Examination of a hybrid beamformer that preserves auditory spatial cues

Virginia Best, Elin Roverud, Christine R. Mason, and Gerald Kidd, Jr.
Department of Speech, Language and Hearing Sciences, Boston University, Boston, Massachusetts 02215, USA
ginbest@bu.edu, erover@bu.edu, cmason@bu.edu, gkidd@bu.edu

Abstract: A hearing-aid strategy that combines a beamforming microphone array in the high frequencies with natural binaural signals in the low frequencies was examined. This strategy attempts to balance the benefits of beamforming (improved signal-to-noise ratio) with the benefits of binaural listening (spatial awareness and location-based segregation). The crossover frequency was varied from 200 to 1200 Hz, and performance was compared to full-spectrum binaural and beamformer conditions. Speech intelligibility in the presence of noise or competing speech was measured in listeners with and without hearing loss. Results showed that the optimal crossover frequency depended on the listener and the nature of the interference.

[QJF]
Date Received: August 3, 2017 **Date Accepted:** September 30, 2017

1. Introduction

Beamforming algorithms, which combine the signals from multiple microphones to create directional tuning, are of great interest to hearing-aid researchers and manufacturers for their potential to improve speech understanding in noise (e.g., Soede *et al.*, 1993a; Stadler and Rabinowitz, 1993; Kates and Weiss, 1996; Desloge *et al.*, 1997; Saunders and Kates, 1997; Doclo *et al.*, 2010; Baumgärtel *et al.*, 2015; Völker *et al.*, 2015). One issue with many beamforming algorithms, however, is that their output is a single-channel signal which no longer contains any binaural information. This is a problem not only because binaural information allows listeners to locate sounds and orient in their environment, but also because differences in perceived location that binaural cues provide can help listeners to segregate competing sounds (e.g., Freyman *et al.*, 2001). Thus, 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 might be partially counteracted by the loss of spatial cues (Kidd *et al.*, 2015; Best *et al.*, 2017).

Several potential strategies for preserving or restoring spatial information have been proposed (e.g., Desloge *et al.*, 1997; Van den Bogaert *et al.*, 2009; Doclo *et al.*, 2010; Picou *et al.*, 2014). In one such strategy, which we explored in the current study, the beamforming is restricted to higher frequencies and the lower frequencies are reserved for the reception of natural speech that includes binaural information. The strength of this approach lies in the fact that beamforming is most effective at higher frequencies, where the wavelengths are short, and thus the consequences of restricting the bandwidth may be minor in functional terms. Conversely, it is well known that interaural time differences provide a robust cue to localization for low-frequency signals. Thus, this combination of low- and high-frequency information could retain much of the acoustic benefit of a full beamformer while preserving some spatial awareness. This prediction depends on the effective integration of the single-channel (diotic) beamformer input with the natural two-channel (binaural) input. Our previous work related to this topic has shown that a hybrid beamformer based on this approach leads to lower speech reception thresholds than a full beamformer for challenging speech-on-speech masking tasks (Kidd *et al.*, 2015; Best *et al.*, 2017). To date, however, there has been no attempt to determine how performance depends on the boundary between the high- and low-frequency regions that the listener must integrate to effectively utilize this hybrid approach, nor has an optimal crossover frequency been determined. Furthermore, recent work has indicated that the benefits of beamforming depend on the type of masking (energetic or informational, cf. Kidd, 2017) that is present in the acoustic environment, with more robust benefits observed when energetic masking dominates. The goal of the current study was to examine the effect of varying the crossover frequency of a hybrid beamformer on speech intelligibility in the presence of

either noise or speech maskers intended to produce primarily energetic or informational masking, respectively.

