communications engineering

A Nature Portfolio journal

Article

https://doi.org/10.1038/s44172-025-00414-5

A brain-inspired algorithm improves "cocktail party" listening for individuals with hearing loss

Alexander D. Boyd [1,2,3] ✉, Virginia Best [1,3] & Kamal Sen [1,2,4]

Selective listening in competing-talker situations is an extraordinarily difficult task for many people. For individuals with hearing loss, this difficulty can be so extreme that it seriously impedes communication and participation in daily life. Directional filtering is one of few proven methods to improve speech understanding in competition, and most hearing devices now incorporate some kind of directional technology, although real-world benefits are modest, and many approaches fail in competing-talker situations. We recently developed a biologically inspired algorithm that is capable of very narrow spatial tuning and can isolate one talker from a mixture of talkers. The algorithm is based on a hierarchical network model of the auditory system, in which binaural sound inputs drive populations of neurons tuned to specific spatial locations and frequencies, and the spiking responses of neurons in the output layer are reconstructed into audible waveforms. Here we evaluated the algorithm in a group of adults with sensorineural hearing loss, using a challenging competing-talker task. The biologically inspired algorithm led to robust intelligibility gains under conditions in which a standard beamforming approach failed. The results provide compelling support for the potential benefits of biologically inspired algorithms for assisting individuals with hearing loss in "cocktail party" situations.

One of the most challenging listening tasks encountered by people in their daily lives is to understand what a talker of interest is saying in an acoustically cluttered environment, especially one that contains other people talking at the same time[1,2] (the proverbial "cocktail party problem"). Listeners with sensorineural hearing loss (including older people with typical age-related declines in hearing) experience extreme difficulties in these situations, and in many cases ease-of-communication and willingness to attend social gatherings is seriously impeded[3,4]. We are now beginning to understand the devastating longer-term consequences of untreated hearing difficulties, which include declines in cognitive health, and the associated societal and economic costs[5].

One proven way to improve speech understanding in noise is via directional filtering[6]. Directional microphones are by now almost ubiquitous in hearing aids and provide broad filtering that attenuates sounds arising from behind the listener to improve the signal-to-noise ratio (SNR) for those in front. Narrower tuning can be achieved by using larger numbers of microphones[7] (e.g., arranged in an array on a headband or eyeglasses). A number of two-channel (binaural) beamforming algorithms achieve narrow tuning by combining the signals at the left and right ears[8]. One issue with most previous solutions is that in order to improve the SNR, they reduce multichannel inputs to a single-channel output, thus sacrificing information related to the spatial location of sound sources that is available with multiple receivers. This means that individual sound sources are not heard at their original locations, which can be disorienting, and also can disrupt location-based segregation of sound mixtures, which is especially critical in complex situations with competing talkers[9,10]. Deep neural network approaches to sound segregation have made impressive leaps in recent years and under the right conditions can effectively isolate a sound of interest from a complex mixture[11,12]. These approaches, however, are computationally expensive and not yet well-suited for low-power real-time applications[13].

We recently developed a biologically oriented sound segregation algorithm[14] (BOSSA), which is designed to separate competing sounds based on differences in spatial location. Taking its inspiration from the binaural auditory system, this algorithm requires only two input signals and produces two output signals that preserve natural spatial cues (i.e., interaural differences in time and level). It is also well-suited for low-power real-time applications and thus could have real utility in wearable hearing-assistive devices. BOSSA was developed and optimized using objective intelligibility

[1]Hearing Research Center, Boston University, Boston, MA, USA. [2]Department of Biomedical Engineering, Boston University, Boston, MA, USA. [3]Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA, USA. [4]Neurophotonics Center, Center for Systems Neuroscience, Boston University, Boston, MA, USA. ✉e-mail: adboyd@bu.edu

measures and has been evaluated behaviorally in a group of young listeners with audiometrically normal hearing[14]. In this population, BOSSA provided robust improvements in the intelligibility of a target talker embedded in a challenging speech mixture. To date, BOSSA has not been evaluated in the population who most stand to benefit from assistance in cocktail party situations, which is individuals with hearing loss.

In the current study, we recruited adults with bilateral sensorineural hearing loss and measured the benefits provided by BOSSA for the task of understanding one talker in a mixture of five competing talkers. We compared the performance of BOSSA to a binaural implementation of the current industry standard beamforming approach (minimum variance distortionless response, or MVDR) employed in hearing aids[8].

Results

Figure 1 shows the average proportion of words correctly identified as a function of target-masker ratio (TMR) for four different multitalker scenarios (panels A-D), and for four different sound processing conditions (colored lines). These functions show that performance improved systematically with TMR, as expected, but also differed systematically between processing conditions. In each scenario, performance for two different versions of BOSSA (DiffMask and RatioMask) was better than performance for the Natural (unprocessed) condition and was also better than for the standard MVDR beamformer condition.

These patterns are summarized in Fig. 2, which shows group-mean speech reception thresholds (SRTs) for each processing condition in each experiment. For each experiment, a one-way repeated-measures ANOVA found a significant main effect of processing condition (Exp 1, 0° [$F(3,21) = 50.77$, $p < 0.001$, $\eta p^2 = 0.879$], Exp 1, −30° [$F(3,21) = 41.22$, $p < 0.001$, $\eta p^2 = 0.855$], Exp 1, +30° [$F(3,21) = 106.74$, $p < 0.001$, $\eta p^2 = 0.938$], Exp 2 [$F(3,9) = 29.43$, $p = <0.001$, $\eta p^2 = 0.908$]). Planned comparisons (paired t-tests, $p < 0.01$) indicated that for Experiment 1 (0° and −30°) and Experiment 2, both versions of BOSSA resulted in better SRTs than the Natural condition, whereas for Experiment 1 (+ 30°) this was only true for RatioMask. In all cases, SRTs for MVDR and Natural were not significantly different.

