Original Article

[illegible]
[illegible] 24: 1 11
© [illegible] 2020
[illegible]
DOI: 10.1177/2331216519896908
[illegible]

Benefits of Beamforming With Local Spatial-Cue Preservation for Speech Localization and Segregation

Le Wang[1], Virginia Best[2], and Barbara G. Shinn-Cunningham[3]

Abstract

[illegible]

Keywords

[illegible]

[illegible] 6 [illegible] 2019; [illegible] 21 [illegible] 2019; [illegible] 1 [illegible] 2019

Introduction

In many realistic communication settings, a fundamental task of the listener is to perceptually segregate the various sources of sound, select one source upon which to focus attention, and then receive and process the information coming from the chosen source. The ability of listeners to succeed in this task varies widely and depends not only on the properties of the acoustic environment and types of competing sound sources that are present but also on a range of factors specific to individual listeners. For the task of understanding speech in the presence of competing talkers, some factors that have been shown to adversely influence performance include advanced age and hearing loss (e.g., Gallun, Diedesch, Kampel, & Jakien, 2013; Glyde, Cameron, Dillon, Hickson, & Seeto, 2013; Marrone, Mason, & Kidd, 2008) although even young normal-hearing (NH) listeners may vary widely in their abilities (e.g., Kidd et al., 2016; Ruggles & Shinn-Cunningham, 2011).

Many front-end signal-processing approaches have been proposed to improve speech recognition in noisy situations. The most successful of these approaches use directionality to improve the signal-to-noise ratio (SNR) and are a common feature of commercial assistive listening devices such as hearing aids (Launer, Zakis, & Moore, 2016) and cochlear implants (Loizou, 2006). Directional systems make use of multiple microphones and beamforming to emphasize sound sources from one direction and attenuate sound sources from other directions (see reviews in Doclo, Gannot, Moonen, & Spriet, 2010; Greenberg & Zurek, 2001). Such systems can make use of the microphones available in current hearing aids (i.e., two in a single hearing aid or four across a pair of hearing aids) or an array of microphones that may be

[1] [illegible] MA, USA
[2] [illegible] MA, USA
[3] [illegible] PA, USA

Corresponding Author:
[illegible] MA 02215, USA.
[illegible]

mounted on the head or on eyeglasses (e.g., Anderson et al., 2018; Greenberg, Desloge, & Zurek, 2003; Kidd, 2017). The spatial tuning can be extremely narrow in systems based on a large number of microphones, which can dramatically improve the SNR for a sound located in the focus of the beamformer. Under relatively simple conditions with a frontal speech source and one or more spatially separated noise sources, reported improvements in speech reception thresholds (SRTs) for beamformers relative to omnidirectional microphones range from around 5 to 12 dB (e.g., Luts, Maj, Soede, & Wouters, 2004; Saunders & Kates, 1997; Soede, Bilsen, & Berkhout, 1993).

A drawback of many beamforming strategies is that they combine the microphone signals to produce a single-channel output that conveys no binaural information. This obviously compromises the ability to localize sounds and may impede the segregation of competing sounds based on differences in spatial position as well as the ability to selectively attend to (or suppress) different sounds. To mitigate this problem, a variety of strategies have been proposed to preserve or restore spatial cues in beamformer systems (see reviews in Doclo et al., 2010; Kollmeier & Kiessling, 2016). Most of these strategies involve the combination of processed with unprocessed signals or the selective application of beamforming to some parts of the signal. Generally, beamformer systems designed for hearing-aid applications try to reach a balance between SNR improvement and spatial-cue preservation (Van den Bogaert, Doclo, Wouters, & Moonen, 2008, 2009) and thus may not provide speech-in-noise benefits as large as those that are theoretically possible. Indeed, recent studies that evaluated beamforming hearing aids under relatively complex listening situations found rather modest improvements in speech intelligibility relative to standard directional

The following study was designed to compare this full-bandwidth spatialization approach with the hybrid approach and to a full-bandwidth beamformer with no spatialization. These three strategies were compared with a reference condition in which the listener received natural, unprocessed binaural signals. The main task of interest was a speech-on-speech masking task like that we have used previously. A localization experiment was included to confirm that natural binaural cues were conveyed by the full-bandwidth spatialization approach. For each task, both broadband and high-pass speech conditions were included to confirm that useful spatial cues were provided at high frequencies as well as at low frequencies. The hypothesis was that full-bandwidth spatialization would support accurate horizontal localization, which would in turn support the ability of listeners to focus spatial attention effectively on the target, and thus maximize the advantage of beamforming under speech-on-speech masking conditions.

