Trends in Hearing
Volume 21: 1–11
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2331216517722304
journals.sagepub.com/home/tia
SAGE

The Benefit of a Visually Guided Beamformer in a Dynamic Speech Task

Virginia Best[1], Elin Roverud[1], Timothy Streeter[1], Christine R. Mason[1], and Gerald Kidd, Jr[1]

Abstract

The aim of this study was to evaluate the performance of a visually guided hearing aid (VGHA) under conditions designed to capture some aspects of "real-world" communication settings. The VGHA uses eye gaze to steer the acoustic look direction of a highly directional beamforming microphone array. Although the VGHA has been shown to enhance speech intelligibility for fixed-location, frontal targets, it is currently not known whether these benefits persist in the face of frequent changes in location of the target talker that are typical of conversational turn-taking. Participants were 14 young adults, 7 with normal hearing and 7 with bilateral sensorineural hearing impairment. Target stimuli were sequences of 12 question–answer pairs that were embedded in a mixture of competing conversations. The participant's task was to respond via a key press after each answer indicating whether it was correct or not. Spatialization of the stimuli and microphone array processing were done offline using recorded impulse responses, before presentation over headphones. The look direction of the array was steered according to the eye movements of the participant as they followed a visual cue presented on a widescreen monitor. Performance was compared for a "dynamic" condition in which the target stimulus moved between three locations, and a "fixed" condition with a single target location. The benefits of the VGHA over natural binaural listening observed in the fixed condition were reduced in the dynamic condition, largely because visual fixation was less accurate.

Keywords

beamforming, eye tracking, hearing loss, conversations

Date received: 10 March 2017; revised: 7 June 2017; accepted: 26 June 2017

Introduction

One of the few ways of improving speech understanding in noise is via directional hearing aids (Dillon, 2012), which preferentially amplify sounds from one direction (usually the front) relative to sounds from other directions. While directional microphones are by now standard in most hearing aids, a variety of more sophisticated algorithms have also been developed that combine the signals from multiple microphones (e.g., across a pair of hearing aids, or from a microphone array) to create extremely sharp spatial tuning (e.g., Desloge, Rabinowitz, & Zurek, 1997; Doclo, Gannot, Moonen, & Spriet, 2010; Soede, Berkhout, & Bilsen, 1993). In this study, we examine performance using a beamforming microphone array that combines signals from 16 microphones mounted on a headband. The intelligibility-weighted directivity index of this array is around 9 dB, similar to values reported in the literature for other beamformers (e.g., Baumgärtel et al., 2015; Desloge et al., 1997; Kates & Weiss, 1996; Soede, Bilsen, & Berkhout, 1993; Stadler & Rabinowitz, 1993). Previous studies from our laboratory have demonstrated that this type of beamformer can enhance speech intelligibility (relative to natural binaural listening) for fixed-location, frontal speech targets amidst spatially separated maskers (Kidd, 2017; Kidd, Favrot, Desloge, Streeter, & Mason, 2013; Kidd, Mason, Best, & Swaminathan, 2015). For noise maskers, the benefit measured in speech identification tests approaches the benefit predicted acoustically (6–9 dB) and compares favorably with speech-in-noise improvements reported for other

[1] Department of Speech, Language and Hearing Sciences, Boston University, MA, USA

Corresponding author:
Virginia Best, Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA 02215, USA.
Email: ginbest@bu.edu

kinds of beamformer (e.g., Saunders & Kates, 1997; Soede, Bilsen, et al., 1993; Völker, Warzybok, & Ernst, 2015). When the maskers are competing talkers uttering highly similar phrases, however, the benefit is generally smaller than found for noise and performance may even be *worse* than for natural binaural listening. This is because the output of this kind of beamformer is a single-channel signal which, even if delivered to both ears (i.e., diotically), does not contain any binaural information. This is in contrast to the natural listening situation in which differences in time and level between the ears enable the listener to segregate competing sounds based on differences in location. In conditions where this location-based segregation is critical (e.g., when following one talker in the presence of competing talkers), the benefits of beamforming appear to be counteracted by the loss of spatial cues.

To mitigate this problem, some form of natural binaural "cue preservation" can be incorporated into the algorithm (e.g., Desloge et al., 1997; Doclo et al., 2010; Picou, Aspell, & Ricketts, 2014; Van den Bogaert, Doclo, Wouters, & Moonen, 2009). For the microphone array under consideration in this study, a hybrid version has been developed in which the beamforming is restricted to the high frequencies (where it is most effective), leaving natural acoustic cues in the low frequencies (where interaural time differences are known to be useful). Assuming the listener has no trouble binding the two parts of the spectrum containing these different types of information, this version should theoretically maintain much of the acoustic benefit of the full beamformer, while preserving some sense of spatial separation. Our previous work has shown that this hybrid version tends to provide better thresholds when a frontal target talker is masked by competing talkers placed symmetrically to either side (Kidd et al., 2015).

In many typical listening situations, the target of interest is not fixed in front of the listener, and when a listener is engaged in a group conversation, the target talker (and location) may change from moment to moment. Under these conditions, highly directional hearing aids might be detrimental in that, by design, they cause "tunnel hearing." When there is only one dominant talker, automatic steering methods show promising results (Adiloğlu et al., 2015), but this kind of solution is not appropriate when there are also unwanted talkers in the scene. What is needed is for a listener to be able to *steer* the directional beam, and to do so swiftly enough to keep up with the flow of natural conversations. Best et al. (2015) showed that when listeners had to turn their heads to steer a forward-facing binaural beamformer to off-center targets in a sentence test, the benefits were reduced compared with frontal targets. In another evaluation of the same beamformer, however, Mejia et al. (2015) asked listeners to select "acceptable noise levels" while following single-talker monologues or two-person conversations and found that benefits of the device were robust to spatial variations. In our laboratory, we have been investigating an approach in which the acoustic look direction (ALD) of the microphone array described earlier is steered according to the user's eye position, which is detected using an eye tracker. Theoretically, eye movements have advantages over head movements in that they are faster and are quite a natural part of normal conversations. However, the benefits of beamforming under visual guidance for moving or off-center targets have not yet been investigated.

