JSLHR

Research Article

Enhancing Auditory Selective Attention Using a Visually Guided Hearing Aid

Gerald Kidd Jr.[a]

Purpose: Listeners with hearing loss, as well as many listeners with clinically normal hearing, often experience great difficulty segregating talkers in a multiple-talker sound field and selectively attending to the desired "target" talker while ignoring the speech from unwanted "masker" talkers and other sources of sound. This listening situation forms the classic "cocktail party problem" described by Cherry (1953) that has received a great deal of study over the past few decades. In this article, a new approach to improving sound source segregation and enhancing auditory selective attention is described. The conceptual design, current implementation, and results obtained to date are reviewed and discussed in this article.
Method: This approach, embodied in a prototype "visually guided hearing aid" (VGHA) currently used for research, employs acoustic beamforming steered by eye gaze as a means for improving the ability of listeners to segregate and attend to one sound source in the presence of competing sound sources.
Results: The results from several studies demonstrate that listeners with normal hearing are able to use an attention-based "spatial filter" operating primarily on binaural cues to selectively attend to one source among competing spatially distributed sources. Furthermore, listeners with sensorineural hearing loss generally are less able to use this spatial filter as effectively as are listeners with normal hearing especially in conditions high in "informational masking." The VGHA enhances auditory spatial attention for speech-on-speech masking and improves signal-to-noise ratio for conditions high in "energetic masking." Visual steering of the beamformer supports the coordinated actions of vision and audition in selective attention and facilitates following sound source transitions in complex listening situations.
Conclusions: Both listeners with normal hearing and with sensorineural hearing loss may benefit from the acoustic beamforming implemented by the VGHA, especially for nearby sources in less reverberant sound fields. Moreover, guiding the beam using eye gaze can be an effective means of sound source enhancement for listening conditions where the target source changes frequently over time as often occurs during turn-taking in a conversation.
Presentation Video: http://cred.pubs.asha.org/article.aspx?articleid=2601621

This research forum contains papers from the 2016 Research Symposium at the ASHA Convention held in Philadelphia, PA.

The Analogy of Selective Attention Acting as a Spotlight Shined on the Source of Interest Under Volitional Control

Among the simpler explanations for our ability to attend to certain objects or ongoing events while ignoring other objects or events is that the conscious mind is able to select among multiple sources of sensory input and emphasize the processing of the stimulus that is selected. The act of selection and emphasis has historically suggested the analogy of a beam of light that illuminates an otherwise murky, indistinct, or cluttered visual scene. An early precursor to this view is found in the empirical work and introspections of the famous psychologist William James (1890) who observed that "the things we attend to come to us by their own laws. Attention creates no idea; an idea must already be there before we can attend to it. Attention only fixes and retains what the ordinary laws of association bring before the footlights of consciousness" (p. 450). The idea of the conscious mind figuratively shining a light on the subject of interest was a very appealing and broadly accessible formulation of the concept of attention, drawing on our common, everyday experiences of perception in a visual scene. For example, although extensive study has revealed that the analogy does not strictly hold true (e.g.,

[a]Department of Speech, Language, and Hearing Sciences and Hearing Research Center, Boston University, MA

Correspondence to Gerald Kidd, Jr.: gkidd@bu.edu

Presented at the ASHA Research Symposium, November 19, 2016, Philadelphia, PA

Editor-in-Chief: Frederick (Erick) Gallun
Editor: Karen Helfer

Received February 22, 2017
Revision received July 28, 2017
Accepted July 31, 2017
https://doi.org/10.1044/2017_JSLHR-H-17-0071

Disclosure: *The author has declared that no competing interests existed at the time of publication.*

Awh & Pashler, 2000; Driver & Bayless, 1998), selective attention has been characterized as acting like a spotlight that is voluntarily shined on whatever attracts our interest. This was stated many decades after James' work, but still early on in the modern attention literature, by Posner, Snyder, and Davidson (1980): "These findings [from a series of vision perception experiments] are consonant with the idea of attention as an internal eye or spotlight (p. 172)… that enhances the detection of events within its beam…" (p. 171). Figure 1 is a simple schematic illustration of this concept for the visual modality.

The basic idea is that directing attention toward an indistinct source brings forward the features of the visual image. The right panel shows that the focus of attention can be shifted from one source to another, bringing a different image into focus while disengaging from the former image. This selective emphasis of one source out of multiple sources distributed at different locations may be considered to be a type of "spatial filter" because it passes the information of interest within the region of focus while suppressing/attenuating the information of disregard outside the region of focus.

An Auditory Analog of the Visual Attention Spotlight

If a simple analogy for selective attention can be made by invoking a beam of light and the sense of vision, the question arises as to whether there is a parallel in the auditory domain. Can we voluntarily emphasize the input received from one sound source at a particular location while turning down the inputs from competing sound sources at other locations? And if so, how is this accomplished? A good example of the task of auditory source selection occurs when we are trying to listen to only one person speaking in a mixture of other talkers—a common experience in many typical social situations. Although distinguishing among concurrent talkers is a complex problem potentially involving a variety of cues available at different stages of processing (e.g., recent series of reviews in Middlebrooks, Simon, Popper, & Fay, 2017), the spotlight metaphor in vision may usefully be extrapolated to the sense of hearing by considering that the human listener has two receivers (ears) located on opposite sides of the head and that the sources we wish to select—among different talkers—typically are spatially distributed in the sound field. Each point of origin of a sound source creates a specific set of acoustic differences in the waveform of the sound as it is received at the two ears. These interaural values associated with different source locations—in particular, source azimuths specified in degrees from a reference point (e.g., 0° is assigned to the azimuth of a source directly in front of the listener with angles to the right and left of the center assigned positive and negative values, respectively)—provide an acoustic basis (transformed internally into a useable neural code in the binaural auditory system) for directing the "spotlight" of auditory attention toward a specific location under voluntary control. The idea is that binaural cues allow us to emphasize the processing of a sound source at one location and de-emphasize the processing of sources from other locations. Figure 2 illustrates how sound sources from different spatial locations create different patterns of arrival of sounds at the two ears. These interaural differences in the time of arrival are indicated by the different lengths of the arrows to each ear indicating that the distance the sound must travel is shorter to the near ear than the far ear. The head attenuates sounds creating acoustic "head shadow" (affecting higher frequencies more than lower frequencies) that reduces sound levels to the far ear relative to the near ear (depicted by thinner arrows for lower sound levels to the far ear in each case). Note that the two sources create distinct interaural time and level differences.

The idea of an attention-based "spatial filter" that can be tuned in azimuth and directed toward a source at a particular location is consistent with the early views of how listeners can solve the cocktail party problem. However, how is this concept realized in actual human function? And how do we know? The evidence for spatial filtering that will be considered here is drawn from a series of human perceptual and speech recognition experiments that were designed to examine the mechanisms that cause auditory masking and the ways that human listeners can overcome auditory masking. Although many studies have shown changes in performance due to varying the spatial separation between a target

Figure 1. Three-panel schematic of the "spotlight" analogy to selective attention. In the left panel (without spotlight/flashlight), a set of distributed visual images of human shapes is murky; in the center panel, the beam of light enhances the target image illuminating its features; and in the right panel, the focus of the attentional spotlight is redirected to a new source.

