BEAMAR) suffered in the dynamic condition, but only on the first trial after a target location transition. Performance recovered to “Fixed” levels by the second trial after the transition. Decrements in the KEMAR condition were expected given other evidence of switching costs in spatialized speech mixtures (see Brungart and Simpson 2007), however there was only a statistically significant drop in the trial immediately after a move in one condition (NH Sides; see Fig. 5). The significant drop in performance for BEAM and BEAMAR conditions in the dynamic conditions immediately following a transition essentially was consistent with the findings of Best et al. (2017b) where target location transitions occurred between each trial. In contrast, Kidd et al. (2013) found little difference between performance in the BEAM condition and the KEMAR condition for NH listeners tested using a moving target. However, Kidd et al. presented a visual cue 1 second before the auditory stimuli were presented rather than simultaneously with the auditory stimuli. This suggests that advanced warning of a location switch may eliminate the disadvantages of location switches for the BEAM and BEAMAR conditions.

In the present study, the onset of the printed word signaling the target location transition was synchronous with the onset of the auditory stimuli. As such, participants were required to move their eyes *during* the first word after a location switch, meaning that at least some portion of this first word would not have been fully under the focus of the beam. The better performance on the first word following a transition for the KEMAR condition (or, conversely, the smaller decrement in performance) relative to the two VGHA conditions suggests that the observer can redirect the focus of attention to a new spatial location using binaural cues more quickly than by shifting the beam of amplification using eye gaze. An alternative explanation is that the timing of these events may be similar (and relatively slow) but it may be easier to perform a *post hoc* retrieval of the spatialized words stored in memory than determining whether the “wrong” word was amplified while the eyes are moving from one location to the next.

To examine the eye movement timing issue directly, the eye tracker traces during the experiment were analyzed. Unfortunately, due to a coding error, the eye tracker data were only saved for two NH listeners. The bottom panel of Fig. 9 shows a histogram of the word durations for all words spoken by all talkers in the experiment (44 words spoken by 8 talkers). There was considerable variability in the auditory word durations with a minimum word duration of 286 ms (“but”) and a maximum word duration of 1244 ms (“shoes”). The average word duration was 691 ms. The top panel of Fig. 9 shows the timing relationship between the auditory word (for a word of average duration) and the measured movement of the eyes to the correct location the 1st trial after a location transition averaged over the data obtained from these two NH listeners. Note that the eye tracker data analyzed here included the 98-ms eye tracker sampling noted in the methods. The eye tracker data were further broken down into trials in which a target moved 30° (e.g., from Center to Right) or 60° (e.g., from Left to Right). Response times (time of the button press) are also shown for these conditions for reference (as open squares). The latency of initiating eye movement following

onset of the visual cue was approximately 300 ms (288 ms for a cue 30° away and 346 ms for a cue 60° away). Eye sweeps were completed (eye gaze was within 5° of the visual target) by 465 ms for a target 30° away and by 632 ms for a target 60° away. With these eye movement latencies, beamforming would have been suboptimal for the entirety of the shortest words, and for a large part of the average-duration words. At most, the longest-duration words would have received the full amplification benefit from the VGHA for only the latter half of the word duration. The adverse effect on intelligibility that would be expected from optimally amplifying the latter half of the words used in this study is not known. The fact that directing one's eye gaze over these distances occurs on a time scale of around one one-syllable word indicates the potential importance of context and a priori knowledge (e.g., visual talker cues) in signaling source transitions in natural listening (e.g., turn-taking in conversation) for the VGHA. In the present study, no a priori visual cues were provided to simulate the most extreme and difficult case. A logical next step in this line of research is to examine how different types of cues to impending source transitions are used naturally – resulting in the redeployment of auditory selective attention from one spot to another versus sweeping highly directional amplification with an external control device as with the VGHA. In any case, the rapid recovery of performance following a target location transition in the present study suggests that any cues available to the listener in advance of a transition likely would lead to performance equivalent to that found in situations in which the target is fixed in location.