Figure 3 shows group-mean SRT differences for each of the processing conditions compared to the Natural condition. These differences are expressed in terms of a processing "benefit", where a positive value in dB corresponds to a decrease in SRT, and a negative value in dB corresponds to an increase in SRT.

Discussion

Benefits of BOSSA in relation to natural listening and beamforming

We have developed a brain-inspired algorithm (BOSSA) that can selectively enhance one talker in a multitalker mixture according to its spatial location. The algorithm has a wide variety of potential uses, including in hearing-assistive devices, where it could provide a benefit in challenging communication situations to millions of people with hearing difficulties. In the current study we evaluated BOSSA in adults with bilateral sensorineural hearing loss and compared its performance to a standard beamforming approach used in hearing aids.

Our evaluation used a challenging speech mixture consisting of five female talkers uttering sentences that were highly confusable with each other. This amounted to a speech intelligibility task with extremely high "informational masking."[2] We considered two versions of this task (Experiments 1 and 2) to confirm that the results were not dependent on specific details of the task, and we considered three different target locations,

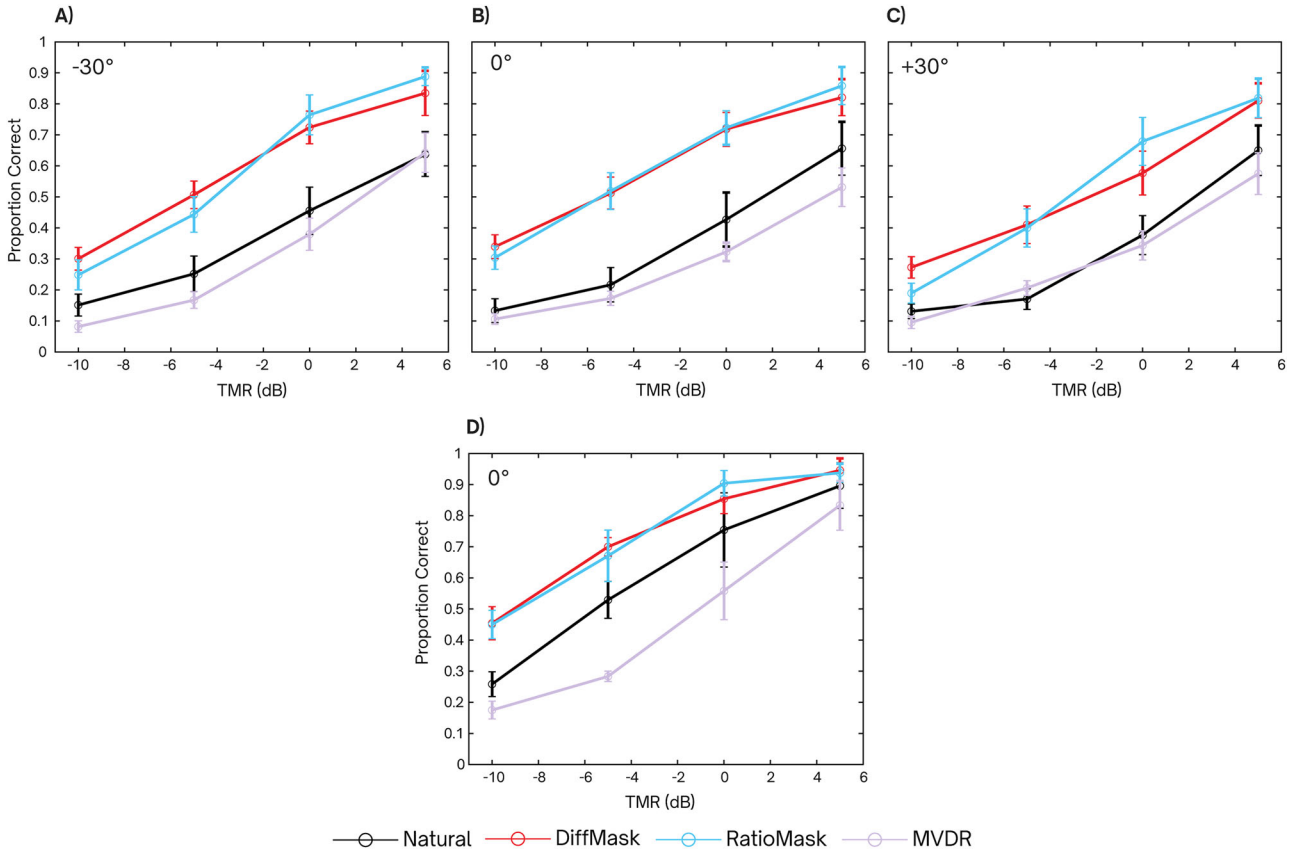

Fig. 1 | All subjects average word recognition scores. A–C Average proportion of correctly recognized words as a function of target-masker-ratio (TMR) for Experiment 1 ($n = 8$) with target locations of −30°, 0°, +30° respectively. **D** Average proportion of correctly recognized words as a function of TMR for Experiment 2 ($n = 4$) with a target location of 0°. Error bars show across-subject standard deviations.