2. Methods

2.1 Participants

The participants in the study were fourteen young adults, seven with normal hearing (NH; mean age 27 years), and seven with bilateral sensorineural hearing impairment (HI; mean age 28 years). The NH listeners had pure-tone averages (PTA; mean threshold across both ears at 0.5, 1, and 2 kHz) that ranged from 0 to 5 dB hearing level (HL) (mean 2 dB HL). The HI listeners had a range of losses with PTAs from 19 to 74 dB HL (mean 47 dB HL). The losses were relatively symmetric, with a PTA difference between the ears of no more than 10 dB. Five of the seven HI participants were regular hearing aid users (four bilateral, one unilateral), while the remaining two had never worn hearing aids. No participant had any experience outside of the laboratory using a beamforming hearing aid. 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).

2.2 Microphone array

The microphone array was created by Sensimetrics Corporation (Malden, MA). It is 200 mm long, and consists of 16 digital omnidirectional microphones arranged in four front-back oriented rows along a flexible headband. 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. Figure 1 shows patterns of attenuation provided by the beamformer for an acoustic look direction of 0° and a range of source directions from −90° to +90°. The different lines show the attenuation pattern for octave-wide bands of noise centered at frequencies from 125 to 4000 Hz. The attenuation patterns show that the spatial selectivity of the microphone array increases systematically with frequency.

The hybrid beamformer was realized by means of a "virtual" headphone simulation. Briefly, 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 (at a distance of 5 ft). One set of impulse responses captured the signals received

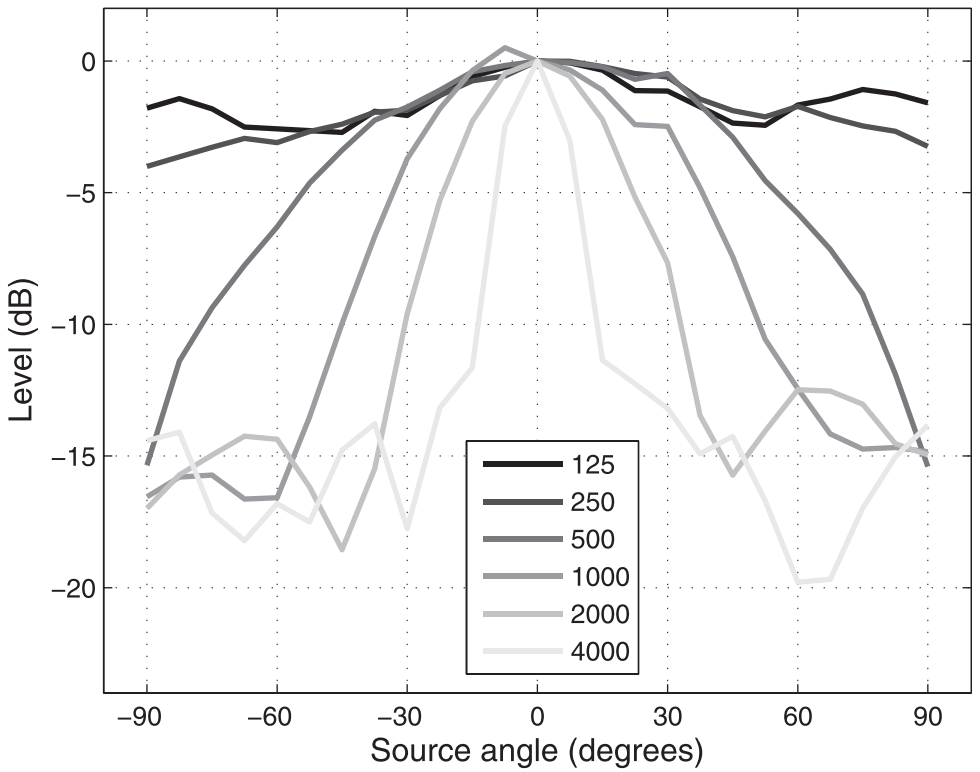

Fig. 1. Directional attenuation patterns for an acoustic look direction of 0° and a range of octave-band noises centered at frequencies from 125 to 4000 Hz.