The goal of the current study was to test the "visually guided hearing aid" (VGHA) under conditions that capture some aspects of real-world communication settings and, critically, incorporate spatially dynamic stimuli. We made use of a new speech test based on question–answer pairs (Best, Streeter, Roverud, Mason, & Kidd, 2016) in which the participant is required to indicate whether the answer is true or false. This test is well suited to the current purposes for several reasons. First, the simple binary response may be obtained rapidly via a keypress, which allows multiple trials to be presented in succession to create ongoing listening conditions. Second, because the information on each trial is distributed across two parts (the question and the answer), it is possible to introduce intratrial changes in location that challenge the visual guidance component by requiring users to steer the VGHA rapidly. Finally, the response mode does not engage the eyes, leaving them free for operating the VGHA.

Methods

Participants

The participants in the study were 14 young adults, 7 of whom had normal hearing (NH; seven female) and ranged in age from 21 to 24 years (mean ± standard deviation 22 ± 1 years). The other seven had bilateral sensorineural hearing impairment (HI, two female) and ranged in age from 19 to 41 years (mean ± standard deviation 25 ± 8 years). Recruitment was deliberately focused on young HI listeners, even though it restricted the pool of participants, to avoid age-related factors that might have introduced potentially confounding variables had we recruited from the more common older HI population. Pure-tone averages (PTA; mean threshold across 0.5, 1, and 2 kHz) ranged from −3 to 12 dB HL in the NH group and from 21 to 59 dB HL in the HI group. The HI listeners had relatively symmetric losses, defined as a difference between the ears of no more than 10 dB HL at any of the standard audiometric frequencies. Audiograms for each listener (averaged across the ears) are shown in Figure 1, along with the mean audiograms

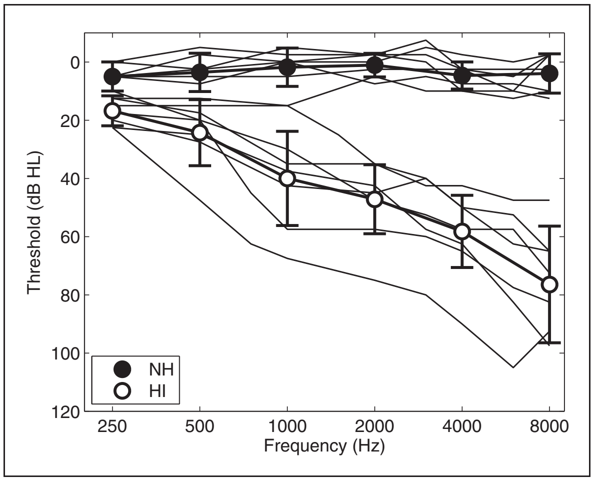

Figure 1. Audiograms (averaged over left and right ears) for each of the 14 participants (thin lines). Also shown are mean audiograms for the NH (filled circles and thick lines) and HI (open circles and thick lines) groups, with error bars showing across-subject standard deviations.
NH = normal hearing; HI = hearing impairment.

for each group. Six of the seven HI listeners were regular hearing-aid wearers, but their hearing aids were removed for the purposes of this study, and individualized linear amplification was provided to the stimuli using the National Acoustic Laboratories' revised formula with profound correction factor (NAL-RP) (Byrne, Parkinson, & Newall, 1991). All participants were college students or recent graduates, and all were native speakers of American English. They all had normal or corrected-to-normal vision. Participants were paid for their participation, gave informed consent, and all procedures were approved by the Boston University Institutional Review Board (Protocols 2633E and 3409E).

VGHA Simulation

The microphone array was designed by Sensimetrics Corporation (Malden, MA) and is functionally similar to that described in our previous publications (Kidd, 2017; Kidd et al., 2013; Kidd et al., 2015). The current version, however, uses 16 digital omnidirectional microphones instead of 8 analog cardioid microphones. The 16 microphones are arranged in four front–back-oriented rows along a flexible headband. The total length of the array is 200 mm, with a spacing of 67 mm between rows. Within each row, the microphones are arranged into two pairs with 10 mm spacing (15 mm spacing between the pairs). The outputs of the 16 microphones are weighted using the optimal-directivity algorithm of Stadler and Rabinowitz (1993) and combined to give a single-channel array output. The microphone array processing is designed to attenuate sounds arriving from directions away from the ALD (which can be set to any angle) and has an intelligibility-weighted directivity index of approximately 9 dB.

For the purposes of this experiment, two sets of impulse responses were recorded from an acoustic manikin (KEMAR) fitted with the microphone array and seated in a large sound-treated room. These impulse responses were obtained for multiple source locations in the horizontal plane (spaced at 7.5° intervals from −90° to + 90° azimuth at a distance of 5 feet) and were used to create ''virtual'' stimuli that were presented via headphones to the listener during the experiment. One set of impulse responses captured the signals received by two microphones situated in the ear canals of the manikin and were used to simulate a natural binaural listening situation (''KEMAR'' condition). The other set of impulse responses captured the 16-channel output of the microphone array for each source location. These outputs were then weighted and combined according to the ALD to give a single-channel impulse response that was used to simulate listening through the microphone array (''BEAM'' condition). A continuous range of ALDs was simulated in the range −40° to + 40° with a resolution of 2°. A hybrid configuration (''BEAMAR'') was also implemented in which low-pass filtered KEMAR impulse responses were combined with high-pass filtered BEAM impulse responses (with a crossover point at 689 Hz). The low- and high-pass filters were created by applying a Hann window to ideal frequency domain filters.

Stimuli were controlled in MATLAB (MathWorks Inc., Natick, MA) and presented via a 24-bit soundcard (RME HDSP 9632) through a pair of headphones (Sennheiser HD280 Pro). The headphone transfer function (measured on KEMAR) was compensated for prior to stimulus delivery. The participant was seated in a double-walled sound-treated booth (Industrial Acoustics Company) in front of a widescreen computer monitor (34 in. UltraWide, LG 34UM64-P). Eye position was tracked using an eye tracker (TOBII eyex) mounted on the bottom frame of the monitor. To facilitate accurate eye tracking, the head was stabilized by a chair-mounted neck rest and positioned such that the eyes were approximately 21 in. from the monitor. Calibration of the eye tracker was then done using the accompanying software at the start of each new session, and special care was taken to verify that eye gaze could be reliably tracked for participants who wore glasses or contact lenses. In the BEAM and BEAMAR conditions, the ALD of the microphone array was updated in real time according to the gaze direction of the user. Eye position was queried (and the impulse responses updated) every 98 ms, which produced a relatively smooth and continuous steering of the ALD. The total delay in the processing chain after accessing eye position (including buffering of the signal, update of the array weights, and low- and high-pass filtering) was approximately 23 ms.

