

2022

A biologically orientated algorithm for spatial sound segregation

K. Chou, V. Best, H.S. Colburn, K. Sen. 2022. "A biologically orientated algorithm for spatial sound segregation." *Frontiers in Neuroscience*. <https://doi.org/10.3389/fnins.2022.1004071>
<https://hdl.handle.net/2144/46938>

"Downloaded from OpenBU. Boston University's institutional repository."



OPEN ACCESS

EDITED BY
Yi Zhou,
Arizona State University, United States

REVIEWED BY
Yan Gai,
Saint Louis University, United States
John C. Middlebrooks,
University of California, Irvine,
United States

*CORRESPONDENCE
Kamal Sen
kamalsen@bu.edu

SPECIALTY SECTION
This article was submitted to
Auditory Cognitive Neuroscience,
a section of the journal
Frontiers in Neuroscience

RECEIVED 26 July 2022
ACCEPTED 28 September 2022
PUBLISHED 14 October 2022

CITATION
Chou KF, Boyd AD, Best V, Colburn HS
and Sen K (2022) A biologically
oriented algorithm for spatial sound
segregation.
Front. Neurosci. 16:1004071.
doi: 10.3389/fnins.2022.1004071

COPYRIGHT
© 2022 Chou, Boyd, Best, Colburn and
Sen. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](#). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

A biologically oriented algorithm for spatial sound segregation

Kenny F. Chou¹, Alexander D. Boyd¹, Virginia Best²,
H. Steven Colburn¹ and Kamal Sen^{1*}

¹Department of Biomedical Engineering, Boston University, Boston, MA, United States, ²Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA, United States

Listening in an acoustically cluttered scene remains a difficult task for both machines and hearing-impaired listeners. Normal-hearing listeners accomplish this task with relative ease by segregating the scene into its constituent sound sources, then selecting and attending to a target source. An assistive listening device that mimics the biological mechanisms underlying this behavior may provide an effective solution for those with difficulty listening in acoustically cluttered environments (e.g., a cocktail party). Here, we present a binaural sound segregation algorithm based on a hierarchical network model of the auditory system. In the algorithm, binaural sound inputs first drive populations of neurons tuned to specific spatial locations and frequencies. The spiking response of neurons in the output layer are then reconstructed into audible waveforms *via* a novel reconstruction method. We evaluate the performance of the algorithm with a speech-on-speech intelligibility task in normal-hearing listeners. This two-microphone-input algorithm is shown to provide listeners with perceptual benefit similar to that of a 16-microphone acoustic beamformer. These results demonstrate the promise of this biologically inspired algorithm for enhancing selective listening in challenging multi-talker scenes.

KEYWORDS

multitalker speech perception, sound (audio) processing, sound segregation, cocktail party problem, binaural hearing, spatial listening, hearing loss

Introduction

Attending to a single conversation partner in the presence of multiple distracting talkers (i.e., the Cocktail Party Problem, CPP) is a complicated and difficult task for machines and humans (Haykin and Chen, 2005; McDermott, 2009; Qian et al., 2018). While some listeners can accomplish this task with relative ease, other groups of listeners report great difficulty—such as those with sensorineural hearing loss (Kochkin, 2000, 2007; Shinn-Cunningham and Best, 2008), cochlear implant users (Bernstein et al., 2016; Goupell et al., 2016, 2018; Litovsky et al., 2017), subgroups of children (Dhamani et al., 2013), persons with aphasia (Villard and Kidd, 2019) and adults with “hidden

hearing loss” (Pichora-Fuller et al., 2017; Shinn-Cunningham, 2017; Parthasarathy et al., 2019). At a cocktail party, talkers are distributed in space. Listeners use spatial cues (i.e., interaural timing and level differences, or ITDs and ILDs, respectively) for sound localization. Additionally, normal-hearing listeners appear to make use of spatial cues in addition to a variety of other talker-related cues, to perceptually segregate the competing talkers and attend to the one of most interest. Indeed, spatial listening has been shown to provide enormous benefit to listeners in cocktail-party scenarios (Litovsky, 2012; Rennie and Kidd, 2018).

Sound processing algorithms can be designed with the distinct goals of sound localization or spatial sound segregation. Specifically, spatial processing plays a key role in several sound segregation algorithms that aim to help hearing-impaired listeners overcome the CPP. For example, acoustic beamforming techniques utilize multiple microphones to selectively enhance signals from a desired direction (Gannot et al., 2017; Chiariotti et al., 2019), and are often employed in hearing aids (Greenberg and Zurek, 2001; Chung, 2004; Doclo et al., 2010; Picou et al., 2014; Launer et al., 2016). Machine learning approaches such as clustering using Gaussian mixture models (MESSL) (Mandel et al., 2010) and deep neural networks (DNN) (Wang et al., 2014), among others, also make use of ITDs and ILDs to localize the target sound.