It is interesting to consider whether the redirection of auditory spatial attention is “swept” continuously in azimuth just as the eyes are swept while redirecting the VGHA beam, or if the redirection/deployment of attention to another spatial location is done discretely without sweeping across physically intervening sound sources. To attempt to answer this question, the proportion of false alarms (FAs) was calculated for each microphone condition for the trials immediately following a target location transition of 60° (left to right or right to left) when the *masker congruence* occurred for the center auditory word. If auditory spatial attention is redirected like a sweeping spotlight, one might expect a high FA rate for the KEMAR condition on par with the FA rate for a visually swept beamformer in this case. Alternatively, if auditory spatial attention is not swept in this manner, one might expect a higher FA rate for BEAM than for KEMAR. Averaged across all listeners, the BEAM FA rate (Mean: 0.38, SD: 0.16) was significantly higher than the KEMAR FA rate (Mean: 0.26, SD: 0.17) ($t = -2.415, p = 0.01$). This suggests that listeners were indeed less influenced by an intervening masker congruence when redirecting auditory spatial attention from left to right or right to left than when physically steering the microphone array via eye gaze. This is consistent with the idea that redirection of auditory spatial attention does not behave precisely like a “sweeping spotlight” but may be likened more to a spotlight that is turned off at one location and then turned on at another internally determined location.

Conclusions

The findings of this study support the conclusion that the benefit of the VGHA, as measured in our experimental simulation using the present task, is largely preserved with visual steering for conditions in which the target source moves unpredictably from one location to another. These findings are qualified by the reduced performance found for the VGHA

conditions immediately after a transition, which appears to result from the unexpected, sudden and un-cued nature of the transitions used in this task.

Additional work is needed to evaluate the effectiveness of the VGHA in realistic communication settings. In such settings, there may in fact be visual or contextual cues signaling an upcoming transition of talkers in a conversation available to the listener which may lessen disadvantages of the VGHA during transitions. On the other hand, other potential disadvantages for the VGHA may be revealed. For example, eye movements may only follow a target source during certain rapid-change segments of communication; at other times, listeners may not maintain visual attention on an intended target. Thus, future work should consider how listeners move their eyes over the long term during communication. Furthermore, this future work should involve a comparison of performance with a head-worn VGHA to that with a beamformer that is solely redirected via torso/head movements. As a final point, the current study simulated all microphone conditions under headphones, allowing for precise stimulus control. However, in future studies with a head-worn VGHA device, some unprocessed sound would reach the listener's ears via the direct sound path. Future studies of a wearable VGHA device are needed to fully understand the influence that mixtures of indirect-processed and direct-unprocessed sound would play on any benefits the VGHA may provide.

Acknowledgments

E.R. helped to design the experiment, performed the experiment, analyzed/interpreted data, and wrote the paper; V.B. assisted in interpretation of data and provided critical manuscript revision; C.R.M. assisted in interpretation of data and provided critical manuscript revision; T.S. helped to design and perform the experiment and assisted with analysis; G.K. designed the experiment, assisted in interpretation of data, and provided critical manuscript revision. We wish to thank Lorraine Delhorne for her assistance with subject recruitment. This work was supported by a grant from NIH/NIDCD (to G.K.) and by a grant from DoD/AFOSR (to G.K.). Portions of this work were presented at the 39th meeting of the Association for Research in Otolaryngology in San Diego, CA, and at the 171st meetings of the Acoustical Society of American in Salt Lake City, UT.

References

- Adiloglu K, Kayser H, Baumgartel RM, et al. A binaural steering beamforming system for enhancing a moving speech source. *Trends in Hearing*. 2015; 19:1–13.
- Amlani AM. Efficacy of directional microphone hearing aids: a meta-analytic perspective. *J Am Acad Audiol*. 2001; 12:202–214. [PubMed: 11332520]
- Bentler RA. Effectiveness of directional microphones and noise reduction schemes in hearing aids: A systematic review of the evidence. *J Am Acad Audiol*. 2005; 16:473–484. [PubMed: 16295234]
- Best V, Mason CR, Swaminathan J, et al. Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures. *J Acoust Soc Am*. 2017a; 141:81–91. [PubMed: 28147587]
- Best V, Roverud E, Streeter T, et al. The benefit of a visually guided beamformer in a dynamic speech task. *Trends in Hearing*. 2017b; 20:1–11.
- Best V, Roverud E, Mason CR, Kidd G Jr. Examination of a hybrid beamformer that preserves auditory spatial cues. *J Acoust Soc Am Express Letters*. 2017c; 142:EL369–374.
- Best V, Mejia J, Freeston K, et al. An evaluation of the performance of two binaural beamformers in complex and dynamic multitalker environments. *Int J Audiol*. 2015; 54:727–735. [PubMed: 26140298]
- Best V, Ozmeral EJ, Kopco N, et al. Object continuity enhances selective auditory attention. *PNAS*. 2008a; 105(35):13174–13178. [PubMed: 18719099]