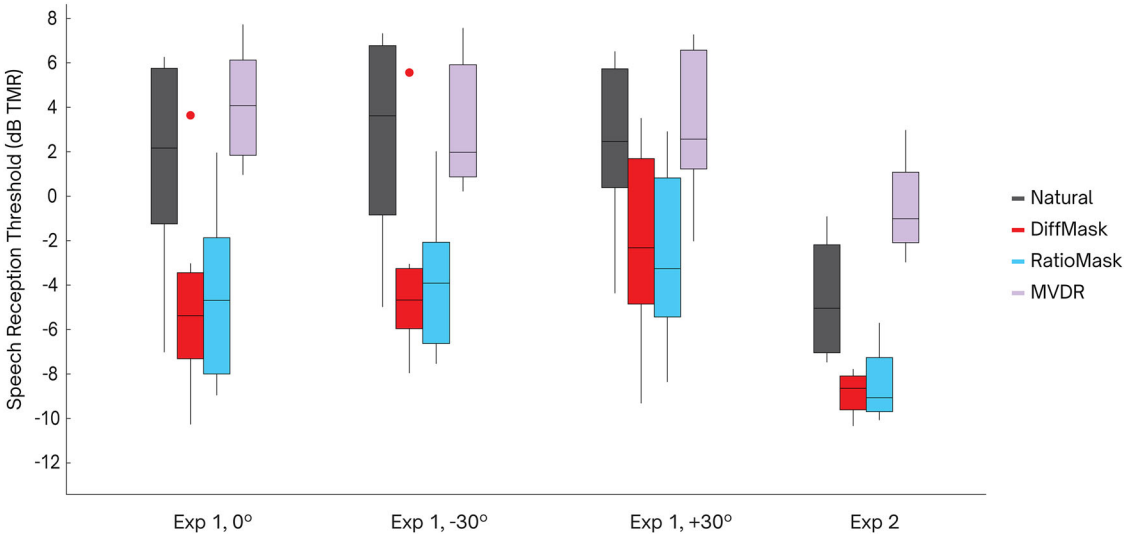

Fig. 2 | Speech reception thresholds (SRT) are shown as boxplots for each processing condition in Experiment 1 ($n = 8$) and Experiment 2 ($n = 4$). Box limits represent upper and lower quartiles, and the box centerline indicates the median. Whiskers represent maximum and minimum SRT benefits, excepting outliers demarcated with a dot that exceeded 1.5 times the interquartile range.

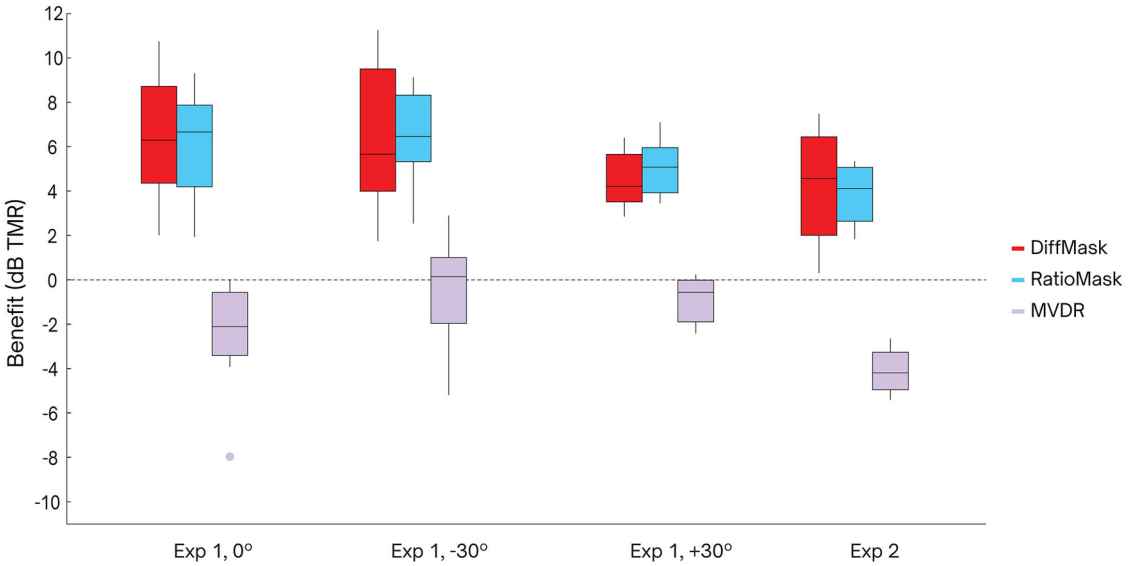

Fig. 3 | Processing benefits are shown as boxplots for each processing condition in Experiment 1 ($n = 8$) and Experiment 2 ($n = 4$). Box limits represent upper and lower quartiles, and the box centerline indicates the median. Whiskers represent maximum and minimum SRT benefits, excepting outliers demarcated with a dot that exceeded 1.5 times the interquartile range.

motivated by the idea that a listener may wish to listen to a target that is not directly in front and an assistive algorithm may need to be "steered" to that location. BOSSA improved speech reception thresholds consistently across these variations relative to the Natural condition with no processing. While the benefit was robust, its magnitude varied across participants and configurations, ranging from 0.3 to 11.3 dB. No participant performed worse with BOSSA than in the Natural condition in any configuration.

Multichannel beamformers like MVDR are designed to provide robust benefits in situations where speech is masked by stationary noise ("energetic masking"; air conditioning noise, etc.) and in such cases can improve speech intelligibility, speech quality, listening effort, and numerous objective measures[15]. At the outset, it was unclear how the MVDR algorithm would perform in our challenging multitalker scenario. We found that this approach did not provide a significant benefit, identifying a potential factor in the failure of current hearing aids to provide robust benefits in cocktail party settings. To confirm our binaural implementation of MVDR was working as expected, we conducted a control experiment using a more traditional speech-in-noise design. For this experiment, we brought back four participants from Experiment 1 (the same four who completed Experiment 2). The target was identical to the 0° target from Experiment 1, and the maskers were two independent speech-shaped noises positioned at -90° and +90° azimuth. The spectrum of these noises was matched to the long-term average spectrum of the set of female talkers in the corpus, and the noises were matched in length to the target on each trial. A repeated-measures ANOVA found a significant effect of processing condition [$F(3,9) = 36.69$, $p < 0.001$, $\eta p^2 = 0.924$]. Planned comparisons showed performance for MVDR was better than for the Natural condition (benefit of 1.6 dB; $p = 0.041$), confirming that MVDR was working as expected

(Supplemental Fig. 1). The performance for the two variations of BOSSA was lower than the Natural condition for these stimuli (DiffMask by 3.9 dB; $p = 0.002$, RatioMask by 3.6 dB; $p = 0.004$), confirming that the algorithm in its current form performs best in conditions involving fluctuating maskers. Future versions of BOSSA that are optimized for stationary noise conditions and/or which incorporate postprocessing like that commonly found in beamforming may improve the performance of BOSSA under these conditions.