The ability of human listeners with normal hearing to solve the CPP is quite remarkable. Many animals, too, appear to have robust solutions to their own versions of the CPP (Bee and Micheyl, 2008). Unlike beamformers, which benefit from using microphone arrays, humans and animals require only two inputs—the left and right ear. These listeners are also able to solve the CPP in novel and unpredictable settings, a challenge for algorithms that rely on supervised learning (Bentsen et al., 2018; Wang and Chen, 2018). This raises the idea that spatially selective algorithms may benefit from incorporating insights from the human and/or animal brain. From a practical standpoint, biological processing, which is based on neural spikes, also has practical advantages that make it uniquely suited for always-on, portable devices such as hearing aids. Spike-based processing is computationally efficient and can be implemented with higher temporal resolution than algorithms operating on sampled waveforms (Ghosh-Dastidar and Adeli, 2009), especially when implemented on specialized neuromorphic hardware (Roy et al., 2019).

We recently proposed a biologically inspired algorithm for sound processing. The primary goal of this algorithm was to use spatial cues to perform sound segregation and selection, not sound localization. In this algorithm, sound mixtures were segregated by spatially selective model neurons, and selection was achieved by selective integration *via* a cortical network model (Chou et al., 2019). For the tested conditions, which included a frontal target talker and two symmetrically placed masker talkers, the algorithm showed segregation performance

similar to MESSL and DNN, and provided proof-of-concept for a biologically based speech processing algorithm. However, the algorithm operated in the spiking domain, and employed a linear decoding algorithm to recover the target speech (Mesgarani et al., 2009), which resulted in low objective speech intelligibility. Like many typical beamformers, the algorithm also did not preserve binaural cues in the output, which can be particularly problematic in multitalker mixtures (Best et al., 2017; Wang et al., 2020). These drawbacks limited its practical use for applications in hearing-assistive devices and machine hearing.

In this study, we present a new biologically oriented sound segregation algorithm (BOSSA) that overcomes specific limitations of our previous algorithm. We introduce a time-frequency mask estimation method for decoding processed neural spikes that improves the quality of recovered target speech compared to the current standard approach (Mesgarani et al., 2009). We compared the proposed two-channel algorithm to a 16-microphone super-directional beamformer, using both objective measures and human psychophysics, and showed equivalent performance. Our algorithm overcomes some of the challenges faced by current state-of-the-art technologies, and provides an alternative, biologically based approach to the CPP.

Algorithm design and implementation

The proposed BOSSA algorithm contains three modules (Figure 1) that together generate neural output patterns that are inputs to the target-reconstruction stage. The first module resembles peripheral filtering by the cochlea. The second module performs spatial segregation by constructing model neurons sensitive to specific spatial cues in narrow frequency bands. Ensembles of neurons then encode sounds that share the same spatial cues. In the third module, the spiking activity of output neurons are decoded into intelligible waveforms using a novel reconstruction approach. All modules are implemented in MATLAB (MathWorks, Natick, MA, United States).

Peripheral filtering

Left and right channels of the input audio are filtered with a gammatone equivalent-rectangular-bandwidth (ERB) filterbank, implemented using the auditory toolbox in MATLAB (Slaney, 1998). The bandwidths were calculated using $ERB = [(f_c/Q)^x + b^x]^{\frac{1}{x}}$ with parameters $Q = 9.26449$ (Glasberg and Moore, 1990), minimum bandwidth (b) = 24.7 Hz, order (x) = 1. The filterbank used here has 64 channels with center frequencies ranging from $f_1 = 200$ Hz to $f_{64} = 20$ kHz. The filterbank outputs are two sets of 64