- Best V, Marrone N, Mason CR, et al. Effects of sensorineural hearing loss on visually guided attention in a multitalker environment. *J Assoc Res Otolaryngol*. 2008b; 10:142–148. [PubMed: 19009321]
- Bronkhorst AW. The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Atten Percept Psychophys*. 2015; 77(5):1465–1487. [PubMed: 25828463]
- Bronkhorst AW, Plomp R. Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *J Acoust Soc Am*. 1992; 92(6):3132–3139. [PubMed: 1474228]
- Brungart DS, Simpson BD. Cocktail party listening in a dynamic multitalker environment. *Perception & Psychophysics*. 2007; 69(1):79–91. [PubMed: 17515218]
- Byrne, D., Parkinson, A., Newall, P. Modified hearing aid selection procedures for severe/profound hearing losses. In: Studebaker, G.Bess, F., Beck, L., editors. *The Vanderbilt Hearing Aid Report II*. York Press; Parkton: 1991. p. 295-300.
- Chalupper J, Wu YH, Weber J. New algorithm automatically adjusts directional system for special situations. *The Hearing Journal*. 2011; 64(1):26–33.
- Desloge JG, Rabinowitz WM, Zurek PM. Microphone-array hearing aids with binaural output – part I: Fixed-processing systems. *IEEE Transactions on Speech and Audio Processing*. 1997; 5(6):529–542.
- Dijkstra T, Schreuder R, Frauenfelder UH. Grapheme context effects on phonemic processing. *Language and Speech*. 1989; 32(2):89–108.
- Doclo, S., Gannot, S., Moonen, M., et al. Acoustic beamforming for hearing aid applications. In: Haykin, S., Liu, KR., editors. *Handbook on Array Processing and Sensor Networks*, (ch 9). New York, NY: Wiley; 2008.
- Durlach NI, Colburn HS. Binaural phenomena. *Handbook of perception*. 1978; 4:365–466.
- Favrot S, Mason CR, Streeter TM, et al. Performance of a highly directional microphone array in a multi-talker reverberant environment. *Proc Mtgs Acoust POMA*. 2013:1–8. 050145.
- Frost R, Repp BH, Katz L. Can speech perception be influenced by simultaneous presentation of print? *Journal of Memory and Language*. 1988; 27:741–755.
- Gelfand SA, Ross L, Miller S. Sentence reception in noise from one versus two sources: Effects of aging and hearing loss. *J Acoust Soc Am*. 1988; 83(1):248–256. [PubMed: 3343444]
- Getzmann S, Hanenberg C, Lewald J, et al. Effects of age on electrophysiological correlates of speech processing in a dynamic “cocktail-party” situation. *Frontiers in Neuroscience*. 2015; 9:1–17. [PubMed: 25653585]
- Glyde H, Cameron S, Dillon H, et al. The effects of hearing impairment and aging on spatial processing. *Ear Hear*. 2013; 34(1):15–28. [PubMed: 22941406]
- Glyde H, Buchholz JM, Nielsen L, et al. Effect of audibility on spatial release from speech-on-speech masking. *J Acoust Soc Am*. 2015; 138(5):3311–3319. [PubMed: 26627803]
- Hanson VL. Processing of written and spoken words: Evidence for common coding. *Memory and Cognition*. 1981; 9(1):93–100. [PubMed: 7231173]
- Kidd G Jr, Mason CR, Brughera A, et al. The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acta Acust Acust*. 2005a; 91:526–536.
- Kidd G Jr, Arbogast TL, Mason CR, et al. The advantage of knowing where to listen. *J Acoust Soc Am*. 2005b; 118(6):3804–3815. [PubMed: 16419825]
- Kidd G Jr, Best V, Mason CR. Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *J Acoust Soc Am*. 2008; 124(6):3793–3802. [PubMed: 19206805]
- Kidd G Jr, Favrot S, Desloge JG, et al. Design and preliminary testing of a visually guided hearing aid. *J Acoust Soc Am*. 2013; 133(3):EL202–EL207. [PubMed: 23464129]
- Kidd G Jr, Mason CR, Best V, et al. Benefits of acoustic beamforming for solving the cocktail party problem. *Trends in Hearing*. 2015; 19:1–15.
- Kidd G Jr, Mason CR, Swaminathan J, et al. Determining the energetic and informational components of speech-on-speech masking. *J Acoust Soc Am*. 2016; 140:132–144. [PubMed: 27475139]
- Kidd G Jr. Enhancing auditory selective attention using a visually guided hearing aid. *J Speech Lang Hear Res*. (In Press).