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, which were weighted and combined to give a single-channel beamformer output ("BEAM" condition). The hybrid configuration ("BEAMAR") was created by combining low-pass filtered KEMAR impulse responses with high-pass filtered BEAM impulse responses, with a crossover point that was varied parametrically from 200 to 1200 Hz in 200 Hz steps. The low- and high-pass filters were created by applying a Hann window to ideal frequency-domain filters.

2.3 Stimuli

Target stimuli were five-word sentences of the form name-verb-number-adjective-noun (e.g., "Sue found two red shoes"), which were created by concatenating individually recorded monosyllabic words. These words were taken from a 40-word corpus containing eight words in each of the five word categories (Kidd *et al.*, 2008). The target sentence was spoken by a female voice (chosen randomly on each trial from a set of eight) and was identified on the basis of the first keyword (which was always "Sue"). The target was presented simultaneously with four maskers, which were comprised of either speech or noise. The speech maskers were competing sentences, assembled in the same manner as the target sentence, but using different female talkers and different keywords. In this condition, all five sentences were time-aligned at the beginning of the first word, but tended naturally to become staggered by the end of the utterances. The noise maskers were speech-shaped speech-modulated noises. They were generated by modulating a random speech-shaped noise (having the same spectrum as the average of all of the female words in the corpus) with the envelope of a random speech masker (extracted using a fourth order 300-Hz low-pass Butterworth filter).

The target was always located at 0° azimuth, and the four maskers were located at −60°, −30°, +30°, and +60° azimuth. Each masker was presented at 55 dB sound pressure level (SPL) and the level of the target was varied to set the target-to-maker ratio (TMR) to −20, −15, −10, −5, or 0 dB. For HI listeners, individualized linear amplification according to the NAL-RP prescription (Byrne *et al.*, 1991) was applied to each stimulus just prior to presentation.

Informal listening in six of the 14 participants (four NH and two HI) confirmed that for the BEAM condition, all sounds were perceived at the center, and that for the KEMAR condition, the lateral maskers were perceived at or near their intended azimuth. Perceived laterality of the maskers was intermediate for the BEAMAR conditions, and varied systematically with the crossover frequency, which determined the ratio of binaural to diotic information in the mixed signal.

2.4 Procedures

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 listener was seated in a double-walled sound-treated booth (Industrial Acoustics Company) in front of a computer monitor and provided responses by clicking on a custom-made user interface containing a grid of the 40 words.

The listeners attended two sessions of approximately two hours each. In these sessions they completed three blocks of each of the 16 conditions (two masker conditions × eight listening conditions) for a total of 48 blocks. The blocks were presented in a random order for each listener. Within each block, a single condition was tested five times at each of the five TMRs (25 trials). Psychometric functions were generated for each listener in each condition by plotting percent correct (calculated across all four keywords in all trials) as a function of TMR, and fitting a logistic function. Thresholds corresponding to the TMR at 50% correct were extracted from each function.

3. Results

The upper panels of Fig. 2 show thresholds as a function of crossover frequency for the speech-masker condition (A) and the noise-masker condition (B). Each line in these panels represents a different listener, with the NH listeners shown in black and HI listeners shown in gray. The dashed grey lines in panel (A) indicate two listeners whose performance was so poor in the KEMAR condition that a threshold could not be extracted from their data; the asterisk indicates our estimate of those thresholds based on extrapolation. The lower panels of Fig. 2 show the same data as in panels (A) and (B), but normalized by each individual's KEMAR threshold such that the ordinate represents the "benefit" provided by the beamformer relative to the KEMAR condition. Figure 3