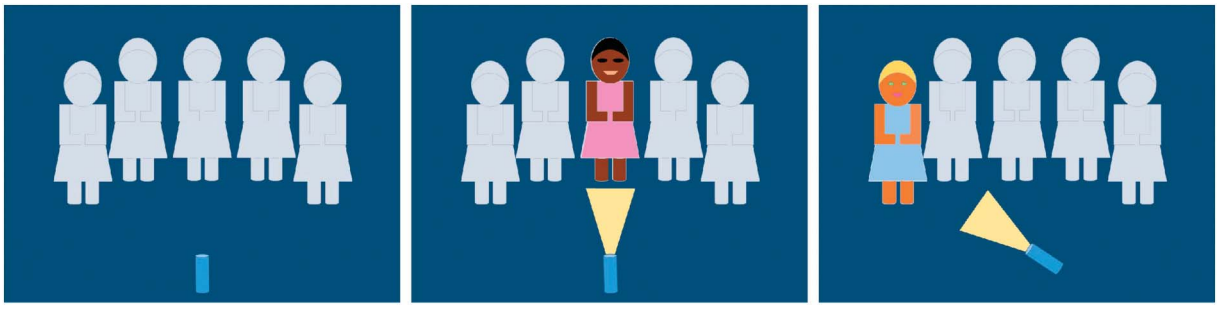
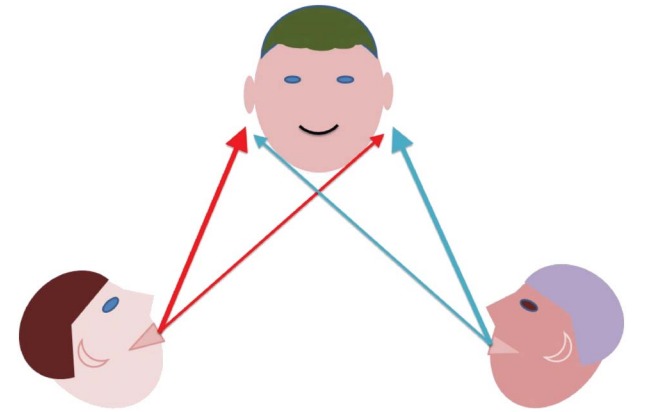


Figure 2. Schematic showing two speech sources at different locations and the interaural differences (time of arrival and sound intensity) they create at the listener.



speech source and one or more competing sources (e.g., Zurek, 1993), the direct measurement of the properties of an attention-based filter tuned in azimuth is relatively recent (and differs from the simple acoustic attenuation effects due to the head; i.e., "head shadow"). Among the first studies to address this topic directly was that of Arbogast and Kidd (2000) who adapted the probe-signal method originally applied to the frequency domain by Greenberg and Larkin (1968) to assess "tuning" in source azimuth.

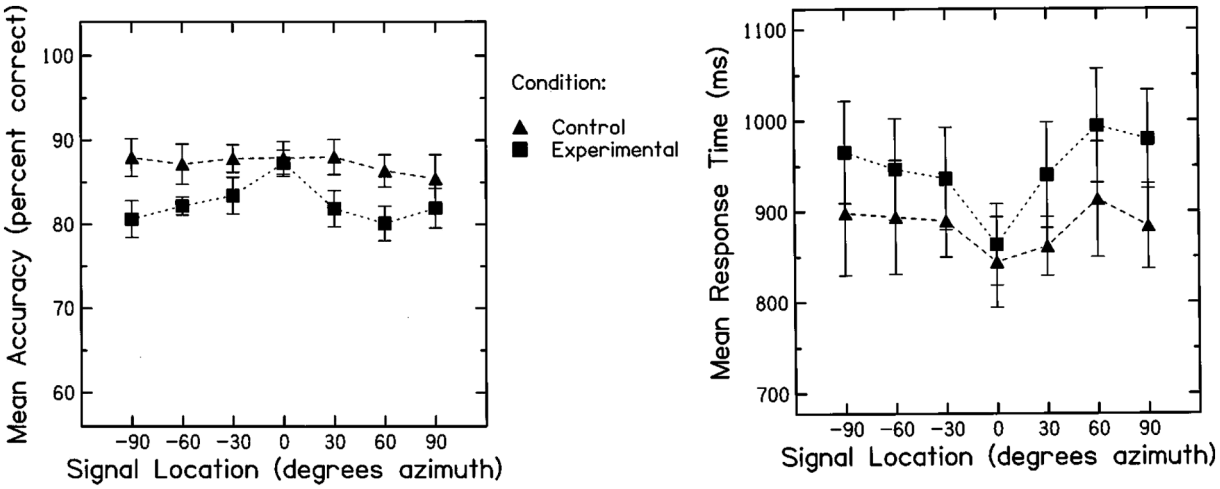
In the Arbogast and Kidd (2000) study, the task was to discriminate an upward versus a downward glide in frequency using a one-interval, two-alternative, forced-choice procedure in which the target stimulus was composed of a sequence of brief pure tones confined to a narrow and known frequency band. The target sequence was presented from a loudspeaker at one location while narrowband masking sounds comprising asynchronous randomized-frequency pure-tone sequences centered at different frequency regions (i.e., competing independent auditory "streams") were presented separately to six other loudspeakers. The seven loudspeakers were spatially distributed in 30° steps spanning a range of 180°. When the target stimulus was presented from a known location (i.e., any one of the seven loudspeakers) that was fixed across trials, masked discrimination performance was high and roughly the same for all loudspeaker locations. However, when the probe-signal procedure was employed such that the target location was uncertain (0.750 probability of occurrence from the loudspeaker directly in front of the listener and 0.042 from each of the other six locations), the most accurate performance and fastest responses were observed when the target occurred at the most likely location. Furthermore, performance gradually decreased in accuracy and increased in response time as the target location occurred increasingly distant from the most likely location. These findings are illustrated in Figure 3.

The left panel of Figure 3 shows accuracy as a function of target location in degrees azimuth. The two functions plot performance when the target was fixed in location across trials in a block and when the target was presented most often at 0° azimuth and at the other locations in a quasi-random manner on a combined 25% of presentations throughout a block of trials. The right panel shows response times for the same conditions. The "spatial tuning" attributed to an attention-based filter operating on interaural differences is supported by the findings of better accuracy and faster response times in the probe-signal condition at the expected (most likely) location. In that case, performance was equal to that which was observed in the fixed condition with reduced accuracy and longer response times (compared to the fixed condition) found for the unexpected (less likely) locations—findings that are compatible with the attenuation caused by a filter and that are comparable with the results of Greenberg and Larkin (1968) and others (cf. Scharf, Quigley, Aoki, Peachey, & Reeves, 1997; Wright & Dai, 1994) using the probe-signal method for other stimulus dimensions. These psychophysical findings are consistent with the results from a variety of other studies examining spatial tuning using different methods. For example, Teder-Salejarvi and Hillyard (1998) found evidence for sharply tuned (in azimuth) human evoked potential responses to broadband sounds that were varied probabilistically in location. Subsequent physiological work and modeling (e.g., Dong, Colburn, & Sen, 2016; Maddox, Billimoria, Perrone, Shinn-Cunningham, & Sen, 2012) have supported the idea of spatially selective neural responses at the level of the cortex notably under conditions where there are multiple sound sources present originating from distinct source azimuths.