Question-and-Answer Task

The stimuli for this task are 227 simple questions (e.g., “What day comes before Monday?”) and their single-word answers (e.g., “Sunday”). Each question and answer was spoken by each of 22 talkers (11 men and 11 women) and recorded by Sensimetrics Corporation (Malden, MA). The questions ranged in duration from 1,388 to 3,189 ms (mean 2,091 ms). The answers ranged in duration from 448 to 1,006 ms (mean 720 ms). A run consisted of 12 question–answer pairs (“trials”) with a gap of 0.5 s between the question and answer within a trial and also between trials (i.e., between the end of one answer and the beginning of the next question). The answer was correct on 50% of the trials and incorrect (but valid) on the remaining 50% of the trials. Listeners indicated “correct” or “incorrect” after each question–answer pair using a hand-held keypad. Trials in which a response was not registered within a 2-s window starting with the onset of the answer were excluded, but this happened rarely (on 1.9% of trials).

Note that this particular implementation of the question-and-answer task has been described in detail elsewhere (Best, Streeter et al., 2016), and a subset of the current data appeared in that paper by way of example (the KEMAR condition for the NH group).

Stimuli and Conditions

Two spatial conditions were examined, as illustrated schematically in Figure 2. In the dynamic condition (top), the location of the questions and answers moved unpredictably across three target locations (−30°, 0°, and + 30° azimuth). There was a forced transition on every question and every answer such that no utterance ever occupied the same location as that preceding it. In the fixed condition (bottom), all questions and answers were presented from one of the target locations throughout a run. The questions and answers were spoken by three randomly selected target talkers in a run. In the dynamic condition, each of the three target voices was associated with one of the three target locations.

The targets were presented simultaneously with three maskers, each of which consisted of a conversation between a different male/female pair. The maskers were located at −60°/−45°, −15°/+15°, and + 45°/+60° azimuth. They were ramped on 1 s before the first question and ramped off 1.5 s after the final answer in a run. The target stimuli were presented at 55 dB sound pressure level (as measured at the headphones for a frontal sound) and the level of each masker conversation was varied to set the target-to-maker ratio (TMR) to one of four values. These values were chosen on the basis of pilot testing separately for the NH (−10, −5, 0, and + 5 dB) and HI (−5, 0, + 5, and + 10 dB) groups.

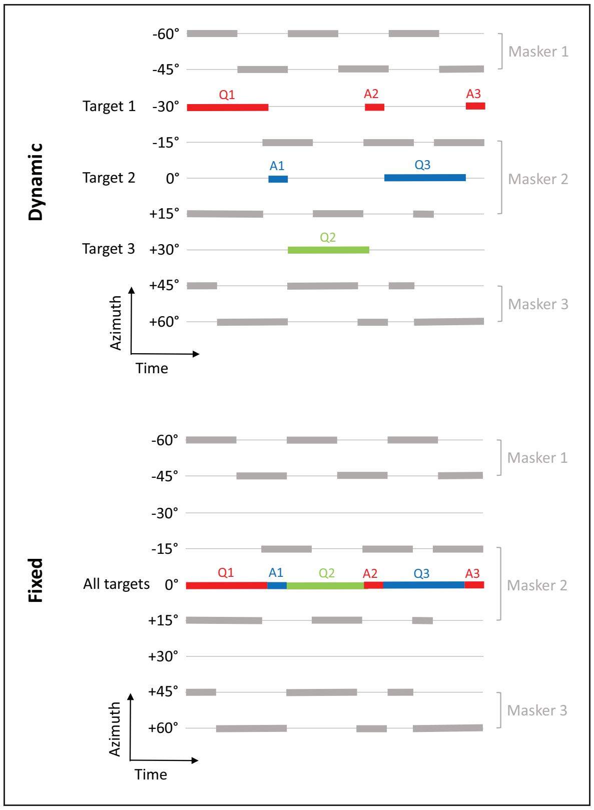

Figure 2. Example stimulus configurations in the dynamic (top) and fixed (bottom) conditions. The nine horizontal lines show the nine possible stimulus positions (from −60 to + 60° azimuth) as a function of time. Colored bars indicate the location and timing of the three target talkers (with each color representing a different talker), and the gray bars represent masker conversations. In these examples, three questions and answers (labeled Q$_1$, A$_1$, etc.) out of the sequence of 12 are shown.

A visual cue (the letter “Q” or “A”) was provided on the monitor synchronously with each question and answer (within 50 ms of the acoustic onset). The location of the cue on the monitor was calculated to correspond to the azimuth of the relevant target. This cue served both to indicate the presence of a target (so that listeners knew to respond, even at low TMRs when it may be difficult to hear the questions and answers) and to guide the eyes to the appropriate location to steer the VGHA.