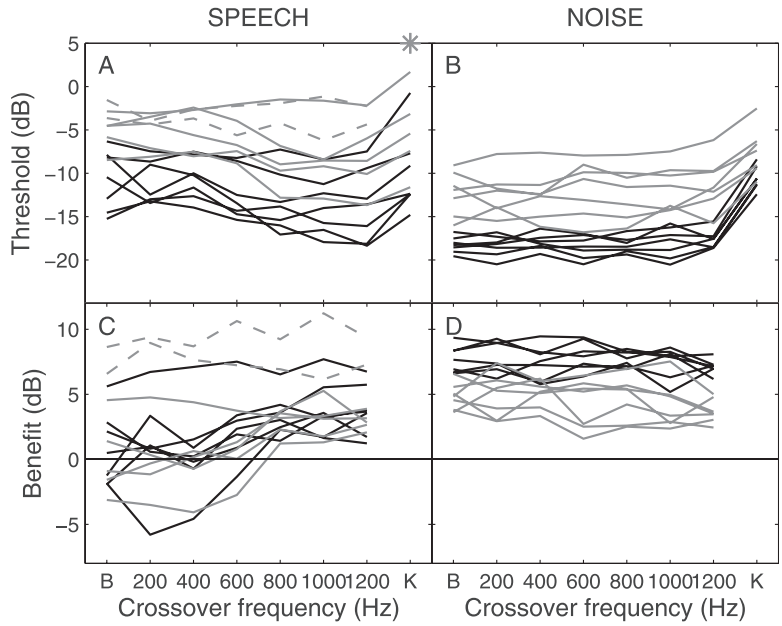

Fig. 2. (A) Individual speech reception thresholds as a function of crossover frequency for speech maskers. (B) As per panel (A) but for noise maskers. (C) Individual benefits (beamformer threshold relative to KEMAR threshold) for speech maskers. (D) As per panel (C) but for noise maskers. In all panels, NH and HI listeners are shown in black and gray, respectively. Dashed lines indicate two listeners for whom a threshold had to be estimated in the KEMAR condition for speech maskers (asterisk). The ends of the crossover frequency axis represent the full BEAM condition ("B") and the full KEMAR condition ("K").

shows group-mean data (and across-subject standard deviations) in a similar format to Fig. 2.

Thresholds were generally higher for the HI group than for the NH group, for both speech and noise maskers [Figs. 2(A), 2(B), 3(A), 3(B)]. The group difference for the KEMAR condition corroborates a wealth of previous studies reporting that HI listeners have trouble understanding speech in masked conditions, particularly when the maskers are competing talkers (e.g., Marrone *et al.*, 2008; Glyde *et al.*, 2013; Glyde *et al.*, 2015). The presence of a group difference in the BEAM and BEAMAR conditions is consistent with our previous findings (Kidd *et al.*, 2015; Best *et al.*, 2017), but here we show that it persists across all of the crossover frequencies tested.

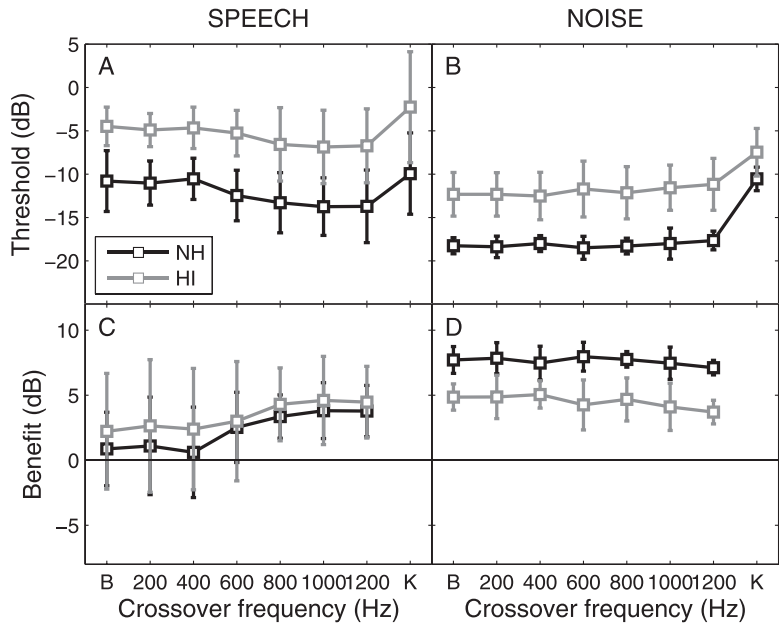

Fig. 3. Group mean speech reception thresholds and benefits plotted as per Fig. 2. Error bars show across-subject standard deviations.