Speech-on-Speech Masking and the Cocktail Party Problem

As noted above, one common and important use of spatial selectivity is to choose one specific talker to attend while ignoring or suppressing the speech of other talkers (e.g., Arbogast, Mason, & Kidd, 2002; Freyman, Helfer, McCall, & Clifton, 1999). Unlike the stimuli and task used in the report by Arbogast and Kidd (2000) indicating spatial tuning in the discrimination of nonspeech sounds, speech comprehension requires linguistic processing, which may be a factor for both target and masker speech (e.g., Brouwer, Van Engen, Calandruccio, & Bradlow, 2012). The idea that human listeners can use differences in spatial location to select one talker to be under the focus of attention has been discussed in the auditory and speech research literature for decades. Among the earliest and best known works advancing this idea is an article by Colin Cherry in 1953. He posed the following two questions "How do we recognize what one person is saying when others are speaking at the same time (the 'cocktail party problem')?" and "On what logical basis could one design a machine ("filter") for carrying out such an operation?" (p. 976). Although Cherry (1953) listed several cues that human listeners could use to solve the cocktail party problem, his experiments on dichotic listening (i.e., presenting an attended voice in one ear and other sounds that often were only partly ignored in the other ear via earphones) illustrated how the inputs to the two ears

Figure 3. The results from the probe-signal experiment reported by Arbogast and Kidd (2000). The left panel shows accuracy as a function of target source azimuth, whereas the right panel shows the associated response times. The two curves in each panel are for the control condition (triangles) where location is fixed throughout a block of trials and for the probe-signal condition (squares) where 0° azimuth is the most likely location. Reprinted with permission from Arbogast and Kidd. Copyright 2000, Acoustical Society of America.



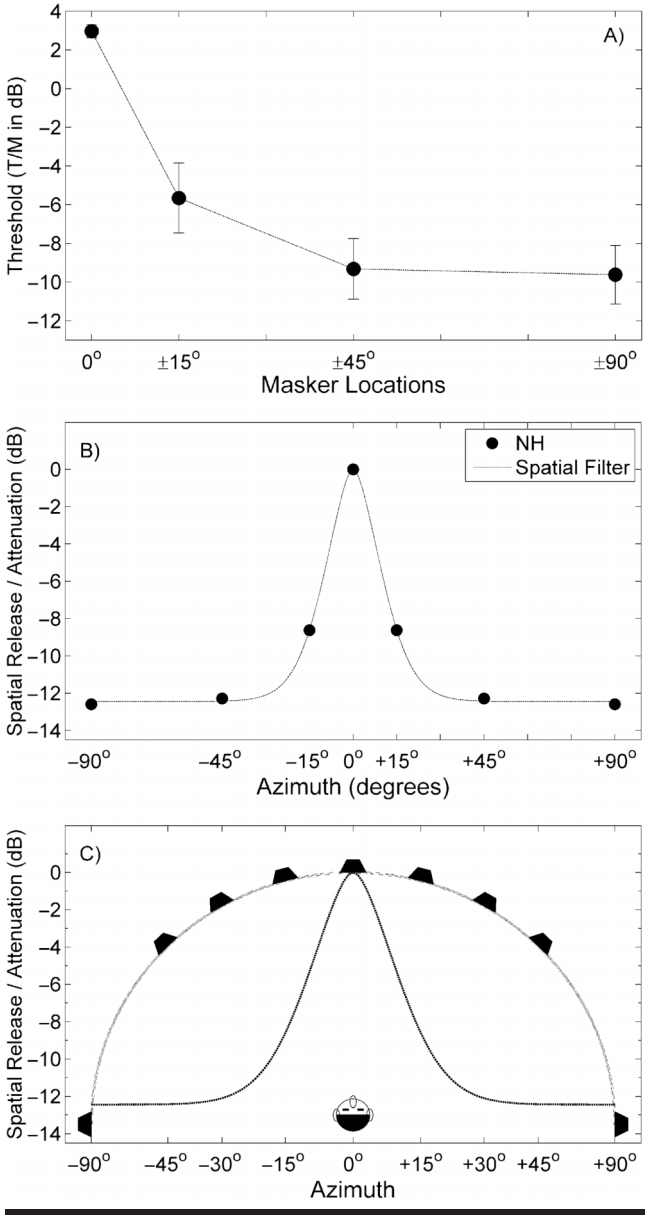
could be consciously attended and also how the strength of selective attention varied depending on the stimulus in each ear. In the dichotic listening paradigm he employed, the two ears were considered to be independent channels of the peripheral auditory system stimulated separately by earphones, but the relevance for the perception of sounds in the natural environment that produce different inputs to the two ears depending on spatial location (i.e., Figure 2) was obvious. Cherry's work motivated many subsequent studies examining how the interaural differences associated with spatially distributed sound sources can be used to improve the reception of one sound source at a specific location in the environment and to focus attention on that particular source in preference to others. As with the visual spotlight, the idea that we can aim our sense of hearing toward a source we have selected to emphasize the processing of the sound emanating from that source is both appealing and accessible based on our common auditory experience in the sound field. For the sense of hearing, the dynamic aspect of this process is of particular importance because the information we wish to extract from an acoustic message typically is organized in sequential units (e.g., words into sentences) with meaning that unfolds over time. The importance of the temporal dimension of audition requires that we maintain the focus of attention on a source over extended periods so that the items placed in memory can be fully processed and used in the exploitation of predictability that allows, in Cherry's words, "…noise or disturbances to be combatted" (p. 276). Consider, for example, the task of following the flow of conversation among two or more communication partners. To be successful, we often must redirect our attention from one talker to another as they take turns speaking. The dynamic aspects of both source selection and attentional focus have been studied less commonly in audition than have static fixed-location conditions in part because of the challenges associated with devising well-controlled and valid experiments to gauge our ability to perform such tasks.

Another famous scientist, Donald Broadbent, specifically proposed early on that we could apply an attentional filter operating on differences in spatial location to select a particular speech source of interest in the midst of competing speech sources, all of which contained potentially relevant information (i.e., comprehensible messages). His consideration of this problem led to empirical work supporting the "filter theory" of selective attention. He writes: "The main advantage of spatial separation [of speech sources] comes in the case when one or more irrelevant messages have to be ignored… [selective attention makes it] easier to pass one [message] and reject another [message] that also contains relevant items…this is effectively a filter theory" (Broadbent, 1958, pp. 42–43).

Perhaps the most straightforward approach to determining the characteristics of an attention-based spatial filter for a speech signal is to measure the difference in speech intelligibility that occurs when the target talker is masked by another voice nearby or even at the same location versus when the masking talker is moved away from the target talker causing spatial separation of sources. A priori, the expectation would be that a spatially tuned filter would emphasize sounds from the direction at which it is aimed while progressively attenuating surrounding sounds as they are increasingly separated in azimuth from the target. Does this happen when a listener is faced with solving the cocktail party problem? Figure 4 illustrates how a speech-on-speech (SOS) masking experiment can be used to infer the properties of an attention-based spatial filter.