Comparison to other sound source segregation approaches

BOSSA is not the first algorithm to leverage binaural cues to isolate a sound of interest. Roman et al.[16] proposed a related mask-based approach that uses supervised learning, applied to specific scenarios and within frequency bands, to estimate a binary mask from distributions of interaural time and level differences. While this approach is not highly practical for a hearing-aid application due to the complexity of its mask calculation and reliance on training data, it provides robust SNR improvements, increases automatic speech recognition accuracy, and shows intelligibility improvements in NH listeners under some conditions. The approach was later extended to reverberant situations, with good results according to objective metrics, but no intelligibility data were presented[17].

Another class of algorithms uses much larger numbers of microphones to achieve spatial filtering. To give one example, Kidd and colleagues[18] developed a hybrid beamformer ("BEAMAR"), which combines the output of a 16-channel beamforming microphone array with natural low-frequency cues to preserve spatial information. Best et al.[9] tested this beamformer for a frontal speech target against four symmetrically separated maskers (a very similar setup to the current study) and produced robust benefits for NH and HI populations. A direct comparison of BOSSA and BEAMAR in a new group of NH listeners[14] confirmed that the benefits were comparable for the two methods, despite BOSSA using only two microphones compared to 16 microphones for BEAMAR.

In their large comparative study, Volker et al.[19] fitted NH and HI participants with a hearing-aid simulator programmed to run eight state-of-the-art "pre-processing strategies" that included binaural algorithms as well as several single and multichannel noise-reduction algorithms. SRT benefits were broadly similar across strategies (on the order of 2 to 5 dB across three different background noises). The current study suggests that BOSSA would outperform these state-of-the-art strategies for listening scenarios containing competing talkers and substantial amounts of informational masking. Future work is needed, however, to comprehensively test BOSSA in a wide variety of listening scenarios and in a larger and more diverse listener population.

Deep neural network (DNN) approaches to sound source segregation are rapidly evolving. In a recent review of their single-channel DNN-based noise-reduction strategy, Healy et al.[11] point out the large strides that have been made since its inception a decade ago, both in terms of efficacy for HI listeners and viability for real-time implementation. Similar results are emerging from other groups[12], and many of the major hearing-aid manufacturers are now incorporating DNN-based noise reduction into their premium hearing devices[20,21]. It is worth noting that although these approaches achieve impressive results for speech in noise, they remain challenged by competing talkers, an important real-world scenario for listeners. Another challenge for DNN-based approaches is the requirement of large labeled datasets for training. Generating such a dataset is labor-intensive and potentially costly. Moreover, even when such training datasets are available, generalization to scenarios that are not part of the training has not been convincingly demonstrated even for state-of-the-art DNNs. Part of the issue here is that it is difficult to thoroughly sample the vast number of possible configurations of targets and maskers in a complex scene with multiple sources in a training dataset. Finally, despite their impressive performance under many conditions, the power consumption required for the highly intensive computational demands of DNNs is also a factor that continues to impact their adoption in miniature wearable devices such as hearing aids. In this context, it is worth noting that the neurally inspired architecture of BOSSA makes it well-suited for power-efficient neuromorphic implementations.

Limitations and future work

While BOSSA provided robust improvements in speech intelligibility under the conditions of our experiments, there are current limitations that could hinder its performance in real-world scenarios. For example, in this evaluation, the number and location of spatially tuned neurons (STNs) were fixed according to the known locations of the competing talkers in the mixture. While the extent to which STN to source alignment impacts model performance is not yet fully understood, simulations in three talker scenes suggest modest changes in objective measures as a function of STN misalignment. Moreover, STN misalignment effects may be minimized with a dense array of STNs. Computational headroom currently limits the number of STNs that contribute to the reconstruction mask calculation with acceptable latency, but intelligent approaches for selecting the ideal set of STNs for reconstruction may yield robust real-world performance while maintaining the algorithm's high efficiency. Future versions of BOSSA could use a dynamic set of STNs capable of selectively isolating sounds from any direction given any arbitrary mixture of competitor locations. This could be accomplished by decoding the neural activity of STNs to first perform source localization and prioritization before proceeding to source isolation.

Another avenue we are actively pursuing involves making BOSSA responsive to the changing needs of the user. For example, future implementations of BOSSA will take input directly from the user to dynamically change focus instead of assuming a fixed look direction. One such method that is actively being explored is to use eye gaze patterns as an input source, creating a means for informed steering or tuning of the spatial filter. Behavioral studies using such an approach suggest that listeners with and without hearing loss are able to use and gain speech intelligibility benefits from this kind of system[22–24].

The results provide compelling support for the potential benefits of biologically inspired algorithms for providing hearing assistance in cocktail party situations. Of course, our investigation was limited to a small sample of individuals with hearing loss, and a more comprehensive evaluation study is warranted. In particular, it will be important to examine whether the benefits we observe generalize to older listeners with age-related hearing loss, and to individuals with highly asymmetric losses. Moreover, with a hearing-aid application in mind, our implementation included linear gain to broadly compensate for different hearing-loss profiles. It remains to be seen whether benefits could be further increased with the use of compressive gain like that used in current hearing aids.