In the presence of noise maskers, all listeners obtained a benefit of the beamformer [Figs. 2(D), 3(D)]. This benefit was robust across the entire range of crossover frequencies tested, declining by only 1 dB or so from the BEAM condition to the BEAMAR condition with the highest crossover frequency. A mixed analysis of variance (ANOVA) on the benefits in noise confirmed that there were significant main effects of group [$F(1,12) = 32.3$, $p < 0.001$] and frequency [$F(6,72) = 2.3$, $p = 0.04$], but no significant interaction [$F(6,72) = 0.9$, $p = 0.497$]. Planned comparisons (t-tests, $p < 0.05$) confirmed that the benefits were significantly greater than zero at all crossover frequencies for both groups. Moreover, the BEAMAR benefit at 1200 Hz was significantly lower than the full-spectrum BEAM benefit in the HI group, while there was no significant decline in benefit for any of the BEAMAR conditions relative to BEAM in the NH group (paired t-tests, $p < 0.05$). Overall, benefits of the beamformer in noise were smaller in HI listeners (4.5 dB) relative to NH listeners (7.6 dB). This may reflect reduced audibility in the high-frequency region where the beamformer provides the biggest SNR improvement. In support of this speculation, average benefits in the beamformer conditions showed a significant negative correlation ($r = -0.78$, $p = 0.04$) with high-frequency pure-tone averages (mean threshold across both ears at 4, 6, and 8 kHz) in the HI group.

In the presence of speech maskers, the benefit of the beamformer, and the patterns of performance as a function of crossover frequency, varied dramatically across listeners [Fig. 2(C)]. Some listeners showed a benefit for all crossover frequencies, but others showed a benefit only for the higher crossover frequencies. In the average data [Fig. 3(C)], a benefit was only apparent for crossover frequencies of 800, 1000, and 1200 Hz. The average benefit across these frequencies was 3.6 and 4.4 dB for NH and HI listeners, respectively. A mixed ANOVA on the benefits in speech maskers confirmed that there was a significant main effect of frequency [$F(6,72) = 9.9$, $p < 0.001$] but no main effect of group [$F(1,12) = 0.4$, $p = 0.540$] and no significant interaction [$F(6,72) = 0.4$, $p = 0.882$]. Planned comparisons (t-tests, $p < 0.05$) confirmed that the benefits were significantly greater than zero at 800, 1000, and 1200 Hz for both groups. Moreover, the benefits at these high crossover frequencies (800–1200 Hz in the NH group, 1000–1200 Hz in the HI group) were significantly larger than for the BEAM condition (paired t-tests, $p < 0.05$).

4. Conclusions

This study explored the optimal crossover frequency for a hybrid signal-processing strategy that combines beamforming at high frequencies with natural binaural signals at low frequencies. While the 16-channel microphone array used for beamforming is not representative of the beamformers present in current hearing aids, we believe these findings have implications for any system that involves a trade-off between directionality and spatial cues.

The optimal configuration of the hybrid beamformer depended on the listener, but overall the results were distinctly different for the two different kinds of interference. For speech masked by noise, the best performance was obtained when the full bandwidth was used for beamforming, thus maximizing the SNR, although performance suffered very little when the beamformer processing was restricted to frequencies above 1200 Hz. For speech masked by competing speech, on the other hand, the inclusion of natural binaural information at the low frequencies up to at least 800 Hz was necessary to obtain a reliable benefit from the beamformer.

These results suggest that the trade-off between SNR and spatial cues depends on the listening situation. When target intelligibility is hampered primarily by energetic masking (e.g., with background noise), the increase in SNR provided by a directional system can provide robust benefits, which is consistent with many studies and clinical reports showing benefits of beamformers for speech recognition in noise (e.g., Soede *et al.*, 1993b; Saunders and Kates, 1997; Picou *et al.*, 2014). However, for more complex listening situations where informational masking plays a substantial role (e.g., multitalker mixtures), the preservation of spatial cues that can support location-based segregation may be increasingly important.