The upper panel in Figure 4 shows the group mean results (speech reception thresholds plotted in target-to-masker ratio, T/M, in dB) from a typical multiple-talker

Figure 4. (A, upper panel) Group mean results from the speech-on-speech masking experiment of Marrone et al. (2008a) plotted as threshold target-to-masker ratios (T/M) in dB and standard errors as a function of the target-masker separation in azimuth. (B, middle panel) The data shown in A are replotted in dB attenuation/spatial release from masking with the values reflected around 0° azimuth. The dotted line connecting the data points is a best-fitting rounded exponential function illustrating the concept of a spatial filter. (C, lower panel) Spatial filter plotted as in B with the addition of a schematic overlay of the loudspeaker array and subject situated in the sound field. The distances shown in the sound-field schematic are independent of the values of attenuation/spatial release from masking on the ordinate. NH = normal hearing.



speech masking experiment as a function of the separation in azimuth between the target and masker talkers (adapted from Marrone, Mason, & Kidd, 2008a). In this case, the target talker is always located in front of the listener, and there are two independent masker talkers who are either colocated with the target at 0° azimuth or symmetrically separated in azimuth (one to the left and the other to the right) by the angles indicated along the abscissa. Spatial release from masking (SRM) is computed as the difference in target speech reception thresholds between the colocated and separated conditions. In the middle panel, the connection between SRM results and the attention-based spatial filter is shown. The data from the upper panel are replotted reflected around 0° representing the two symmetrically placed masker locations and presumed symmetric placement of the filter. The values on the ordinate are dB plotted relative to the T/M at threshold in the colocated condition (i.e., 0°, 0 dB). Thus, the filter attenuation is assumed to be equal to the SRM. In the lower panel, a schematic illustration of the laboratory listening environment is superimposed over the data from the middle panel so that the spatial filter may be visualized in reference to the listener seated in the center of a semicircle of loudspeakers. Note that the implied ordinate for the sound-field layout is distance and is independent of the filter attenuation.

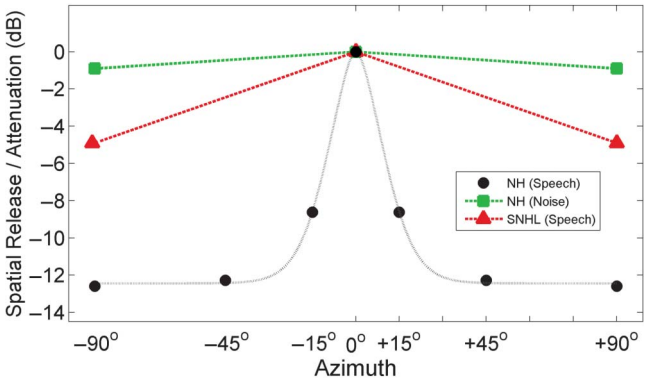
The spatial filter that is estimated using this method and the data from Marrone et al. (2008a) has a bandwidth of 10°–15° and filter slopes that attenuate about 0.6 dB/degree. An important point to be made here is that, unlike an actual filter, the characteristics of this attention-based spatial filter depend crucially on the interactions of a complex set of variables including the information about the sources that is preserved in the auditory periphery, the ability to use the peripheral inputs to segregate the different sound sources, and the higher-level functions leading to informational masking (IM) and the release from IM (see Kidd & Colburn, 2017, for a recent review).

Reduced Spatial Tuning in Listeners With Sensorineural Hearing Loss

Studies of SRM for speech targets in various types of "noise" have revealed that many listeners with sensorineural hearing loss (SNHL) are less able to exploit the spatial separation of sound sources to overcome masking than are listeners with normal hearing (NH). Furthermore, both hearing loss and age may affect the ability to use spatial separation to obtain a release from masking, especially from IM (e.g., Srinivasan, Jakien, & Gallun, 2016). This empirical observation is consistent with the common experience of those with SNHL who often report great difficulty "hearing in noise" and, in particular, understanding speech in complex sound environments such as the multiple-talker "cocktail party" situation (e.g., Agus, Akeroyd, Noble, & Bhullar, 2009; Noble & Gatehouse, 2006; see also Shinn-Cunningham & Best, 2008). In the laboratory, the reduced SRM commonly found in SNHL is apparent in SOS masking experiments when the target-masker separation in azimuth is varied. The results from one such study are shown in Figure 5.

There are three sets of data plotted here. First, the NH data from Marrone et al. (2008a) used in Figure 4 are

Figure 5. Group mean "spatial release from masking" (SRM) for listeners with sensorineural hearing loss (SNHL) plotted in dB attenuation at the spatial separation of ±90° of the maskers from the target (triangles). The thresholds used to derive the spatial filter for listeners with normal hearing (NH) for speech-on-speech masking conditions are also plotted and connected with the dotted line indicating the filter. Also shown are group mean thresholds for listeners with NH in speech envelope–modulated Gaussian noise (squares). The values are replotted from Marrone et al. (2008a, 2008b).



replotted together with the fitted filter function. Second, corresponding results from Marrone et al. (2008a) for two independent noise maskers using identical methods are shown for listeners with NH for the extreme spatial separation of ±90°. Only the extreme value was tested because of the small SRM observed (about 1 dB) and the assumption that the intermediate separations would yield the same or less benefit. The other data set is for listeners with SNHL for two speech maskers also colocated and ±90° separations from Marrone, Mason, and Kidd (2008b) again using the same methods as for the listeners with NH but with gain applied to compensate for the hearing loss. The main points from this figure are that spatial tuning, as reflected by release from masking, is strongest (greater attenuation) for listeners with NH for SOS masking conditions and that, by comparison, listeners with SNHL as a group show markedly reduced SRM. Although it is not indicated in the figure, the colocated thresholds for SOS masking are only slightly higher for the listeners with SNHL than for those with NH, meaning that the difference between groups is due primarily to the thresholds for the spatially separated masker conditions. The noise masker results for NH are shown because they produce primarily energetic masking (EM; cf. Kidd, Mason, Richards, Gallun, & Durlach, 2008) and so are less amenable to attenuation imposed by attentional mechanisms. For those maskers, the attenuation due to the attention-based spatial filter is much less than for natural speech. Also, neither acoustic head shadow nor binaural analysis—a process by which neural computations on the inputs from the two ears improve the effective within-channel signal-to-noise ratio in the auditory system (e.g., Colburn & Durlach, 1978)—provides much benefit for these symmetrically placed noise maskers. Furthermore, the colocated thresholds (T/Ms) for noise masker conditions for both NH and SNHL (not shown) are significantly lower than those for speech masking conditions reflecting the much reduced IM caused by noise (a point considered again in a later section). The interactions of these different factors, as well as others such as reduced audibility and differences in the ability to use brief glimpses of target energy in masker envelope minima (cf. Best, Mason, Swaminathan, Roverud, & Kidd, 2017), complicate the interpretation of the reduced SRM for listeners with SNHL. A full discussion of this issue is beyond the scope of the current article. Instead, the following sections provide a description of a new way of compensating for the reduced benefit of spatial separation of sources observed in most listeners with SNHL and some listeners with NH.

The Potential Benefits of Spatially Tuned Amplification

The problem addressed by the new hearing aid approach discussed in the remainder of this article is the reduced spatial benefit or lack of spatial tuning for listeners with SNHL apparent in Figure 5 by comparison of the findings from groups with NH and SNHL. As noted above, this problem manifests primarily as higher T/Ms at threshold for spatially separated sound sources. This finding leads to the question as to how auditory prostheses can remediate this problem in listeners with SNHL and restore the normal ability to segregate and select one sound source in multiple-source environments. Current hearing aids provide a number of signal-processing capabilities that may benefit listeners with SNHL in multiple-source listening situations. These capabilities include boosting sound levels to improve audibility, amplitude compression to compensate for loudness growth issues, noise reduction to reduce the levels of some types of background sounds, and directionality to emphasize the acoustic input to the listener from sources at particular azimuths relative to the listener's head.