Methods

Participants

Participants in the experiments were eight adults (ages 20-42 years) with bilateral sensorineural hearing loss. Their hearing losses represented a wide variety of configurations and severities but were all bilaterally symmetric; across-ear average audiograms are shown in Fig. 4. Participants were paid for their participation and gave written informed consent. All procedures were approved by the Boston University Institutional Review Board, and all methods were performed in accordance with relevant guidelines and regulations. All participants completed Experiment 1, and a subset of four participants completed Experiment 2.

Stimuli

Five-word sentences were constructed from a corpus of monosyllabic words[25] with the form [name-verb-number-adjective-noun] (e.g., "Sue found three red hats"). The corpus contains eight words in each of the five categories. On each trial, a target sentence was mixed with four masker sentences. The target sentence was designated by the name "Sue". The five sentences were simulated to originate from five spatial locations (0°, ±30°, and ±60° azimuth) by convolving with anechoic head-related transfer functions measured on an acoustic manikin[26] at a distance of 1 m. The level

of the target was varied to achieve target-to-masker ratios (TMRs) of −10, −5, 0, and 5 dB. The nominal presentation level of the mixture (post-processing, see below) was 62 dB SPL. On top of this, each listener was given linear frequency-dependent gain to compensate for their audiogram according to the NAL-RP prescriptive formula[6].

In Experiment 1 (Fig. 5A), each word in a sentence was spoken by a different female talker, randomly chosen from a set of eight female talkers, without repetition. In this experiment, for each word category, the target and masker words were time-aligned at their onsets, and zero-padding was applied to align their offsets. This resulted in highly synchronized sentences. The design of these stimuli was intended to reduce the availability of voice and timing-related cues and, as such, increase the listener's reliance on spatial information to solve the task. In three separate sub-experiments, the target sentence was presented from 0°, −30°, and +30° azimuth, and the four maskers were presented from the other four non-target locations in the set.

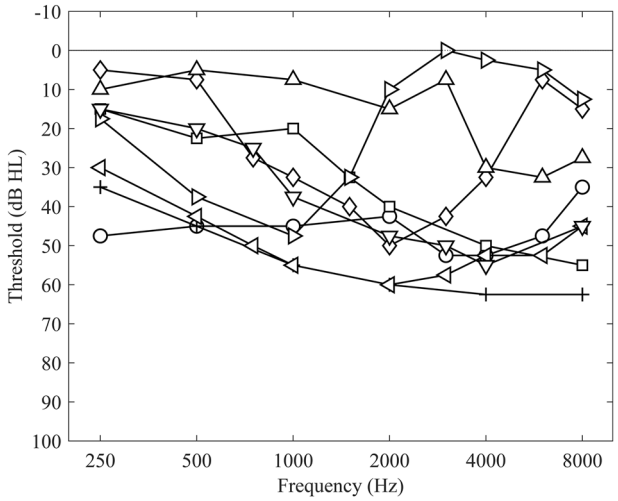

Fig. 4 | Individual participant audiograms. The different curves show pure-tone thresholds for each of the eight participants (averaged over left and right ears). Unique symbols distinguish individual subjects.

In Experiment 2 (Fig. 5B), two modifications were made to ensure that the results were not dependent on specific choices made in Experiment 1. In this experiment, the words in each sentence were spoken by the same talker, such that there was voice continuity as well as spatial continuity within each competing sentence. In addition, the onset of words was not time-aligned across sentences, but individual word recordings were simply concatenated (with no additional gaps) to create each sentence. For Experiment 2, only the center target location was examined.

Procedures

Each experiment was comprised of 12 blocks of trials (three blocks for each of the four processing conditions, see below). Each block contained five trials at each of the four TMRs (20 total trials per block). The order of presentation of TMRs within a block, and the order of blocks for each participant, were chosen at random. Each experiment took approximately one hour to complete.

Stimuli were controlled in MATLAB 2019a (MathWorks Inc., Natick, MA) and presented at 44.1 kHz sampling rate through an RME HDSP 9632 24-bit soundcard (RME Audio, Bayern, Germany) to Sennheiser HD280 Pro headphones (Sennheiser Electronic GmbH & Co., Wedemark, Germany). The sound system was calibrated at the headphones with a sound level meter (type 2250; Brüel & Kjær, Nærum, Denmark). Participants were seated in a double-walled sound-treated booth. A computer monitor inside the booth displayed a graphical user interface containing a grid of 40 words (five columns of eight words, each column corresponding to one-word category). On each trial, participants were presented with a mixture and were instructed to listen for the target sentence. Responses were provided by choosing one word from each column on the grid with a computer mouse. The stimulus setup and the target location (left, center, right) were described to the participant prior to each experiment.

Performance was evaluated by calculating the percentage of correctly answered words (excluding the first word, which was always "Sue") across all trials at each TMR. Psychometric functions were generated by plotting the percent correct as a function of TMR and fitting a logistic function to those data. SRTs, defined as the TMRs corresponding to 50% correct, were extracted from each function using the psignifit toolbox[27].