Methods

Participants

Fourteen adults participated in the study, seven with normal hearing (aged 18–40 years, mean age 23 years) and seven with bilateral sensorineural hearing impairment (aged 20–56 years, mean age 36 years). There was no significant age difference between the NH and HI groups, (12) = 2.02, = .07. The NH participants had pure-tone averages (PTAs; mean threshold across both ears at 0.5, 1 and 2 kHz) that ranged from 0 to 6.7 dB hearing level (HL; mean 3.8 dB HL). The HI participants had a range of losses with PTAs from 2.5 to 73.3 dB HL (mean 35.8 dB HL). The losses were relatively symmetric, with a PTA difference between the ears of no more than 10 dB. Participants were paid for their participation, gave informed consent, and all procedures were approved by the Boston University Institutional Review Board. Total testing time for the localization and speech intelligibility experiments was approximately 2.5 hr. We note that one NH participant was unable to complete the high-pass condition for both localization and speech intelligibility experiments, and another was unable to complete the broadband speech intelligibility experiment. Results are based on only six NH participants in these cases.

Beamforming and Spatialization

The different listening conditions were tested using a headphone simulation. Impulse responses were measured on an acoustic manikin (KEMAR) seated in a large sound-treated booth (IAC Acoustics). The inner dimensions of the booth were approximately 3.75 m 4 m 2.25 (Length Width Height). The manikin was seated halfway along one of the walls with a distance of about 0.6 m between its back and the wall. An array of loudspeakers (Acoustic Research 215PS) was arranged in front of the manikin at ear-height, at a distance of about 1.5 m. Loudspeakers were positioned between 90 and +90 azimuth at 7.5 intervals for the impulse response recordings, although only a subset of five positions was used for this study (see later). The manikin was fitted with a flexible headband that ran from ear-to-ear across the top of the head. The headband housed a microphone array (Sensimetrics Corporation), which consisted of 16 omnidirectional microphones arranged in four front-back-oriented rows. The rows were evenly spaced with a separation of 66.67 mm, for a total array length of 200 mm. More details about the array, including images of the microphone layout, can be found elsewhere (Kidd, 2017, Figure 6; Roverud, Best, Mason, Streeter, & Kidd, 2018, Figure 2).

Figure 1 provides an overview of the processing steps used to create stimuli for each condition. Two sets of impulse responses were recorded. One set of impulse responses, which captured the signals received by the manikin's in-ear microphones, 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

Figure 1. O a i i a i i. T i a i i IR a i KEMAR' a a a a KEMAR i i a i IR a i a a a BEAM i i. IR BEAMAR i i a i i LP i KEMAR IR a HP i i a a IR. F a i a -BEAM i i, IPD a ILD a KEMAR i a a a i BEAM i a. LP = - a ; HP = i - a ; IPD = i a a a i ; ILD = i a a i ; IR = i !

source location. These outputs were weighted and combined according to the optimal-directivity algorithm of Stadler and Rabinowitz (1993) for a look direction of 0 azimuth. The single-channel output for this condition ("BEAM") was presented diotically. The hybrid condition described earlier was simulated by combining low-pass-filtered binaural KEMAR impulse responses with high-pass-filtered BEAM impulse responses and is referred to as "BEAMAR." The crossover frequency was chosen to be 800 Hz, which was shown previously to be optimal (Best et al., 2017; Desloge et al., 1997).