The latter capability—directional amplification—currently appears to hold the most promise for directly compensating for the reduced spatial tuning that characteristically is observed in SNHL. Of the various approaches to providing directional amplification, acoustic beamforming is appealing because of the strong directional response and sharp spatial tuning that may be achieved in some frequency regions. The possible use of beamforming microphone arrays in hearing aids has been investigated by a number of researchers in the past, and discussions of the possible benefits of the various approaches for incorporating beamforming into auditory prostheses may be found in Greenberg and Zurek (1992); Desloge, Rabinowitz, and Zurek (1997); and Goldsworthy, Delhorne, Desloge, and Braida (2014), among others.

Recently, we have developed a new type of hearing aid—currently a research prototype—that relies on acoustic beamforming created by a head-worn microphone array. The initial design and testing of this device have been described in articles by Kidd, Favrot, Desloge, Streeter, and Mason (2013) and Favrot, Mason, Streeter, Desloge,

and Kidd (2013), and preliminary findings for SOS masking conditions for both listeners with NH and SNHL have been reported in Kidd, Mason, Best, and Swaminathan (2015); Best, Streeter, Roverud, Mason, and Kidd (2017); Best, Roverud, Mason, and Kidd (in press); and Roverud, Best, Mason, Streeter, and Kidd (in press). Figure 6 is a photograph showing the various components of the current laboratory prototype.

This figure shows the circuit board (lower right) with the four sets of four microphones used to implement beamforming circled in red. This flexible circuit board supporting the microphone array is mounted on the head underneath a band running across the top of the head from ear to ear (upper right viewed from behind placed on the KEMAR manikin). The spatial layout of the microphone array contributes to the directional response, which is obtained by computing filters (directionally dependent weights across frequency) that optimize the response to the intended "acoustic look" direction (cf. Desloge et al., 1997; Stadler & Rabinowitz, 1993). The prototype currently depends on connections to a microcomputer and various analog components and thus is neither portable nor feasible for clinical use; it is, however, sufficient in its present form for the intended purpose of evaluating the potential functional advantages of an acoustic beamformer steered by eye gaze.

The directional response of this beamformer is shown in Figure 7 plotted along with the human filter function derived from the data shown earlier from Marrone et al. (2008a).

The typical selectivity of the human attentional filter expressed in degrees azimuth at the −3 dB points is 10°–15°, and it is about 15°–20° for this beamformer at the same points. The attenuation from 0° to ±30° is about 12 dB for both. To obtain the response of the beamformer shown in Figure 7, measurements were made of the array output for an acoustic look direction (the direction the beamformer is aimed relative to the head) of 0° (straight ahead) and a broadband noise source that varied in location across several azimuths including, and to the left and right of, 0° as indicated in the figure. It should be noted, though, that the response varies as a function of frequency with broader tuning apparent in the lows and sharper tuning apparent in the highs (cf. Kidd et al., 2015). Given this composite response, the question arises as to how a human listener would benefit from listening through the beamforming microphone array to one target source located among multiple masking sources. To ascertain this, we tested groups of listeners with NH and SNHL in a speech-in-noise (speech spectrum–shaped, speech envelope–modulated noise) task that paralleled the SOS task described earlier. We used a different closed-set speech corpus (BU speech matrix corpus; Kidd, Best, & Mason, 2008) than that used in Marrone et al. (2008b), and the stimuli were presented using head-related transfer functions recorded via the KEMAR manikin in the sound field for both "natural" (microphones in KEMAR's ear canal affording normal binaural cues) and BEAM (beamforming microphone array mounted on KEMAR's head as shown in Figure 6) as described by Kidd et al. (2015). The results of these speech-in-noise masking conditions are shown in the left panel of Figure 8 with the results plotted as T/M at threshold.

The two noise maskers were generated by taking the broadband amplitude envelopes of the words spoken by two separate talkers (syntactically correct five-word utterances from the BU speech matrix corpus, same as the target sentences) and using those to modulate independent speech spectrum–shaped noises. When spatially separated using natural binaural cues, the noise sources are perceptually distinct but unintelligible and essentially are two single-channel noise vocoders. The thresholds measured in the colocated KEMAR condition are at −13 dB T/M for listeners with NH and −11 dB for listeners with SNHL. These values are much lower than those measured for the speech maskers shown in Figure 4. The difference in these thresholds is

Figure 6. Photographs of the components comprising the visually guided hearing aid mounted on the KEMAR manikin. (A) Image on the left shows the eye tracker, microphone array, insert earphones, and associated electronics; separate photos show the microphone array mounted on a headband (B) and the circuit board (C) underneath the headband that contains the array and electronics. The arrows between B and C indicate the positions of the four rows of four microphones on the circuit board.

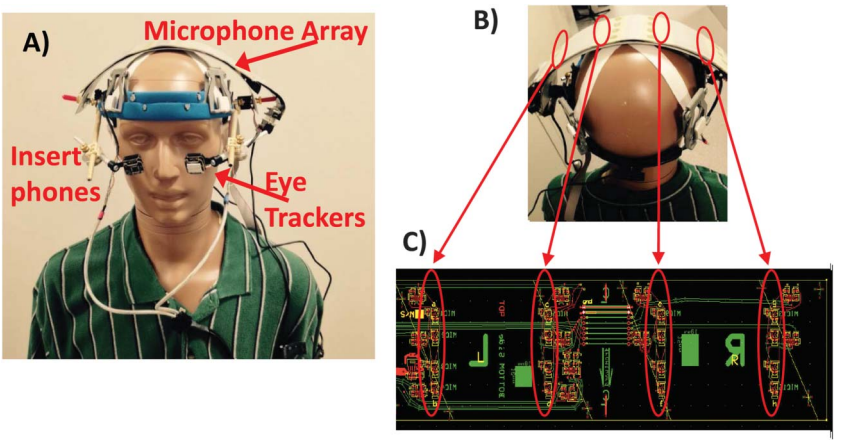
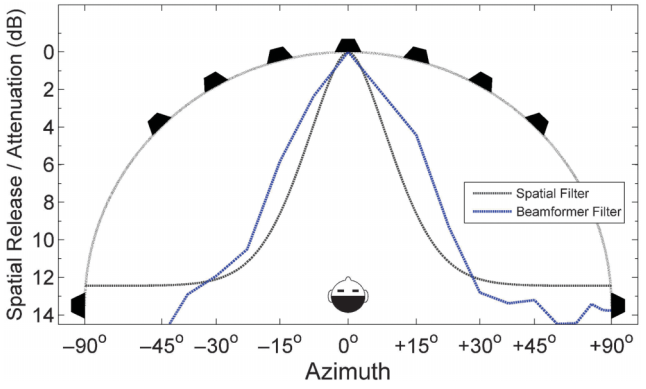


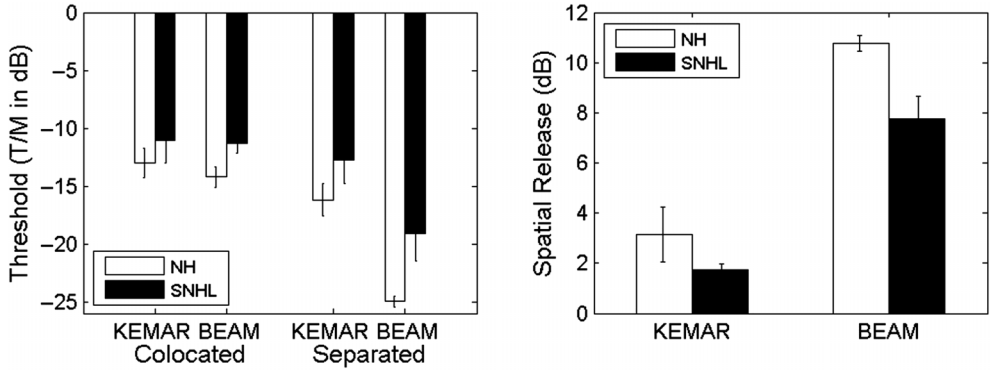
Figure 7. Measured spatial tuning characteristics of the beamforming microphone array (dashed blue line) compared with that estimated from human listeners (dotted gray line) in a speech-on-speech masking task from Marrone et al. (2008a). The schematic of the sound field with the listener and loudspeaker array also is superimposed on the data.