Processing conditions

The four processing conditions included one control condition that simulated the natural listening condition ("Natural"), two variations of BOSSA

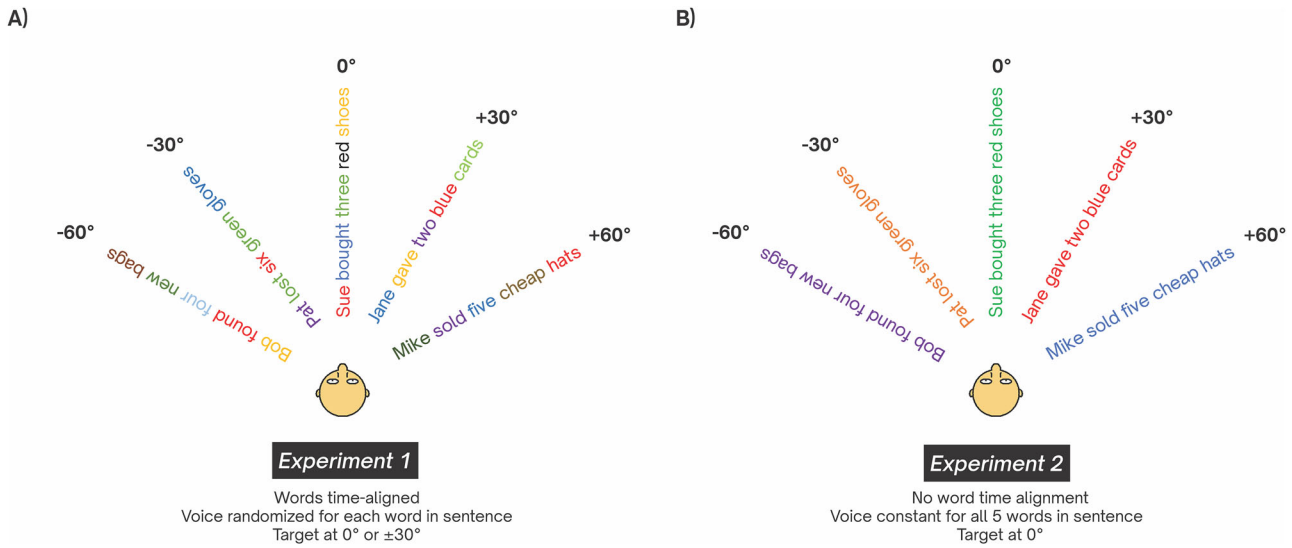

Fig. 5 | Stimulus layouts for Experiments 1 and 2. A Talker structure and spatial configuration of the five competing sentences in Experiment 1. The target source locations tested were 0°, −30°, and +30° azimuth. **B** Talker structure and spatial configuration of the five competing sentences in Experiment 2. The target source location tested was 0°. Different colors depict different female talkers.

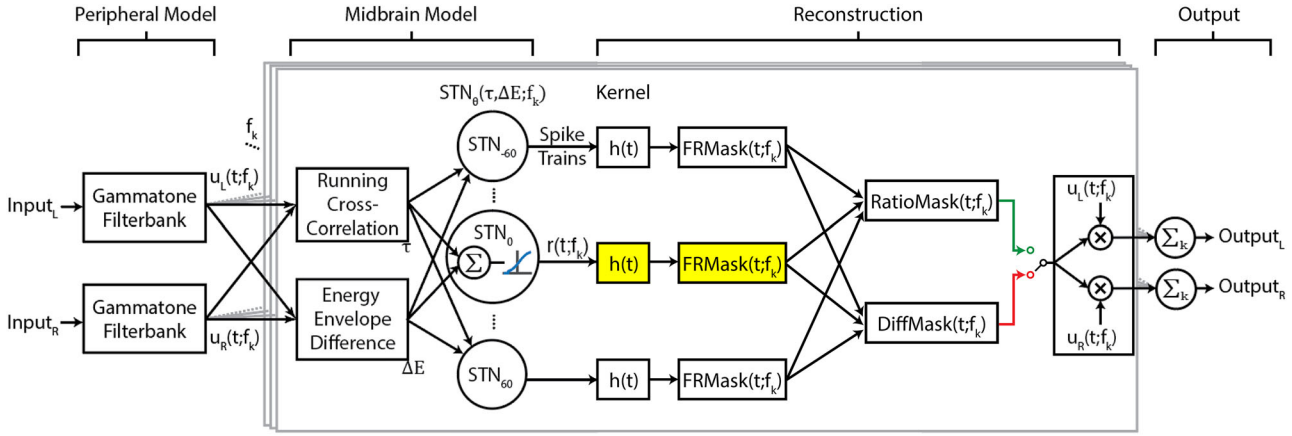

Fig. 6 | Flow diagram of the biologically oriented sound segregation algorithm (BOSSA). Central boxes, outlined in gray, show processing for a single frequency band. The functions $u_L(t;f_k)$ and $u_R(t;f_k)$ are the narrowband signals of the left and right input channels for each frequency channel, and f_k denotes the k^{th} frequency channel. The midbrain model is based on spatially tuned neurons (STNs), where each STN has a "best" interaural timing and level difference (ITD and ILD), denoted τ and ΔE, respectively. The best ITD and ILD values of a neuron depend on the direction θ and frequency f_k to which the STN is tuned. $h(t)$ represents the reconstruction kernel that converts spike trains to waveforms. Either RatioMask (green line) or DiffMask (red line) was used for reconstruction, as indicated by the switch. The target STN for reconstruction (yellow highlight) is manually selected as a parameter. The implementations of DiffMask and RatioMask in our analysis involve five sets of STNs, where $\theta \in \{0, \pm 30, \pm 60\}$; however, other implementations of the model may involve different sets of STNs.

("DiffMask" and "RatioMask"), and a standard binaural beamformer ("MVDR").

BOSSA. Using an approach inspired by ideal time-frequency mask estimation[28], BOSSA separates an incoming binaural audio signal into time-frequency bins and then selectively applies gain to bins such that sound energy arising from a prescribed target location is preserved while sound energy from non-target locations is suppressed. As described in Chou et al.[14], the gain for each time-frequency bin was calculated by combining the spiking activity of five sets of STNs with target angles $\theta \in \{0°, \pm 30°, \pm 60°\}$. Each STN's directional selectivity was based on binaural cues (interaural time and level differences) for the desired angle.