Figure 3 shows directional patterns of broadband attenuation provided by the beamformer operating with ALD = −30°, 0°, or + 30°. The symbols in each panel show the attenuation of the different potential masker sources (black and gray) relative to the target source (white). The attenuation patterns were calculated

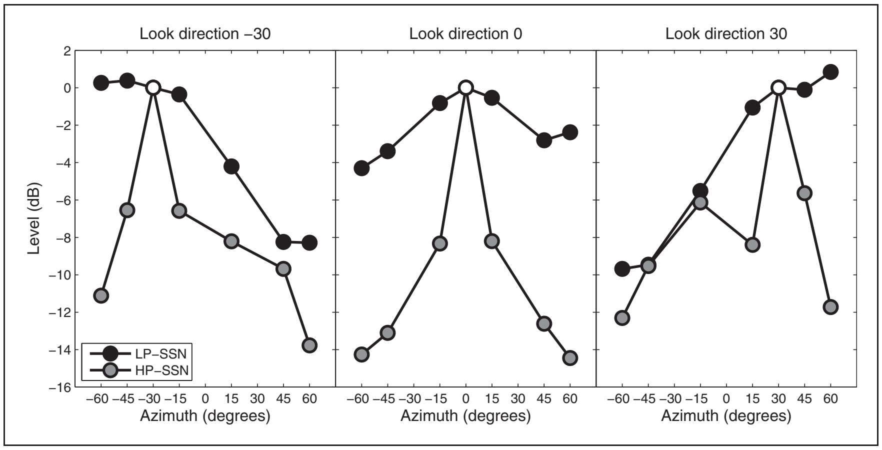

Figure 3. Broadband directional attenuation patterns for an ALD of -30° (left), 0° (middle), or $+30^\circ$ (right). To illustrate how these relate to the experimental stimuli, targets at each ALD are indicated by white symbols, and the individual potential masker sources by black and gray symbols. The attenuation patterns were calculated separately for a low- and high-passed speech-shaped noise, respectively, with a cutoff frequency of 689 Hz corresponding to that used to divide the spectrum in the BEAMAR condition.
ALD = acoustic look direction; LP-SSN = low-passed speech-shaped noise; HP-SSN = high-passed speech-shaped noise.

separately for a speech-shaped noise that was low passed at the BEAMAR cutoff of 689 Hz (black symbols) and a speech-shaped noise that was high passed at the same cutoff (gray symbols). This figure demonstrates that while the beamformer is rather broadly tuned at low frequencies, it is quite narrowly tuned at higher frequencies and can provide substantial attenuation of even the nearest maskers. To predict the gain in intelligibility that might be expected from the beamformer for the different target/masker configurations, we calculated the intelligibility-weighted SNR gain (as per Greenberg, Petersen, & Zurek, 1993) with reference to the average SNR across ears in the KEMAR condition. The predicted gains were 4.3 ± 1.6, 5.0 ± 0.4, and 3.9 ± 1.5 dB for the -30°, 0°, and 30° target locations, respectively (values represent averages and standard deviations over the eight possible masker configurations). These values were slightly lower when the reference was the better of the two ears for KEMAR (3.9 ± 1.5, 4.7 ± 0.4, and 3.5 ± 1.7 dB).

Procedures

The listeners attended four sessions of approximately 2 hr each. Across these sessions, five blocks of each of the three listening conditions (KEMAR, BEAM, and BEAMAR) were completed in a random order. The first block per listening condition was counted as training and was not included in the analysis. Within each block, the fixed condition was tested for each of the three target locations, and the random condition was tested three times to ensure that each of the three target locations was sampled as often as in the fixed condition. These six runs were tested at each of the four TMRs for a total of 24 runs per block. The order of the runs within a block was random and different for each participant. At the start of each block, listeners were informed as to which listening condition would be tested.

Results

Group Mean Performance

Figure 4 shows psychometric functions for the NH and HI groups under fixed and dynamic spatial conditions. For the KEMAR listening condition (left panel), performance was better for the NH than the HI group but within a group was very similar for the fixed and dynamic conditions. For the BEAM and BEAMAR conditions (middle and right panels), performance was also better for the NH than the HI group but, for both groups, performance for the dynamic condition was worse than for the fixed condition.

Logistic functions were fit to the raw data for each listener, and 75% thresholds were estimated. Figure 5 shows mean thresholds for each group for the fixed condition (left panel) and the dynamic condition (right panel). A mixed analysis of variance (ANOVA) found significant main effects of microphone condition, KEMAR/BEAM/BEAMAR, $F(2, 24) = 11.57$, $p < .001$; spatial condition, fixed/dynamic, $F(1, 12) = 30.84$, $p < .001$; and group, NH/HI, $F(1, 12) = 52.25$, $p < .001$, with a significant interaction between microphone condition and spatial condition, $F(2, 24) = 5.73$, $p = .009$. The factor of group did not interact significantly with microphone condition, $F\ (2, 24) = 1.82$, $p = .2$, or with spatial condition, $F(1, 12) = 0.04$, $p = .9$, and the three-way interaction was not significant, $F(2, 24) = 1.90$, $p = .2$. Planned comparisons determined whether

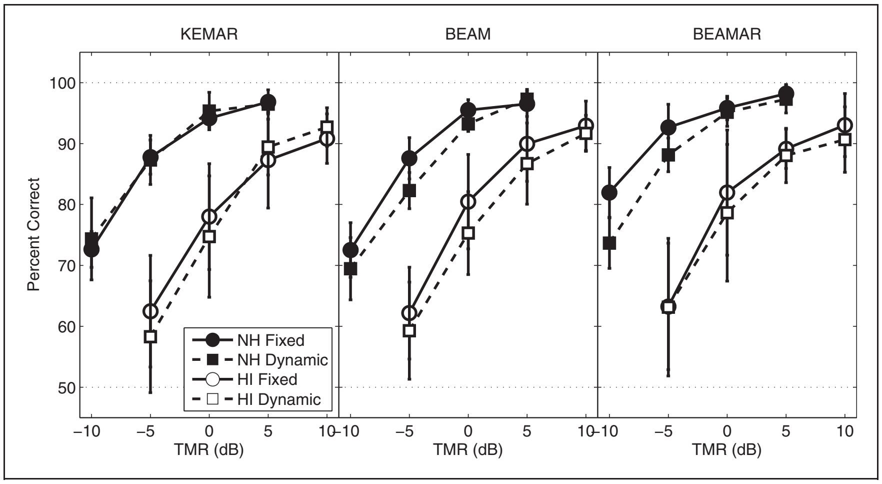

Figure 4. Average psychometric functions in the KEMAR (left), BEAM (middle), and BEAMAR (right) conditions (pooled across all target locations and all listeners in each group). Error bars show across-subject standard deviations.