due to the greater IM caused by the speech masker than by the noise masker. Because there is so little IM, spatially separating the noise maskers does not produce a large source segregation advantage nor does binaural analysis or head shadow for these symmetric masker placements yield large reductions in thresholds. The thresholds for the spatially separated KEMAR conditions were −16.2 dB for listeners with NH and −12.8 dB for listeners with SNHL yielding SRMs of 3.2 and 1.8 dB, respectively. For the beamformer listening conditions, the colocated thresholds were very similar to natural listening and were −14.2 and −11.3 dB for the groups with NH and SNHL, respectively. The beamformer, however, operates before the sound input to the listener and improves the T/M for the spatially separated sources reaching the ears. Those thresholds were −25 dB for listeners with NH and −19.1 dB for listeners with SNHL, which were equal to SRMs of 10.8 and 7.8 dB, respectively. These SRMs are plotted in the right panel of Figure 8. These results show that spatially separated thresholds for both listeners with NH and SNHL in noise are significantly lower when listening through the beamformer than when using natural binaural cues. It also illustrates the important point that the "attentional filter" human listeners employ to selectively listen to one source among competing, spatially distributed sources does not operate on the sound input in the same manner as an external device with a spatially tuned response. Instead, the attention-based spatial filter operates internally using binaural input to sharpen the focus on the desired sound location. However, the information from the periphery must be present for the internal processing to achieve the benefit depicted in Figures 4 and 5 above. Another way of stating this is that the attentional filter can overcome IM but not EM. When EM dominates, binaural analysis and acoustic head shadow may be the only factors a listener is able to exploit to achieve SRM. The symmetrically placed noise maskers used here limit the benefit of these two factors. Real-world listening situations likely are a complex mixture of EM and IM that depend on many factors that may be difficult to quantify and include significant effects of context and a priori knowledge (e.g., discussion in Kidd & Colburn, 2017).

For SOS masking, the listening situation often is quite different. The information from the target talker may be present but is not utilized effectively by the listener because of IM. This difference between low-IM noise-masked and high-IM speech-masked listening conditions has been quantified using "ideal time–frequency segregation" by Brungart, Chang, Simpson, and Wang (2006; see also Kidd et al., 2016). The processing implemented by ideal time–frequency segregation essentially performs source segregation for the listener and has shown that high-IM conditions often retain sufficient target speech information to solve the SOS task but that listeners are unable to take advantage of that information. In contrast, for noise-masking conditions, the limitation on performance is due to insufficient information

Figure 8. The left panel shows group mean target-to-masker ratios (T/M) at threshold for speech in noise for colocated and spatially separated conditions for natural binaural cues (via the KEMAR manikin) and for the beamforming microphone array (BEAM). In each case, results are plotted for groups of four young adult listeners with normal hearing (NH) and sensorineural hearing loss (SNHL). In the right panel, spatial release from masking is plotted (threshold in the colocated condition subtracted from the threshold in the separated condition) for the two subject groups and two microphone conditions.

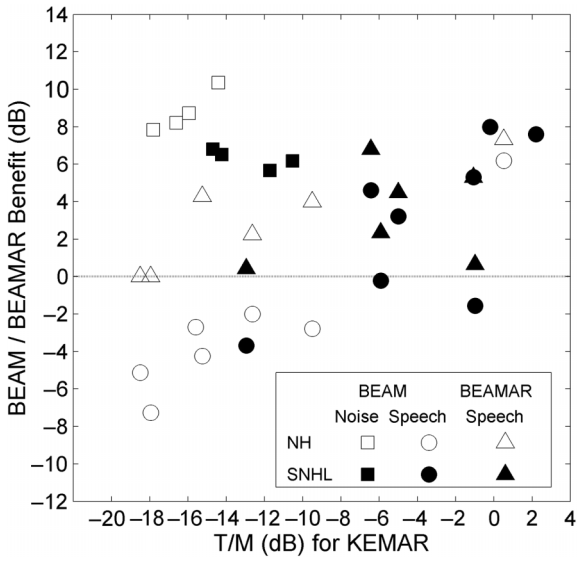


available to the listener. The important point here is that an attention-based filter operating on binaural cues can enhance speech recognition in speech maskers when the information is available to the listener and can be recovered through perceptual segregation and selection of the sources. For noise-masked speech, perceptual processing yields little benefit as compared with the physical enhancement of the stimulus available through an external device such as a beamformer that operates on the acoustic input to the listener.

Previous work comparing performance with acoustic beamforming to that with natural binaural listening for SOS masking has found large differences across subjects. This variation—when specified as a measure of "benefit" (e.g., performance with the beamformer vs. aided binaural listening)—is due primarily to the large range of performance found using natural binaural cues to solve the SOS masking task, which is typical for both groups with NH and SNHL. An example of the wide range of benefit found using the beamformer, and also the hybrid BEAMAR condition (cf. Kidd et al., 2015) that combines beamforming in the mid- to high-frequency range with natural binaural cues in the low frequencies, is shown in Figure 9.

In general, the benefit of the two beamformer conditions for speech maskers, BEAM and BEAMAR, increases as the values along the abscissa (the threshold measured for the natural binaural/KEMAR condition) increase. Positive values (those lying above the dotted horizontal line at 0 dB) indicate a benefit of the beamformer (in either BEAM or BEAMAR conditions) compared with aided binaural listening via KEMAR. The spread of the data points evident in Figure 9 reflects the large intersubject differences in performance using natural binaural listening for the spatially separated conditions and, we believe, represents a complex combination of factors including intrinsic perceptual segregation abilities as well as access to information due to hearing loss and other factors such as the ability to exploit brief time–frequency glimpses of target speech (e.g., Best, Mason, et al., 2017). Consistent with the discussion above, the microphone conditions incorporating the beamformer uniformly provide a benefit for the symmetric noise masker conditions (squares). It should be noted that the BEAMAR condition, which provides natural binaural cues in the low frequencies in combination with the signal-to-noise ratio boost from the beamformer in the high frequencies, can provide much better spatial awareness and sound source localization than the single-channel beamformer alone. Current work in our laboratory aims to determine the optimal combination of these two types of information: natural binaural cues and spatially tuned beamforming (Best, Roverud, et al., in press).

Figure 9. The benefit of listening through the beamforming microphone array (BEAM or BEAMAR) compared with natural binaural listening (through KEMAR manikin) plotted as a function of the thresholds for the natural binaural listening condition. Positive values (above the dotted horizontal line) indicate that the thresholds were better (lower) for BEAM or BEAMAR than for KEMAR conditions. The data points are for individual subjects for speech maskers (circles and triangles) and for noise maskers (squares). NH = normal hearing; SNHL = sensorineural hearing loss; T/M = target-to-masker ratios.