The full-bandwidth spatialization strategy, referred to as "spatial BEAM," combined the spatial information from the KEMAR condition and the noise suppression from the BEAM condition across all frequencies. Specifically, for each spatial configuration, two spatial cues (interaural phase difference [IPD] and interaural level difference [ILD]) in each time–frequency tile were extracted from the binaural KEMAR signals. To do this, the KEMAR signal was broken down into time frames using a 92.9-ms hamming window (4,096 samples) that shifted by 23.2 ms (1,024 samples) for each frame. Within each frame, the spectrum of the left signal and the right signal was computed (frequency resolution = 10.8 Hz). The IPD and ILD were then calculated as the phase difference and the magnitude difference between the left and right signal for each frequency bin in the spectrum. These frequency-dependent IPDs and ILDs were then imposed on the corresponding time slice in the BEAM signal, creating a left and right signal per time frame. To do this, half of the IPD and ILD values were applied to the BEAM signal to create a left-ear signal; half of the IPD and ILD values inverted in polarity were applied to another copy of the BEAM signal to create a right-ear signal. The resynthesized binaural signals in each time frame were then summed to create a continuous output without additional temporal smoothing.

Stimuli

Target stimuli were taken from a 40-word corpus containing eight monosyllabic words in each of five distinct word categories (Kidd, Best, & Mason, 2008). Eight female voices were used in this study. Two speech bandwidth conditions were tested: broadband speech and high-pass speech. The broadband speech condition used the full spectrum speech signals in each processing condition, while the high-pass speech condition removed the frequency content below 800 Hz. The high-pass speech condition served as a control condition to test whether the [illegible] spatial information preserved in the spatial-BEAM condition offers useful information for localization and speech understanding, or whether benefits can only be obtained from the salient [illegible] spatial information. In the high-pass speech condition, the BEAMAR condition became identical to the BEAM condition because the low-frequency KEMAR portion of the BEAMAR signal was filtered out.

Stimuli were generated using MATLAB software (MathWorks Inc.) and presented via a 24-bit soundcard (RME HDSP 9632) through a pair of circumaural headphones (Sennheiser HD280 Pro). The participant was seated in a small sound-treated booth fitted with a computer monitor and mouse. For HI participants, individualized linear amplification according to the National Acoustic Laboratories' Revised, Profound (NAL-RP) prescription (Dillon, 2012) was applied to each stimulus just prior to presentation. A linear prescription was chosen to avoid potentially complicating interactions between nonlinear amplification and the different processing strategies.

Localization Experiment

A localization experiment was conducted to confirm that the spatial BEAM provided appropriate spatial information and produced lateral percepts in line with those produced by natural binaural stimuli (KEMAR). The stimuli were single words drawn at random from the speech corpus, presented at random from one of the five locations: 60 , 30 , 0 , +30 , and +60 azimuth. The nominal level of each word was 55 dB sound pressure level. Each word was processed according to one of the four conditions described earlier (KEMAR, BEAM, BEAMAR, and spatial BEAM), with the look direction of the beamformer always fixed at 0 . Trials were organized into blocks of 100 (five repetitions for each combination of processing condition and location), presented in a different random order for each participant. One block was completed for each of the two speech bandwidth conditions. Participants reported the perceived location of each stimulus by clicking on a graphical user interface showing a continuous arc representing the azimuthal plane from 90 to +90 . Before the experiment, each participant was given a training demo containing example trials from the KEMAR condition until they were familiar with the procedure.

Speech Intelligibility Experiment

The second experiment tested speech intelligibility in the presence of spatially separated competing talkers. Target stimuli were five-word sentences created by concatenating words from the speech corpus (with no added gaps between words). Each sentence had the form name-verb-number-adjective-noun (e.g., "Sue bought two red shoes"). The target sentence was spoken by one voice (chosen randomly on each trial from the set of eight) and was identified on the basis of the first word (which was always “Sue”). The target was presented simultaneously with four speech maskers. The speech maskers were also five-word sentences, assembled in the same manner as the target sentence. The five presented sentences were spoken by different talkers and had no words in common. The target was located at 0 azimuth, and the four maskers were located at 60 , 30 , +30 , and +60 azimuth. Each masker was presented at 55 dB constitutes a flipped tile, by counting only those in which the SNR was <10 dB before and >10 dB after beamforming, the flip rate never exceeded 4% (dashed line). We can conclude from this analysis that the number of tiles that flip from being masker-dominated to target-dominated was relatively low but not negligible. Thus, we considered it important to gain some intuition about the effect of these flipped tiles on the spatial representation of the target.