- Koch I, Lawo V, Fels J, et al. Switching in the cocktail party: Exploring intentional control of auditory selective attention. *Journal of Experimental Psychology: Human Perception and Performance*. 2011; 37(4):1140–1147. [PubMed: 21553997]
- Kochkin S. MarkeTrak VII: Obstacles to adult non-user adoption of hearing aids. *The Hearing Journal*. 2007; 60(4):24–50.
- Lin G, Carlile S. Costs of switching auditory spatial attention in following conversational turn-taking. *Frontiers in Neuroscience*. 2015; 9:1–11. [PubMed: 25653585]
- Macmillan, NA., Creelman, CD. *Detection Theory: A User's Guide*. 2nd. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers;
- Marrone N, Mason CR, Kidd G Jr. The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *J Acoust Soc Am*. 2008a; 124(5):3064–3075. [PubMed: 19045792]
- Marrone N, Mason CR, Kidd G Jr. Evaluating the benefit of hearing aids in solving the cocktail party problem. *Trends in Amplification*. 2008b; 12(4):300–315. [PubMed: 19010794]
- Moore BCJ. Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids. *Ear Hear*. 1996; 17(2):133–161. [PubMed: 8698160]
- Neher T, Wagener KC, Latzel M. Speech reception with different bilateral directional processing schemes: Influence of binaural hearing, audiometric asymmetry, and acoustic scenario. *Hear Res*. 2017; 353:36–48. [PubMed: 28783570]
- Noble W, Gatehouse S. Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the Speech, Spatial, and Qualities of Hearing scale (SSQ). *International Journal of Audiology*. 2006; 45:172–181. [PubMed: 16579492]
- Peissig J, Kollmeier B. Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *J Acoust Soc Am*. 1997; 101(3):1660–1670. [PubMed: 9069633]
- Picou EM, Aspell E, Ricketts TA. Potential benefits and limitations of three types of directional processing in hearing aids. *Ear Hear*. 2014; 35(3):339–352. [PubMed: 24518429]
- Repp BH, Frost R, Zsiga E. Lexical mediation between sight and sound in speechreading. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*. 1992; 45(1):1–20. DOI: 10.1080/14640749208401313 [PubMed: 1636006]
- Ricketts TA, Hornsby BWY. Directional hearing aid benefit in listeners with severe hearing loss. *International Journal of Audiology*. 2006; 45:190–197. [PubMed: 16579494]
- Rollins HA Jr, Hendricks R. Processing of words presented simultaneously to eye and ear. *Jour Of Exp Psych: Human Perception and Performance*. 1980; 6(1):99–109.
- Soede W, Bilsen FA, Berkhout AJ. Assessment of a directional microphone array for hearing-impaired listeners. *J Acoust Soc Am*. 1993; 94(2):799–808. [PubMed: 8370886]
- Stadler RW, Rabinowitz WM. On the potential of fixed arrays for hearing aids. *J Acoust Soc Am*. 1993; 94(3):1332–1342. [PubMed: 8408974]
- Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. *J Acoust Soc Am*. 1954; 26(2):212–215.
- Takahashi G, Martinez CD, Beamer S, et al. Subjective measures of hearing aid benefit and satisfaction in the NIDCD/VA follow-up study. *J Am Acad Audiol*. 2007; 18:323–349. [PubMed: 17580727]
- Wan R, Durlach NI, Colburn HS. Application of a short-time version of the Equalization-Cancellation model to speech intelligibility with speech maskers. *J Acoust Soc Am*. 2014; 136(2):768–776. [PubMed: 25096111]

Figure 1.
Individual audiograms for 11 HI listeners (dashed lines) and averaged audiograms for 10 NH listeners (open circles connected by dashed lines).

Figure 2.

Schematic top view of the microphone array. Each black box represents one omnidirectional microphone. There are four sets of four microphones. Within each set, four microphones are arranged in the front-back direction as two pairs of microphones. The microphones within a pair are spaced 10 mm apart. The medial two microphones in a set (separating the front-back pairs) are spaced 15 mm apart. Finally, each set is spaced horizontally 66.67 mm apart, yielding a total array length of 200 mm.