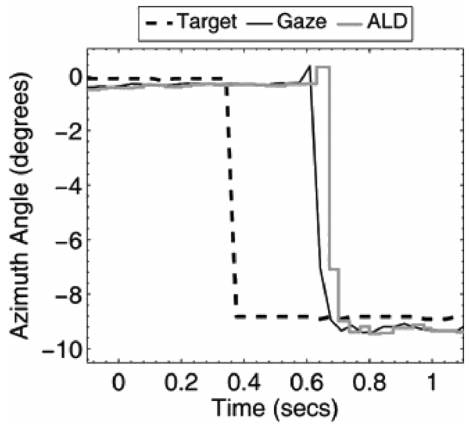
The Visually Guided Hearing Aid

To obtain the source selection benefit of spatially tuned amplification, the listener must be able to steer the acoustic look direction of the hearing aid/device toward the sound source of interest (i.e., the target). In the preceding sections describing the potential benefits of a beamforming microphone array, the listening conditions all were static: The target and masker sources were fixed in location on a given trial while the measurements were obtained. In natural listening situations, however, such as that exemplified by the multiple-talker cocktail party problem, the sound source of interest may change from moment to moment such as what occurs during turn-taking in a conversation. This dynamic variation in target source location means that the acoustic look direction of spatially tuned amplification must vary accordingly and should do so rapidly enough that the information emanating from the new target source is not lost. While human listeners typically turn their heads toward new target sound sources, they also typically locate the source visually, and gaze leads and extends the orientation achieved through head turns. To harness the source selection benefit of acoustic beamforming under dynamic listening conditions, we have developed a system in which the beam is steered by eye gaze sensed by a head-worn, eyeglasses-mounted eye tracker. The prototype version of this "visually guided hearing aid" (VGHA) including eye tracker, microphone array, and audio output/electronic transducer is shown in the left image of Figure 6. Not shown is the computer that implements the processing involved in computing the directional filters that are selected based on the signals from the eye tracker indicating the direction of eye gaze. The idea is that these functions—sensing eye gaze and steering the acoustic look direction of the beamformer—could ultimately be implemented in a portable self-contained instrument worn

by a human listener, although the versions tested so far only have been laboratory prototypes used for research purposes. In our initial efforts in developing this system, a first practical question was whether the speed of operation including all of the steps mentioned above could be implemented in real time. An example of the performance measured for the VGHA is shown in Figure 10.

These measurements indicate that the various steps in guiding the beamformer using eye gaze result in a lag of about 30 ms relative to the actual movement of the eyes in response to a changing visual target. Although this delay may be noticeable, further work from our laboratory has indicated that it does not interfere with speech intelligibility. However, the general problem of determining how well listeners are able to track sound source transitions in dynamic listening situations, such as following a conversation in a noisy room filled with competing conversations, is a challenging task that has not been studied extensively. Recently, we have developed tests for this purpose and have been devising experiments that compare listener performance using the VGHA with that which may be obtained naturally by shifting the focus of attention from one source to another source at different spatial separations. Because the VGHA is steered by eye gaze and because it is often the case that selective attention to a source at a specific location is optimized by the concerted orienting of vision and audition toward the target source, the approach we have taken is to use visual markers to indicate the location and the transition of location for the target. What appears to be crucial in the design of these experiments is the complexity of the task—the extent to which solving the task requires the integration of auditory and visual information, for example—and the a priori knowledge that allows the observer to anticipate source transitions. One approach to this problem has been reported by Roverud et al. (in press) who have tested the VGHA in an "auditory–visual congruence" task. In their paradigm, three separate talkers utter strings of synchronously spoken independent words from three separate locations (i.e., spatially separated loudspeakers implemented virtually using head-related transfer functions). A monitor displays the layout of the loudspeaker array, and a single synchronous string of printed words occurs at one of the loudspeaker locations. The task of the listener is to press a button when the auditory word and the visual word match from the same location, that is, an A–V congruence. The printed words can change location at the onset of any of the words so that a transition in target location occurs, and the probability of that transition is a controlled variable. The preliminary results from this experiment indicate that the benefit of the beamformer is retained under dynamic conditions, although the time it takes to orient eye gaze toward the target may influence performance immediately after a transition. A similar conclusion was reached by Best, Streeter, et al. (2017) who compared natural and VGHA performance in a new dynamic speech comprehension test. In this new test (Best, Streeter, Roverud, Mason, & Kidd, 2016), each trial consists of a simple question (e.g., "What is 2 + 3?") and a one-word answer (e.g., correct "5" or incorrect "8"). Because the trial consists of two separate parts—the question and the answer—it is well suited for implementing target source transitions because the natural conversational form involves two separate talkers at different locations. The test involves comprehension, not only identification or recognition, because the yes/no response of the listener indicates whether the answer was correct or incorrect. Many aspects of this very flexible task may be varied to assess the understanding of speech under dynamic conditions. As in the Roverud et al. (in press) study, Best, Streeter, et al. (2017) found that performance using the VGHA retained the spatial selectivity benefit of the acoustic beamformer and that listeners were able to track target source transitions with minimal loss in performance. Both of these new approaches appear to hold promise for examining dynamic aspects of speech or A–V target recognition/comprehension and for evaluating the performance of spatially tuned amplification devices.

Figure 10. The traces plotted here show the change in location/azimuth of a visual target (dot on a screen) over time (heavy dashed line), the movement of a human subject's eye gaze following the target dot (dark solid line), and the computation and application of the directional filter that implements beamforming directed toward the visual target (the "acoustic look direction" [ALD], light solid line). Reprinted with permission from Kidd et al. (2013).



Future Directions

The current version of the VGHA provides the benefits of acoustic beamforming guided by eye gaze. The system is a research prototype involving several functional subcomponents that may be tested separately or together as a unit. The benefits of acoustic beamforming for source selection augment our natural ability to focus attention on one specific source in the environment while effectively attenuating competing sources at other locations. The extent to which our natural ability to implement selective attention based on interaural difference cues—that is, the extent to which it functions like a spatial filter—depends on various factors including the type of competition the listener experiences in the environment (i.e., whether EM or IM dominates).

The beamforming microphone array described here was designed to fit on the head of a human user and, in its current implementation, extends across the top of the head using a flexible band with the rows of microphones that are used to create the beam of amplification. This configuration generally works best in multiple-source sound fields when the sources are nearby the listener and reverberation is low (e.g., Favrot et al., 2013). Moreover, the type of sensing that is used to determine eye gaze usually has constraints that can limit the usefulness of the device. The sensor used currently does not work well in natural light that has a strong infrared component (i.e., works well indoors but not outdoors). Other ways of sensing eye gaze, such as through electrooculography, or even sensing the direction of attention through the electroencephalography signal may have useful applications but also may have their own constraints (e.g., slow response due to the need for signal averaging).

Highly selective, spatially tuned amplification may have applications other than those considered here. For example, it is clear from past work (e.g., Figure 9) that some listeners, even the young adult college students with NH who comprise our typical subject pool, have great difficulty hearing out one talker among multiple competing talkers. Furthermore, cochlear implant users who experience reduced benefit from spatial cues or persons with attention-related functional deficits could benefit from this type of highly directional amplification. This is because spatially tuned amplification enhances source selection at the acoustic input to the listener and there may be various listening conditions where such amplification is beneficial even when hearing sensitivity is within normal limits. Furthermore, the hybrid beamformer/natural binaural listening condition ("BEAMAR") described in an earlier section holds considerable promise for retaining the benefits of spatially selective amplification while also allowing some degree of natural sound source awareness and localization ability (e.g., Best, Roverud, et al., in press; Kidd et al., 2015).