To this end, another analysis was conducted to estimate the binaural cues associated with the target signal before and after the spatial-BEAM processing. First, time–frequency tiles were identified that were dominated by the target in the original mixture. For these tiles, IPDs and ILDs were calculated from the KEMAR stimuli, and IPDs were transformed to interaural time differences (ITDs). Histograms of ITDs and ILDs are plotted in Figure 2(b) and (c) (black lines) for a mixture with a TMR of 0 dB. As expected for a centrally located source, the histograms are centered on 0-μs ITD and 0-dB ILD (any asymmetries are due to asymmetries in the impulse responses, including minor effects of the room and the alignment of KEMAR within it, etc.). Second, time–frequency tiles were identified that were dominated by the target after beamforming. For these tiles, ITD and ILD distributions were again calculated from the KEMAR stimuli (gray lines in Figure 2(b) and (c)). These distributions capture the binaural cues that would be applied to the target in the spatial-BEAM condition. The first thing to notice is that there are more target-dominated tiles after beamforming. Moreover, these tiles continue to be centered on 0-μs ITD and 0-dB ILD, although some spread in the histograms can be observed. In general, we can conclude that the binaural cues associated with the target talker are not drastically distorted by the tiles that were previously masker-dominated. A similar pattern was found when the same analysis was applied to each of the four masker talkers, although in those cases fewer tiles (and not more) were available after beamforming.

Results

Localization Experiment

Figure 3 shows the group average localization responses as a function of the true location in the broadband speech condition (a and c) and the high-pass speech condition (b and d) for NH participants (a and b) and HI participants (c and d). The gray line indicates a slope of location is equal to the true location. Shallower slopes indicate a weaker spatial percept with a slope of zero indicating that the responses were not at all related to the true location.

When tested with broadband speech, both NH and HI groups were able to localize the stimulus with reasonable accuracy in all conditions except the BEAM condition. A mixed analysis of variance (ANOVA) model using true location and processing condition as within-subjects factors and group as a between-subjects factor found a significant main effect of true location, $F(4, 48) = 287.61$, $< .001$, and a significant interaction between true location and processing condition, $F(12, 144) = 84.67$, $< .001$. No main effect of group was identified, $F(1, 12) = 0.53$, $= .48$, and none of the interactions involving group were significant. Follow-up ANOVAs comparing individual processing conditions indicated that the effect of true location in the spatial-BEAM condition was not significantly different from that in the KEMAR condition, $F(4, 52) = 0.68$, $= .61$, while the effect of true location in the BEAMAR condition was significantly different than in the KEMAR condition, $F(4, 52) = 6.81$, $< .001$. In terms of absolute localization error, when pooled across all trials for all locations for all participants, the mean values were 10 (KEMAR), 12 (spatial BEAM), 13 (BEAMAR), and 40 (BEAM).

With high-pass speech, because the spatial information in the BEAMAR condition was absent, participants were able to perceive late247.3g0rl
BEAM conditi(KEMAR), 1,(8()Tj/[(was)-TD(ditp3mit)-508.9(di38

was normalized by their KEMAR SRT so that the ordinate represents the provided by each beamforming condition relative to a natural binaural listening condition. As expected, the SRTs were generally higher in the HI group than in the NH group, for both the broadband speech (mean SRT 6 dB vs. 12 dB) and high-pass speech (mean SRT 4 dB vs. 9 dB).