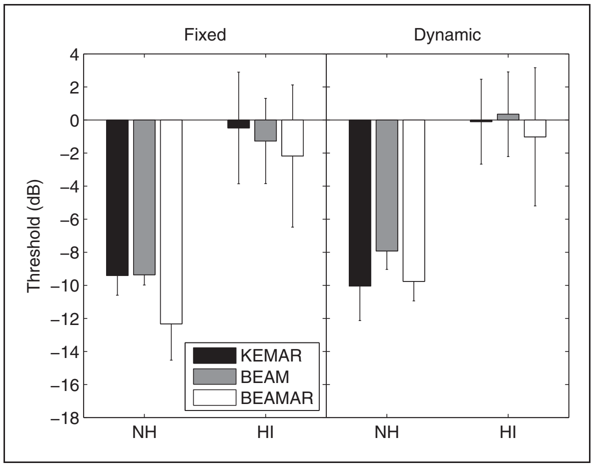

Figure 5. Mean thresholds in the fixed (left) and dynamic (right) conditions. Bars represent different microphone conditions with one cluster of bars per group. Error bars in both panels show across-subject standard deviations.

BEAM/BEAMAR performance differed from KEMAR performance (paired *t* tests using data from both groups as there were no interactions involving group). Results suggested that in the fixed condition, performance was equivalent for KEMAR and BEAM but better for BEAMAR, whereas in the dynamic condition, performance was equivalent for KEMAR and BEAMAR but poorer for BEAM ($p < .05$).

To examine the effect of target location on performance, the data for the fixed condition were subdivided according to location and then thresholds were extracted from the individual fits to the data. Note that a similar analysis could not be done for the dynamic condition, because in that case the target contained a location transition and essentially occupied two of the three locations on each trial. Also, one NH listener had to be excluded from this analysis because a threshold could not be extracted from their BEAMAR data once it was broken down by location. A mixed ANOVA with factors of microphone condition, target side (−30° or + 30°), and group found no significant effects involving target side, and thus the data were collapsed across the two sides and new thresholds were extracted for the "center" and the "sides." These thresholds, shown in Figure 6, reveal that the VGHA benefits were more pronounced for targets in the center than for targets on the sides. This observation was supported by a mixed ANOVA with factors of microphone condition, target location (center or sides), and group which found a significant main effect of microphone condition, $F(2, 22) = 12.00$, $p < .001$, target location, $F(1, 11) = 10.37$, $p = .008$, and group, $F(1, 11) = 39.88$, $p < .001$, as well as a significant interaction between microphone condition and target location, $F(2, 22) = 3.83$, $p = .037$. No other interactions were significant. Paired *t* tests using data from both groups indicated that BEAM performance was better than KEMAR performance for the center only, whereas BEAMAR performance was better than KEMAR performance for both the center and sides ($p < .05$).

Individual Performance

Substantial individual differences in performance were observed in each of the listening conditions. Figure 7 shows individual thresholds (collapsed over center and sides) for each listening condition as a function of PTA. It is clear from this figure that performance in all conditions was related to hearing loss, and this observation was supported by significant correlations ($r = .92$–$.97$,

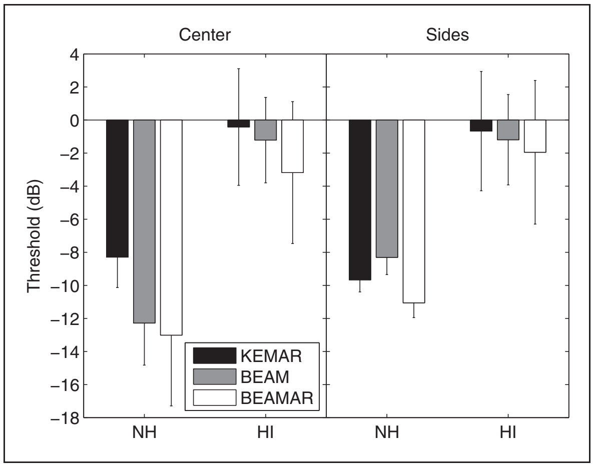

Figure 6. Mean thresholds in the fixed condition for targets located in the center (left) and to the sides (right). Bars represent different microphone conditions with one cluster of bars per group. Error bars in both panels show across-subject standard deviations.

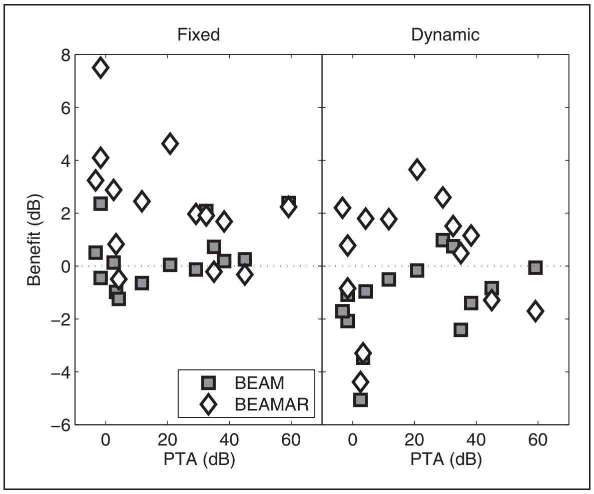

Figure 8. Individual benefits for BEAM and BEAMAR (threshold change relative to KEMAR) as a function of PTA in the fixed (left) and dynamic (right) conditions. PTA = pure-tone average.

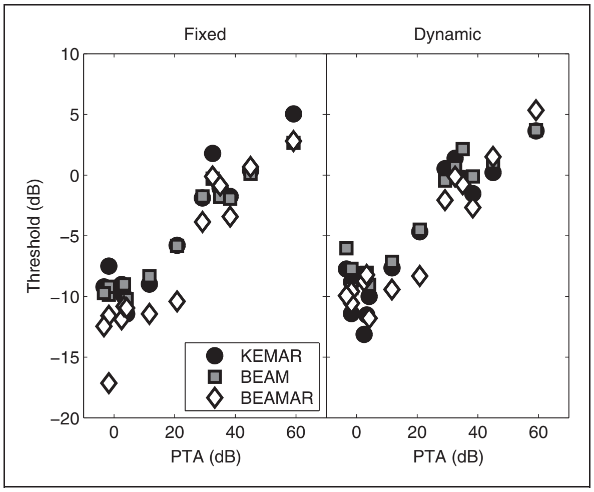

Figure 7. Individual thresholds for each microphone condition as a function of PTA in the fixed (left) and dynamic (right) conditions. PTA = pure-tone average.