Figure 3.

Attenuation characteristics of the beamforming microphone array. **A-** Relative attenuation of a noise source located at three source positions (different line types) as a function of the beam focus. **B-** Relative attenuation of a noise source as a function of source position for three beam focus positions (dashed lines and different open symbols). The dashed lines represent attenuation for intermediate sources, which were measured every 7.5°. For reference, relative KEMAR levels for left (blue) and right (red) ears are also shown. **C-** Relative attenuation for five 1/3-octave bands with center frequencies of 500 to 8000 Hz as a function of source position with a beam focus of 0° azimuth. Symbols show the source positions used in the present study. The dashed lines represent attenuation for intermediate sources, which were measured every 7.5°.

Figure 4.

A schematic of the perceptual differences between the three microphone conditions during a single trial of the experiment. The trial shown is an example of a congruent trial for a target on the left.

Figure 5. Averaged sensitivity (d') results for NH and HI subjects. Results are broken down into targets in the Center (flanked on both sides by maskers) and targets on the Sides (flanked only on one side by maskers). For the Sides results, left and right target results have been averaged. Error bars are ± 1 standard deviation.

Figure 6.

Averaged sensitivity (d') data from Fixed runs and Move runs broken down into trials after a location transition for NH and HI subjects in all three microphone conditions for targets in the Center and on the Sides. Error bars are ± 1 standard deviation. Asterisks indicate conditions in 1 and 2 trials after a transition that were significantly different (lower) than performance in the corresponding condition for Fixed runs.

Figure 7.

Individual sensitivity (d') results plotted as a function of pure-tone average for NH (blue circles) and HI (red squares) listeners for all three microphone conditions for targets in the Center and on the Sides. The line within each panel is the regression fit to the data, with corresponding slope, Pearson correlation coefficient (r), and p-value in each panel.

Figure 8.

Individual performance for BEAM or BEAMAR relative to KEMAR as a function of pure-tone average. The Mean change re: KEMAR of each group is shown in the bounded left column in each panel; error bars are standard errors and asterisks indicate changes re: KEMAR that are statistically significantly greater than or less than 0. Values greater than 0 indicate a benefit of BEAM or BEAMAR, whereas values less than 0 indicate better performance of KEMAR. Dashed lines show fits to the data, with Pearson correlation coefficients and p values in each panel.

Figure 9.

Bottom panel - histogram of all word durations (all 44 words spoken by all 8 talkers in the experiment). Top panel - the temporal relationship between auditory word length, eye movements, and response times averaged for two NH participants in the first trial after a target location transition (across all microphone conditions). In all trials, the printed word onset was synchronous with the auditory word onset (0 ms in this figure). The timing of an auditory word (of average duration) is shown as the dashed blue line, with the offset shown as an asterisk. Note, however, that any given word could end earlier or later within the range shown by the histogram of word durations in the bottom panel. The start of the dashed red and black lines show the initiation of eye movements to a target 30° or 60° away following the visual cue onset (0 ms). The asterisks depict when eye gaze was within 5 degrees of the target look direction. The red and black squares show averaged response times.

Results of the Analysis of Variance performed for the data in Fig. 5. In bold are the significant effects. Tran= Transition (Fixed or Moving); Grp= Group (NH or HD); Loc= Location (Center or Sides); Mic= Microphone Condition (KEMAR, BEAM, or BEAMAR).

TABLE 1

Main (one-way)		Two-way		Three-way		Four-way	
Tran	F(1,19)=165.8, p<.001	Loc × Grp	F(1,19)=.07, p=.80	Loc × Mic × Grp	F(2,38)=10.1, p<.001	Loc × Mic × Tran × Grp	F(2,38)=.88, p=.422
Grp	F(1,19)=19.2, p<.001	Mic × Grp	F(1,18,22.3) =1.2, p=.3	Loc × Tran × Grp	F(1,19)= 6.2, p=.023		
Loc	F(1,19)=75.6, p<.001	Tran × Grp	F(1,19)=9.8, p=.005	Mic × Tran × Grp	F(2,38)= 2.1, p=.14		
Mic	F(1,18, 22.3)= 11.6, p=.002	Loc × Mic	F(1,15,29.17)=5.8, p=.012	Loc × Mic × Tran	F(2,38)=3.6, p=.04		
		Loc × Tran	F(1,19)=7.9, p=.011				
		Mic × Tran	F(2,38)=11.1, p<.001				