Acknowledgments

This work was supported by NIH-NIDCD awards Nos. DC013286 and DC04545. The authors would like to thank Lorraine Delhorne for help with subject recruitment, and Pat Zurek, Jay Desloge, Todd Jennings, and Tim Streeter for technical support and helpful discussions.

References and links

Baumgärtel, R. M., Krawczyk-Becker, M., Marquardt, D., Völker, C., Hu, H., Herzke, T., Coleman, G., Adiloğlu, K., Ernst, S. M. A., Gerkmann, T., Doclo, S., Kollmeier, B., Hohmann, V., and Dietz, M. (**2015**). "Comparing binaural pre-processing strategies I: Instrumental evaluation," Trends Hear. **19**, 1–16.

Best, V., Roverud, E., Streeter, T., Mason, C. R., and Kidd, G., Jr. (**2017**). "The benefit of a visually guided beamformer in a dynamic speech task," Trends Hear. **21**, 1–11.

Byrne, D. J., Parkinson, A., and Newall, P. (**1991**). "Modified hearing aid selection procedures for severe-profound hearing losses," in *The Vanderbilt Hearing Aid Report II*, edited by G. A. Studebaker, F. H. Bess, and L. B. Beck (York Press, Parkton, MD), pp. 295–300.

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

Doclo, S., Gannot, S., Moonen, M., and Spriet, A. (**2010**). "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, edited by S. Haykin and K. J. Ray Liu (Wiley-IEEE, Hoboken, NJ), pp. 269–302.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Glyde, H., Buchholz, J. M., Nielsen, L., Best, V., Dillon, H., Cameron, S., and Hickson, L. (**2015**). "Effect of audibility on spatial release from speech-on-speech masking," J. Acoust. Soc. Am. **138**, 3311–3319.

Glyde, H., Cameron, S., Dillon, H., Hickson, L., and Seeto, M. (**2013**). "The effects of hearing impairment and aging on spatial processing," Ear Hear. **34**, 15–28.

Kates, J. M., and Weiss, M. R. (**1996**). "A comparison of hearing-aid array processing techniques," J. Acoust. Soc. Am. **99**, 3138–3148.

Kidd, G., Jr. (**2017**). "Enhancing auditory selective attention using a visually guided hearing aid," J. Speech Lang. Hear. Res. (in press).

Kidd, G., Jr., Best, V., and Mason, C. R. (**2008**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," J. Acoust. Soc. Am. **124**, 3793–3802.

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

Marrone, N., Mason, C. R., and Kidd, G., Jr. (**2008**). "The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms," J. Acoust. Soc. Am. **124**, 3064–3075.

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

Saunders, G. H., and Kates, J. M. (**1997**). "Speech intelligibility enhancement using hearing-aid array processing," J. Acoust. Soc. Am. **102**, 1827–1837.

Soede, W., Berkhout, A. J., and Bilsen, F. A. (**1993a**). "Development of a directional hearing instrument based on array technology," J. Acoust. Soc. Am. **94**, 785–798.

Soede, W., Bilsen, F. A., and Berkhout, A. J. (**1993b**). "Assessment of a directional microphone array for hearing-impaired listeners," J. Acoust. Soc. Am. **94**, 799–808.

Stadler, R. W., and Rabinowitz, W. M. (**1993**). "On the potential of fixed arrays for hearing aids," J. Acoust. Soc. Am. **94**, 1332–1342.

Van den Bogaert, T., Doclo, S., Wouters, J., and Moonen, M. (**2009**). "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," J. Acoust. Soc. Am. **125**, 360–371.

Völker, C., Warzybok, A., and Ernst, S. M. A. (**2015**). "Comparing binaural pre-processing strategies III: Speech intelligibility of normal-hearing and hearing-impaired listeners," Trends Hear. **19**, 1–18.