Two versions of BOSSA were evaluated. These algorithms differed only in the content of the reconstruction module responsible for converting ensembles of neural spikes into the final binaural audio output, as depicted in Fig. 6.

The first version of BOSSA used a previously studied reconstruction method, DiffMask, which was inspired by lateral inhibition observed in biological networks. In this technique, the scaled sum of non-target STN firing rates ($\sum FRMask_{\theta_{nontarget}}$) was subtracted from the target STN firing rate activity ($FRMask_{\theta_{target}}$). A lower limit of 0 was imposed on the DiffMask output to prevent unrealistic negative firing rates. The subtractive term acts as a mechanism for sharpening the spatial tuning of output neurons, as demonstrated in a previous publication[14]. The scale factor a was adjusted to normalize the mask to a maximum value of 1.

$$Diff\,Mask = a \cdot \max\left\{FRMask_{\theta_{target}} - 0.5 \sum FRMask_{\theta_{nontarget}},\ 0\right\}$$

The second version of BOSSA tested in this study utilized a new reconstruction method called RatioMask, intended to reduce some unnatural artifacts induced by BOSSA with DiffMask.

$$RatioMask = b \cdot FRMask_{\theta_{target}} \cdot \left[\frac{FRMask_{\theta_{target}}}{\left(FRMask_{\theta_{target}} + \sum FRMask_{\theta_{nontarget}}\right)}\right]^{\beta}$$

In contrast to DiffMask, which employs a subtractive operation, RatioMask implements a multiplicative operation to sharpen the spatial tuning of output neurons in the presence of competing noise. The multiplicative term, inspired by the ideal ratio mask (IRM) operation[28], is an estimate of the SNR raised to a power β. Here, the firing rate of the neuron at the target location serves as an estimate of the "signal" and the firing rates at non-target locations serves as an estimate of the "noise". Thus, the multiplicative term boosts time-frequency tiles with a higher SNR and suppresses time-frequency tiles with a lower SNR. We found that a value of $\beta = 1.65$ gave the best results as quantified by an objective intelligibility metric (see below). The normalization scale factor b was adjusted to yield a maximum value of 1 for the RatioMask. After a given mask calculation, the mask was then applied (i.e., point-multiplied) to the left and right channels of the original sound mixture. Lastly, the sum of each frequency channel of the mask-filtered signal was taken to obtain an audible, segregated waveform. In contrast to DiffMask, RatioMask applies a gain factor without any "hard" thresholding operation, which led to a smoother, more natural-sounding output based on our listening experience.

BOSSA parameters. Most model parameters were fixed to biologically plausible values for the implementation of BOSSA evaluated here, rather than chosen based on an extensive optimization process. Variation in some reconstruction parameters, however, was explored using the Short-Time Objective Intelligibility (STOI) measure[29]. STOI ranges between 0 and 1 and can be used to predict speech intelligibility when combined with an appropriate mapping function. The time-constant of the alpha function kernel (τ_h), and the scaling factor for DiffMask (a) (see Chou et al.[14] for details) were chosen by iteratively trying a range of values, quantifying algorithm performance using STOI, then choosing parameter values that produced the highest average STOI. For RatioMask, the beta (β) parameter value of 1.65 was selected using the genetic algorithm (GA) function in the MATLAB Global Optimization Toolbox with "fitness" defined as the average STOI value. While we do not claim that either approach outputs the optimal parameter set for reconstruction, our experience and the behavioral results indicate the chosen parameter values produced adequate reconstructions that supported robust speech intelligibility.

Binaural MVDR. To compare to a widely used spatial processing algorithm, stimuli were also processed with a binaural MVDR beamformer[8,30]. To enable a direct comparison to the two-channel BOSSA approach, the binaural MVDR implementation used two virtual microphones (one per ear). Relative transfer function vectors aimed toward the target source angle were calculated for each ear using the same

KEMAR HRTFs that were used in BOSSA. Log-level voice activity detection was performed for each ear along with the MVDR application followed by a multichannel Wiener filter with the decision directed approach[31]. To aid in the preservation of some spatial cue information, −16 dB of the original unprocessed signal was blended with the MVDR-enhanced output.

Statistics and reproducibility

For each experiment, a one-way repeated-measures ANOVA (sphericity assumed) was conducted on SRTs with processing condition as the within-subjects factor. Planned comparisons were made between each pair of processing conditions using one-sided paired sample *t*-tests. All statistical analysis was performed using IBM SPSS Statistics (Version 29).

Data availability

The raw data collected are available as open data via Figshare with the identifier https://doi.org/10.6084/m9.figshare.28636433.v1.

Received: 6 November 2024; Accepted: 8 April 2025;
Published online: 22 April 2025