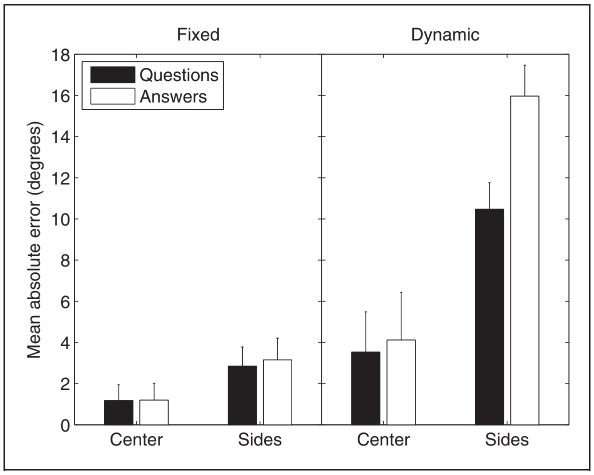

Figure 9. Mean absolute eye-gaze errors in the fixed (left) and dynamic (right) conditions, shown separately for the center and sides, and for questions and answers. Error bars in both panels show across-subject standard deviations.

$p < .001$ for all). These correlations remained significant when only the HI group was considered ($r = .77$–$.91$, $p < .04$ for all).

Figure 8 shows benefits obtained in the BEAM/ BEAMAR conditions relative to KEMAR as a function of PTA. For the fixed condition (left panel), a large range of benefits was observed, but most were positive and larger for BEAMAR than BEAM. For the dynamic condition (right panel), the benefits were reduced and more often negative. Benefits were not significantly correlated with PTA for any of the listening conditions ($r = -.42$–$.47$, $p > .09$ for all).

Eye-Gaze Errors

Figure 9 shows average eye-gaze errors in the fixed (left panel) and dynamic (right panel) listening conditions. These errors were calculated by comparing the eye-tracking data to the actual target position and calculating the mean of the absolute differences. Errors were calculated separately for the question and answer portions of each trial, and for the different target locations (center vs. sides). Because the errors did not vary as a function of TMR or microphone condition the values were collapsed across these factors for each listener. The values shown

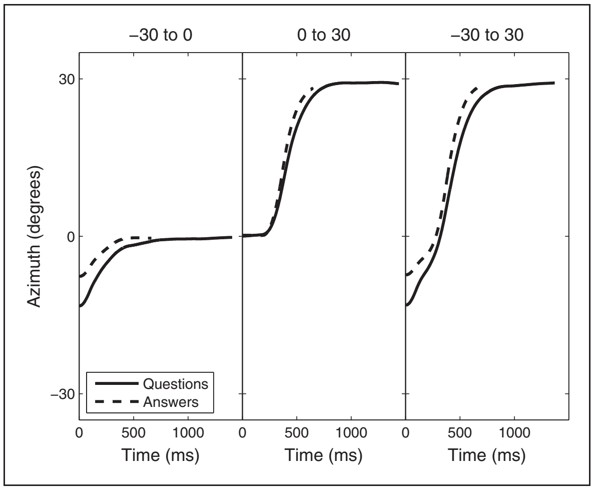

Figure 10. Average eye-traces during transitions in the dynamic condition (pooled across all transitions by all subjects). Traces are shown for three different transitions (left panel: −30° to 0°; middle panel: 0° to 30°; right panel: −30° to 30°), and the zero point in each panel indicates the onset of the word at the new location. Solid and dashed lines show cases in which subjects were transitioning to a question or an answer, respectively.

in Figure 8 represent averages (and standard deviations) across all NH and HI listeners. Eye-gaze errors were larger when the target was dynamic compared with when it stayed fixed. Moreover, eye-gaze errors in both the fixed and dynamic conditions were smaller on average for targets in the center as compared with targets to the side. Finally, on dynamic trials, errors were larger for the answer part of the trial than for the question part of the trial.

Inspection of the eye-gaze data in the dynamic condition revealed several sources of the large errors. The predominant source appeared to be delays in the eye moments. Because the errors are calculated from the onset of each question or answer, which is also the onset of the visual cue, errors for the dynamic condition include the transition of the eyes from the previous (or resting) location to the cued location. This is depicted in Figure 10, which shows average eye-traces during transitions in the dynamic condition (pooled across all transitions by all subjects). The three panels show traces for three different rightward transitions (left panel: −30° to 0°; middle panel: 0° to 30°; right panel: −30° to 30°; note that leftward transitions were similar but flipped) and time zero in each panel indicates the onset of the word at the new location. From these traces, it is clear that when the stimulus transitioned from the center to the side (middle panel), the eyes had not yet begun to move at the onset of the new word. When the stimulus transitioned from the side (left and right panels), the eyes had begun to move back towards the center by the time the new word started, presumably because the listeners knew that the next location would be in that general direction, but were still moving. In either case, the eyes did not stabilize until at least 500 ms after onset (recall that the average duration of the questions and answers was 2,091 and 720 ms, respectively). In addition, inspection of individual eye traces found that, even when eye position had plateaued in the dynamic condition, the traces tended to be less accurate and less stable than in the fixed condition.

To a first approximation, the difference in eye-gaze errors between the fixed and dynamic conditions can explain the differences in performance between these two conditions. In other words, the dynamic condition produced larger eye-gaze errors, which would have resulted in suboptimal steering of the microphone array. This may be at least part of the reason for the reduced BEAM and BEAMAR benefits found in that condition. As an interesting side note, due to a technical error, one additional HI listener completed the experiment under identical conditions but with the microphone array automatically and immediately steered to the target location (i.e., steering was decoupled from eye gaze). That listener showed a robust benefit for both BEAM and BEAMAR, which was equal in magnitude in the fixed and dynamic conditions. This anecdotal observation further supports the interpretation that the reduced benefit seen under dynamic conditions in the main experiment largely was a result of delayed or inaccurate fixation on the target location.