A mixed ANOVA on the SRTs in the broadband speech condition identified a significant main effect of processing condition, $F(3,33) = 38.5$, $< .001$, and a significant main effect of group, $F(1, 11) = 10.44$, $= .008$, but no significant interaction, $F(3, 33) = 1.14$, $= .35$. Post hoc comparisons (paired tests with Bonferroni spatial BEAM provided the most benefit in speech intelligibility among all conditions without degrading localization accuracy. For high-pass speech, results for the BEAMAR condition look essentially the same as the BEAM condition, confirming its limitation in situations when only high-frequency information is available. On the other hand, the spatial-BEAM condition preserves reliable spatial information at high frequencies and thus leads to low localization errors. Because the spatial BEAM also inherits the improved SNR from beamforming, overall performance was better than in the KEMAR condition.

Discussion

This study provided behavioral data to assess the benefits of a signal-processing strategy that reimposes natural, full-bandwidth, binaural information on the output of a highly directional beamformer. This spatial-BEAM strategy may be a promising option for assistive hearing devices because it has the potential to combine the SNR advantage of beamforming with the perceptual benefit of spatialization. Groups of participants with and without hearing loss were tested using this approach on both sound localization and speech intelligibility tasks. As anticipated, the spatial-BEAM strategy supported horizontal localization performance (as measured with single speech sources) that was equivalent to that observed in the natural binaural condition. Moreover, the spatial-BEAM strategy significantly improved speech understanding in the presence of competing speech relative to other implementations of the beamformer.

While the HI group in our study had a poorer mean SRT than the NH group, we found no interaction between group and processing condition, suggesting that the benefit of spatialization was achieved by both transmitted to the wearer for different hearing-aid styles, ear pieces, vent sizes, and so on.

Acknowledgments

acoustic scenario. *H* , *353*, 36–48. doi: 10.1016/j.heares.2017.07.014

Picou, E. M., Aspell, E., & Ricketts, T. A. (2014). Potential benefits and limitations of three types of directional processing in hearing aids. *E* *H* , *35*(3), 339–352. doi: 10.1097/AUD.0000000000000004

Picou, E. M., & Ricketts, T. A. (2019). An evaluation of hearing aid beamforming microphone arrays in a noisy laboratory setting. *A* *A* *A* , *30*, 131–144. doi: 10.3766/jaaa.17090

Roverud, E., Best, V., Mason, C. R., Streeter, T., & Kidd, G. (2018). Evaluating the performance of a visually guided hearing aid using a dynamic auditory-visual word congruence task. *E* *H* , *39*(4), 756–769. doi: 10.1097/AUD.0000000000000532

Ruggles, D., & Shinn-Cunningham, B. (2011). Spatial selective auditory attention in the presence of reverberant energy: Individual differences in normal-hearing listeners. *A* , *12*(3), 395–405. doi: 10.1007/s10162-010-0254-z

Saunders, G. H., & Kates, J. M. (1997). Speech intelligibility enhancement using hearing-aid array processing. *A* *A* , *102*(3), 1827–1837. doi: 10.1121/1.420107

Schoenmaker, E., Brand, T., & van de Par, S. (2016). The multiple contributions of interaural differences to improved speech intelligibility in multitalker scenarios. *A* *A* , *139*(5), 2589–2603. doi: 10.1121/1.4948568

Soede, W., Bilsen, F. A., & Berkhout, A. J. (1993). Assessment of a directional microphone array for hearing-impaired listeners. *A* *A* , *94*(2 Pt 1), 799–808. doi: 10.1121/1.408181

Stadler, R. W., & Rabinowitz, W. M. (1993). On the potential of fixed arrays for hearing aids. *A* *A* , *94*(3 Pt 1), 1332–1342. doi: 10.1121/1.408161

Van den Bogaert, T., Doclo, S., Wouters, J., & Moonen, M. (2008). The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids. *A* *A* , *124*(1), 484–497. doi: 10.1121/1.2931962

Van den Bogaert, T., Doclo, S., Wouters, J., & Moonen, M. (2009). Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. *A* *A* , *125*(1), 360–371. doi: 10.1121/1.3023069

van Hoesel, R. J. (2004). Exploring the benefits of bilateral cochlear implants. *A* , *9*(4), 234–246. doi: 10.1159/000071020Tm26-1y