Acknowledgments

The author acknowledges NIH/NIDCD Grant Awards DC013286 and DC004545 and AFOSR Award FA9550-16-1-0372 for supporting portions of the work described here. The author is grateful to Christine R. Mason for her contributions to this work and to the preparation of this article. He also is grateful to his other colleagues for their collaborations on much of the research described here. The Research Symposium is supported by the National Institute On Deafness and Other Communication Disorders of the National Institutes of Health under Award Number R13DC003383. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Agus, T. L., Akeroyd, M. A., Noble, W., & Bhullar, N. (2009). An analysis of the masking of speech by competing speech using self-report data (L). *The Journal of the Acoustical Society of America, 125,* 23–26.

Arbogast, T. L., & Kidd, G., Jr. (2000). Evidence for spatial tuning in informational masking using the probe-signal method. *The Journal of the Acoustical Society of America, 108,* 1803–1810.

Arbogast, T. L., Mason, C. R., & Kidd, G., Jr. (2002). The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America, 112,* 2086–2098.

Awh, E., & Pashler, H. (2000). Evidence for split attentional foci. *The Journal of Experimental Psychology: Human Perception and Performance, 26,* 834–846.

Best, V., Mason, C. R., Swaminathan, J., Roverud, E., & Kidd, G., Jr. (2017). Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures. *The Journal of the Acoustical Society of America, 141,* 81–91.

Best, V., Roverud, E., Mason, C. R., & Kidd, G., Jr. (in press). Performance of a hybrid beamformer that preserves auditory spatial cues. *Journal of the Acoustical Society of America.*

Best, V., Streeter, T., Roverud, E., Mason, C. R., & Kidd, G., Jr. (2016). A flexible question-answer task for measuring speech understanding. *Trends in Hearing, 20,* 1–8. https://doi.org/10.1177/2331216516678706

Best, V., Streeter, T., Roverud, E., Mason, C. R., & Kidd, G., Jr. (2017). The benefit of a visually guided beamformer in a dynamic speech task. *Trends in Hearing, 21,* 1–11. https://doi.org/10.1177/2331216517722304

Broadbent, D. E. (1958). *Perception and communication*. Oxford, England: Pergamon Press.

Brouwer, S., Van Engen, K., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America, 131,* 1449–1464.

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America, 120,* 4007–4018.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and two ears. *The Journal of the Acoustical Society of America, 25,* 975–979.

Colburn, H. S., & Durlach, N. I. (1978). Models of binaural interaction. In E. Carterette & M. Friedman (Eds.), *Handbook of perception: Hearing* (Vol. 4, Chap. 11). New York, NY: Academic Press.

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

Dong, J., Colburn, H. S., & Sen, K. (2016). Cortical transformation of spatial processing for solving the cocktail party problem: A computational model. *eNeuro, 3*(1), 1–11. https://doi.org/10.1523/ENEURO.0086-0015.2015

Driver, J., & Bayless, G. C. (1998). Movement and visual attention: The spotlight metaphor breaks down. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 448–456.

Favrot, S., Mason, C. R., Streeter, T., Desloge, J., & Kidd, G., Jr. (2013). Performance of a highly directional microphone array in a reverberant environment. *Proceedings of the International Conference on Acoustics/Acoustical Society of America, Montreal, Canada, 18,* 8.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America, 106,* 3578–3588.

Goldsworthy, R. L., Delhorne, L. A., Desloge, J. G., & Braida, L. D. (2014). Two-microphone spatial filtering provides speech reception benefits for cochlear implant users in difficult acoustic environments. *The Journal of the Acoustical Society of America, 136,* 867–876.

Greenberg, G. Z., & Larkin, W. D. (1968). Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method. *The Journal of the Acoustical Society of America, 44,* 1513–1523.

Greenberg, J. E., & Zurek, P. M. (1992). Evaluation of an adaptive beamforming method for hearing aids. *The Journal of the Acoustical Society of America, 91,* 1662–1676.

James, W. (1890). *The principles of psychology*. New York, NY: Holt and Co.

Kidd, G., Jr., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America, 124,* 3793–3802.

Kidd, G., Jr., & Colburn, H. S. (2017). Informational masking in speech recognition. In J. C. Middlebrooks, J. Z. Simon, A.N. Popper, & R. R. Fay (Eds.), *The auditory system at the cocktail party* (pp. 75–109). New York, NY: Springer Nature.

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

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

Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational masking. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 143–190). New York, NY: Springer Science + Business Media, LLC.

Kidd, G., Jr., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., & Best, V. (2016). Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America, 140,* 132–144.

Maddox, R. K., Billimoria, C. P., Perrone, B. P., Shinn-Cunningham, B. G., & Sen, K. (2012). Competing sound sources reveal spatial effects in cortical processing. *PLoS Biology, 10,* e1001319.

Marrone, N. L., Mason, C. R., & Kidd, G., Jr. (2008a). Tuning in the spatial dimension: Evidence from a masked speech identification task. *The Journal of the Acoustical Society of America, 124,* 1146–1158.

Marrone, N. L., Mason, C. R., & Kidd, G., Jr. (2008b). Effect of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *The Journal of the Acoustical Society of America, 124,* 3064–3075.

Middlebrooks, J. C., Simon, J. Z., Popper, A. N., & Fay, R. R. (Eds.). (2017). *The auditory system at the cocktail party, Springer handbook of auditory research, 60*. New York, NY: Springer. https://doi.org/10.1007/978-3-319-51662-2_4

Noble, W., & Gatehouse, S. (2006). Effects of bilateral vs. unilateral hearing aid fitting on abilities measured on the Speech, Spatial, and Qualities of Hearing Scale (SSQ). *International Journal of Audiology, 45,* 172–181.

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *The Journal of Experimental Psychology: General, 109,* 160–174.

Roverud, E., Best, V., Mason, C. R., Streeter, T., & Kidd, G., Jr. (in press). Evaluating the performance of a visually guided hearing aid using a dynamic audio-visual word congruence task. *Ear and Hearing*.

Scharf, B., Quigley, S., Aoki, C., Peachey, N., & Reeves, A. (1987). Focused auditory attention and frequency selectivity. *Perception & Psychophysics, 42,* 215–223.

Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification, 12,* 283–299.

Srinivasan, N. K., Jakien, K. N., & Gallun, F. J. (2016). Release from masking for small spatial separations: Effects of age and hearing loss. *The Journal of the Acoustical Society of America, 140,* EL73–EL78.

Stadler, R. W., & Rabinowitz, W. M. (1993). On the potential of fixed arrays for hearing aids. *The Journal of the Acoustical Society of America, 80,* 1332–1342.

Teder-Salejarvi, W. A., & Hillyard, S. A. (1998). The gradient of spatial auditory attention in free field: An event-related potential study. *Perception & Psychophysics, 60,* 1228–1242.

Wright, B. A., & Dai, H. (1994). Detection of unexpected tones in gated and continuous maskers. *The Journal of the Acoustical Society of America, 95,* 939–948.

Zurek, P. M. (1993). Binaural advantages and directional effects in speech intelligibility. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance* (pp. 255–276). Boston, MA: Allyn and Bacon.