Discussion

Performance for Fixed Targets

Previously, we have evaluated the performance of this kind of beamformer using a fixed frontal target in the presence of symmetrically placed speech and noise maskers using a conventional speech identification test. In the current study, we used a new and different kind of speech test that is based on the comprehension of questions and answers in which the target speech was embedded in a background of competing conversations. To compare performance on this task to the performance measured previously, we examined the thresholds for the center target location in the fixed condition for NH listeners (who we expect to be somewhat homogeneous across different subject pools). Figure 5 showed that performance with the BEAM was superior to KEMAR (by 4 dB on average). This benefit was close to that predicted by the intelligibility-weighted SNR gain for these stimuli in this configuration (5 dB) but smaller than the benefits that can be obtained with two symmetrically placed noise maskers (up to 9 dB; Kidd, 2017). The benefit we observed was larger than that measured previously with four highly confusable speech maskers placed symmetrically around the target (which was often negative; Kidd et al., 2015). The superior performance of the BEAM observed here likely reflects the fact that the speech maskers used in the current study were easier to segregate from the target, which reduced the dependence on natural spatial cues.

Interestingly, the BEAM benefit was eliminated for off-center targets. This appears to be in part because KEMAR performance was better on the sides and in part because of poorer performance of the beamformer on the sides. Reductions in beamformer performance for lateral targets may be explained by differences in the attenuation patterns (see Figure 3; recall also that estimated SNR gains were slightly lower for lateral targets) as well as the increase in eye-gaze errors (see Figure 9).

Although the mean benefit observed for the frontal target was smaller in the HI group than in the NH group, their mean benefit was in fact larger for targets to the side, and overall there was no significant effect of group (and no correlation between hearing loss and benefit, see Figure 8). Across multiple studies using the BEAM condition, however, we have found several HI listeners that appear to receive less benefit than expected. The reason for this is unclear, but it may be related to reduced audibility in the high-frequency region where the beamformer provides the biggest SNR improvement. In future implementations of the VGHA, we will explore amplification strategies that provide more high-frequency gain than the NAL-RP formula, and we predict that more robust benefits will be observed in the HI population.

Consistent with our previous results (Kidd et al., 2015), the BEAMAR condition generally produced better performance than the BEAM condition in the current study for both NH and HI listeners. This suggests that listeners were able to bind the low- and high-frequency portions of the stimuli, despite the differences in spatial characteristics between the two portions, and take advantage of *both* the improved SNR from the beamformer and the spatial information from natural binaural cues (largely interaural time differences). Experiments are underway to explore what the optimal cutoff frequency is for this condition and to compare it to other ways of combining the two kinds of information.

Performance for Dynamic Targets

One appealing feature of the question-and-answer task is that the information on each trial is distributed across two parts (the question and the answer), making it possible to introduce intratrial transitions in voice and location. Moreover, these transitions are tied to the natural dynamics of the speech materials, and thus attention must be switched from talker to talker in a way that captures at least some aspects of typical conversations. The results of the KEMAR condition showed that listeners had no trouble with these dynamic variations under natural listening conditions, as performance was no worse than that for the spatially static condition. In other words, there appeared to be no ‘‘cost’’ associated with switching attention at this conversational rate. This result is consistent with another recent study, which found no impact of increasing the number of target talkers and locations on the comprehension of conversational speech (Best, Keidser, Freeston, & Buchholz, 2016). It seems that the speech materials used in these studies were sufficiently predictable that attention-switching costs were minimized and contained enough redundancy that any loss of information due to switching between streams did not jeopardize comprehension.

The primary goal of the current study was to measure the performance of the VGHA, in NH and HI listeners, in the face of spatially dynamic stimuli. Perhaps unsurprisingly, performance with the device was poorer in the dynamic condition than in the fixed condition. In this case, the listener had to use eye movements to steer the ALD of the array and was less accurate at doing so when the target location changed between each question and answer. This decrease in accuracy was undoubtedly influenced by the time required to detect the visual cue and then move the eyes accordingly. Indeed, an analysis of the eye-gaze data showed that the average time taken to initiate and complete an eye movement during a transition was around 500 ms. This means that at the start of each question and answer, the target stimulus would be outside of the beam and thus would be attenuated by the microphone array. While this may have had little impact on the highly redundant questions (e.g., it is no problem to lose the word ‘‘What’’ in ‘‘What day comes before Monday?’’), it likely would have significantly reduced the audibility of some of the single-word answers, particularly those consisting of only one syllable. Ongoing work in the lab is exploring this issue in more detail using a speech task that allows performance to be analyzed as a function of time (Roverud, Best, Mason, Streeter, & Kidd, 2016).

An open question is how well the dynamics of the speech task used in the current study represent the dynamics of real-world conversations. One relevant characteristic of our task was that the visual cue was synchronized to the onset of the target speech, with no time for preparatory eye movements. In one sense, this could be considered a ‘‘worst case scenario’’ based on the reasoning that there is often some useable predictability in real conversations about who will speak next, which would provide the user of the VGHA with the opportunity to move their eyes in advance. On the other hand, one could easily imagine instances in which transitions of talkers during conversations are unpredictable (e.g., if someone interjects suddenly). Furthermore, the visual cue in our task did provide complete certainty about where the target was located, making it a ‘‘best case scenario’’ in that respect. In real conversations, there could be occasions when a listener has no spatial cue and would have to scan through the participants in search of the current talker, which could lead to even greater losses of information than those observed here.

There are several other aspects of the experimental approach that deserve careful consideration and which may be worth exploring in future studies. First, the reference condition was a simulation based on KEMAR impulse responses, and it is possible that the results would have been different had we used individualized impulse responses to create a more accurate and compelling spatial perception. Interestingly, however, the results suggest that the KEMAR simulation used here was sufficient to support effective switching of spatial attention (as evidenced by the lack of a cost in the dynamic condition). Perhaps more importantly, in all listening conditions we used a ‘‘fixed head’’ simulation and so the results cannot be generalized easily to the more natural case in which listeners make both head and eye movements. Ultimately, we plan to evaluate a wearable version of the VGHA under free-field listening conditions in which case head movements will be possible and the reference condition will be natural binaural listening with one’s own ears.