References

1. Cherry, E. C. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* **25**, 975–979 (1953).
2. Kidd, G. & Colburn, H. S. in *The Auditory System at the Cocktail Party* (eds. Middlebrooks, J. C., Simon, J. Z., Popper, A. N. & Fay, R. R.) 75–109 (Springer International Publishing, Cham, 2017).
3. Arlinger, S. Negative consequences of uncorrected hearing loss–a review. *Int. J. Audiol.* **42**, 2S17–20S17 (2003).
4. Podury, A., Jiam, N. T., Kim, M., Donnenfield, J. I. & Dhand, A. Hearing and sociality: the implications of hearing loss on social life. *Front. Neurosci*. **17**, 1245434 (2023).
5. WHO. *World Report on Hearing*. https://www.who.int/publications-detail-redirect/9789240020481(2021).
6. Dillon, H. *Hearing Aids*. (Thieme, Sydney, 2012).
7. Desloge, J. G., Rabinowitz, W. M. & Zurek, P. M. Microphone-array hearing aids with binaural output. I. Fixed-processing systems. *IEEE Trans. Speech Audio Process.* **5**, 529–542 (1997).
8. Doclo, S., Kellermann, W., Makino, S. & Nordholm, S. Multichannel signal enhancement algorithms for assisted listening devices. *Signal Process. Mag. IEEE* **32**, 18–30 (2015).
9. Best, V., Roverud, E., Mason, C. R. & Kidd, G. Examination of a hybrid beamformer that preserves auditory spatial cues. *J. Acoust. Soc. Am.* **142**, EL369–EL374 (2017).
10. Wang, L., Best, V. & Shinn-Cunningham, B. G. Benefits of beamforming with local spatial-cue preservation for speech localization and segregation. *Trends Hear*. **24**, 2331216519896908 (2020).
11. Healy, E. W., Johnson, E. M., Pandey, A. & Wang, D. Progress made in the efficacy and viability of deep-learning-based noise reduction. *J. Acoust. Soc. Am.* **153**, 2751–2768 (2023).
12. Diehl, P. U. et al. Restoring speech intelligibility for hearing aid users with deep learning. *Sci. Rep.* **13**, 2719 (2023).
13. Drgas, S. A survey on low-latency DNN-based speech enhancement. *Sensors* **23**, 1380 (2023).
14. Chou, K. F., Boyd, A. D., Best, V., Colburn, H. S. & Sen, K. A biologically oriented algorithm for spatial sound segregation. *Front. Neurosci*. **16**, 1004071 (2022).
15. S. Doclo, S. Gannot, M. Moonen & A. Spriet. in *Handbook on Array Processing and Sensor Networks* (eds. Haykin, S. & Ray Liu, K. J.) 269–302 (Wiley-IEEE Press, Hoboken, New Jersey, 2010).
16. Roman, N., Wang, D. & Brown, G. J. Speech segregation based on sound localization. *J. Acoust. Soc. Am.* **114**, 2236–2252 (2003).
17. Roman, N. & Wang, D. Pitch-based monaural segregation of reverberant speech. *J. Acoust. Soc. Am.* **120**, 458–469 (2006).
18. Kidd, G., Favrot, S., Desloge, J. G., Streeter, T. M. & Mason, C. R. Design and preliminary testing of a visually guided hearing aid. *J. Acoust. Soc. Am.* **133**, EL202–EL207 (2013).
19. Völker, C., Warzybok, A. & Ernst, S. M. Comparing binaural pre-processing strategies III: Speech intelligibility of normal-hearing and hearing-impaired listeners. *Trends Hear*. **19**, 2331216515618609 (2015).
20. Christensen, J. H. et al. Evaluating real-world benefits of hearing aids with deep neural network-based noise reduction: an ecological momentary assessment study. *Am. J. Audio.* **33**, 242–253 (2024).
21. Andersen, A. H. et al. Creating clarity in noisy environments by using deep learning in hearing aids. *Semin. Hear.* **42**, 260–281 (2021).
22. Hládek, Ľ., Porr, B., Naylor, G., Lunner, T. & Owen Brimijoin, W. On the interaction of head and gaze control with acoustic beam width of a simulated beamformer in a two-talker scenario. *Trends Hear*. **23**, 2331216519876795 (2019).
23. Favre-Félix, A., Graversen, C., Hietkamp, R. K., Dau, T. & Lunner, T. Improving speech intelligibility by hearing aid eye-gaze steering: conditions with head fixated in a multitalker environment. *Trends Hear.* **22**, 2331216518814388 (2018).
24. Kidd, G. Enhancing auditory selective attention using a visually guided hearing aid. *J. Speech Lang. Hear. Res.* **60**, 3027–3038 (2017).
25. Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J. & Durlach, N. I. in *Auditory Perception of Sound Sources* (eds. W. A. Yost, A. N. Popper & R. R. Fay) 143–189 (Springer Handbook of Auditory Research, New York, 2008).
26. Gardner, W. G. & Martin, K. D. HRTF measurements of a KEMAR. *J. Acoust. Soc. Am.* **97**, 3907–3908 (1995).
27. Schütt, H. H., Harmeling, S., Macke, J. H. & Wichmann, F. A. Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vis. Res* **122**, 105–123 (2016).
28. Wang, D. in *Speech Separation by Humans and Machines* (ed. P. Divenyi) 181–197 (Kluwer Academic, Norwell, M. A., 2005).
29. Taal, C. H., Hendriks, R. C., Heusdens, R. & Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 4214–4217 (IEEE, Dallas, TX, USA, 2010).
30. Gößling, N. & Doclo, S. RTF-steered Binaural MVDR Beamforming Incorporating an External Microphone for Dynamic Acoustic Scenarios. *in ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 416–420 (IEEE, 2019).
31. Abdelli, O. & Merazka, F. Denoising speech signal using decision directed approach. *Int. J. Intormat. Appl. Math.* **3**, 70–83 (2020).

Acknowledgements

This work was supported by grants from the National Institutes of Health (Award No. R01 DC013286), the National Science Foundation (Award No. 2319321) and the Demant Foundation. The authors would like to thank Sergi Rotger-Griful, Martin Skogland, Paol Hoang and Gerald Kidd for their input and support as well as Kenny Chou for his prior work on the BOSSA algorithm.

Author contributions

A.B. and V.B. designed the psychophysical experiment, conducted the experiment, analyzed the data, and wrote the first version of the manuscript, with editing by K.S. A.B. designed the algorithm under the supervision of K.S. All authors contributed to the article and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44172-025-00414-5.

Correspondence and requests for materials should be addressed to Alexander D. Boyd.

Peer review information *Communications Engineering* thanks Yiya Hao and Etienne Gaudrain for their contribution to the peer review of this work. Primary handling editors: [Or Perlman] and [Miranda Vinay and Rosamund Daw]. Peer reviewer reports are available.