Conclusions

A new question-and-answer task was used to evaluate the VGHA under conditions that capture some aspects of real-world communication situations. The results confirmed that the VGHA can provide a benefit for a fixed target talker for both NH and HI listeners. The VGHA benefits were reduced under the dynamic conditions tested here, in which a synchronized visual cue indicated the current location of the target talker. The results highlight some of the limits of listener-controlled beamforming in conversational settings and should be useful for guiding future investigations.

Acknowledgments

The authors like to acknowledge Lorraine Delhorne for help with subject recruitment, and Pat Zurek, Jay Desloge, and Todd Jennings for technical support and helpful discussions.

Author Note

Portions of the work were presented at the ARO MidWinter Meeting (San Diego, February 2016) and the International Hearing Aid Conference (Lake Tahoe, August 2016).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by NIH-NIDCD awards DC04545 and DC013286.

References

Adiloğlu, K., Kayser, H., Baumgärtel, R. M., Rennebeck, S., Dietz, M., & Hohmann, V. (2015). A binaural steering beamformer system for enhancing a moving speech source. *Trends in Hearing*, *19*, 1–13. doi: 10.1177/2331216515618903.

Baumgärtel, R. M., Krawczyk-Becker, M., Marquardt, D., Völker, C., Hu, H., Herzke, T., Coleman, G., . . . Dietz, M. (2015). Comparing binaural pre-processing strategies I: Instrumental evaluation. *Trends in Hearing*, *19*, 1–16. doi: 10.1177/2331216515617916.

Best, V., Keidser, G., Freeston, K., & Buchholz, J. M. (2016). A dynamic speech comprehension test for assessing real-world listening ability. *Journal of the American Academy of Audiology*, *27*, 515–526.

Best, V., Mejia, J., Freeston, K., van Hoesel, R. J., & Dillon, H. (2015). An evaluation of the performance of two binaural beamformers in complex and dynamic multitalker environments. *International Journal of Audiology*, *54*, 727–735.

Best, V., Streeter, T., Roverud, E., Mason, C. R., & Kidd, G. (2016). A flexible question-and-answer task for measuring speech understanding. *Trends in Hearing*, *20*, 1–8.

Byrne, D. J., Parkinson, A., & Newall, P. (1991). Modified hearing aid selection procedures for severe-profound hearing losses. In G. A. Studebaker, F. H. Bess, & L. B. Beck (Eds.), *The Vanderbilt hearing aid report II* (pp. 295–300). Parkton, MD: York Press.

Desloge, J. G., Rabinowitz, W. M., & Zurek, P. M. (1997). Microphone-array hearing aids with binaural output. I. Fixed-processing systems. *IEEE Transactions on Audio Speech and Language Processing*, *5*, 529–542.

Dillon, H. (2012). *Hearing aids*. Turramurra, Australia: Boomerang Press.

Doclo, S., Gannot, S., Moonen, M., & Spriet, A. (2010). Acoustic beamforming for hearing aid applications. In S. Haykin, & K. J. Ray Liu (Eds.), *Handbook on array processing and sensor networks* (pp. 269–302). Hoboken, NJ: Wiley-IEEE Press.

Greenberg, J. E., Petersen, P. M., & Zurek, P. M. (1993). Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. *Journal of the Acoustical Society of America*, *94*, 3009–3010.

Kates, J. M., & Weiss, M. R. (1996). A comparison of hearing-aid array processing techniques. *Journal of the Acoustical Society of America*, *99*, 3138–3148.

Kidd, G. (2017). Enhancing auditory selective attention using a visually guided hearing aid. Manuscript submitted for publication. *Journal of Speech, Language, and Hearing Research*.

Kidd, G., Favrot, S., Desloge, J. G., Streeter, T. M., & Mason, C. R. (2013). Design and preliminary testing of a visually guided hearing aid. *Journal of the Acoustical Society of America*, *133*, EL202–EL207.

Kidd, G., Mason, C. R., Best, V., & Swaminathan, J. (2015). Benefits of acoustic beamforming for solving the cocktail party problem. *Trends in Hearing*, *19*, 1–15.

Mejia, J., Dillon, H., Van Hoesel, R., Beach, E., Glyde, H., Yeend, I., . . . Williams, W. (2015). Loss of speech perception in noise – Causes and compensation. In S. Santurette, T. Dau, J. C. Dalsgaard, L. Tranebjærg, & T. Andersen (Eds.), *Proceedings of ISAAR 2015: Individual Hearing Loss – Characterization, Modelling, Compensation Strategies (5th Symposium on Auditory and Audiological Research)* (p. S5.3). Nyborg, Denmark: The Danavox Jubilee Foundation.

Picou, E. M., Aspell, E., & Ricketts, T. A. (2014). Potential benefits and limitations of three types of directional processing in hearing aids. *Ear and Hearing*, *35*, 339–352.

Roverud, E., Best, V., Mason, C. R., Streeter, T., & Kidd, G. (2016). *Evaluating the efficacy of a visually-guided hearing aid using a dynamic audio-visual congruence task*. Paper presented at the Mid-Winter Meeting of the Association for Research in Otolaryngology, San Diego, CA.

Saunders, G. H., & Kates, J. M. (1997). Speech intelligibility enhancement using hearing-aid array processing. *Journal of the Acoustical Society of America*, *102*, 1827–1837.

Soede, W., Berkhout, A. J., & Bilsen, F. A. (1993). Development of a directional hearing instrument based on array technology. *Journal of the Acoustical Society of America*, *94*, 785–798.

Soede, W., Bilsen, F. A., & Berkhout, A. J. (1993). Assessment of a directional microphone array for hearing-impaired listeners. *Journal of the Acoustical Society of America*, *94*, 799–808.

Stadler, R. W., & Rabinowitz, W. M. (1993). On the potential of fixed arrays for hearing aids. *Journal of the Acoustical Society of America*, *94*, 1332–1342.

Van den Bogaert, T., Doclo, S., Wouters, J., & Moonen, M. (2009). Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. *Journal of the Acoustical Society of America*, *125*, 360–371.

Völker, C., Warzybok, A., & Ernst, S. M. A. (2015). Comparing binaural pre-processing strategies III: Speech intelligibility of normal-hearing and hearing-impaired listeners. *Trends in Hearing*, *19*, 1–18. doi:10.1177/2331216515618609.