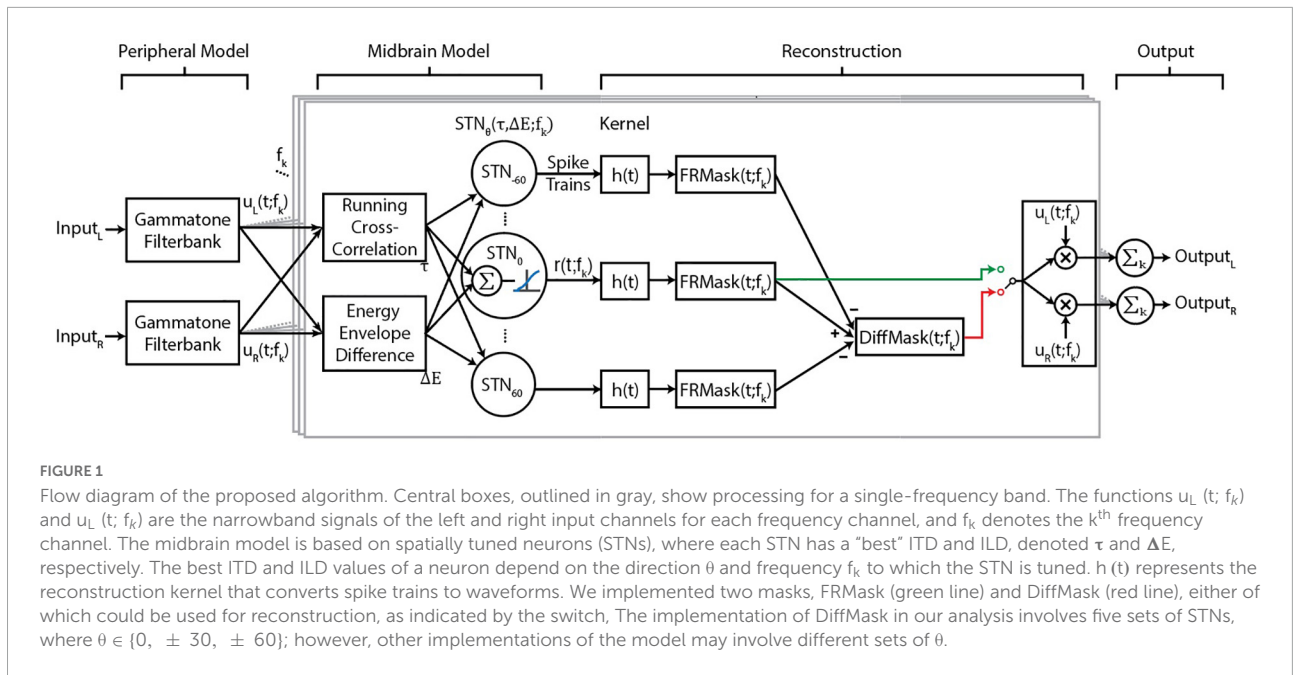


FIGURE 1
 Flow diagram of the proposed algorithm. Central boxes, outlined in gray, show processing for a single-frequency band. The functions $u_L(t; f_k)$ and $u_R(t; f_k)$ are the narrowband signals of the left and right input channels for each frequency channel, and f_k denotes the k^{th} frequency channel. The midbrain model is based on spatially tuned neurons (STNs), where each STN has a “best” 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. We implemented two masks, FRMask (green line) and DiffMask (red line), either of which could be used for reconstruction, as indicated by the switch. The implementation of DiffMask in our analysis involves five sets of STNs, where $\theta \in \{0, \pm 30, \pm 60\}$; however, other implementations of the model may involve different sets of θ .

channels of narrowband signals, $u_L(t; f_k)$ and $u_R(t; f_k)$, corresponding to the left and right channels, respectively.

Midbrain model

First, binaural cues of input signals are extracted based on a model of the barn-owl inferior colliculus (Fischer et al., 2009). ITD was calculated as a short-time running cross correlation between the energy-normalized $u_L(t; f_k)$ and $u_R(t; f_k)$ and ILD as the energy envelope difference between $u_L(t; f_k)$ and $u_R(t; f_k)$. Gain modulation steps matching those used in Fischer et al. (2009) were applied to the filterbank outputs such that the inputs to the cross correlation calculation, ($u_L(t; f_k)$ and $u_R(t; f_k)$), varied as a linear function of stimulus level. Further gain control applied during the cross correlation calculation in conjunction with a logarithmic energy envelope calculation resulted in an approximately stimulus level invariant ILD representation. For a detailed description of the mathematical operations and their physiological basis, we refer interested readers to Fischer et al. (2009).

We then constructed sets of spatially tuned neurons (STNs), where each set consists of 64 neurons tuned to f_k of the previous module. The 64 neurons in each set are sensitive to the same specific direction θ in the horizontal plane (STN_{θ} , Figure 1), and each neuron has specific parameters $\tau(\theta; f_k)$ and $\Delta E(\theta, f_k)$, corresponding to the ITD and ILD for that specific θ . Each neuron’s preferred time-lag τ was calculated using the Woodworth formulation (Woodworth, 1938), with the approximation that ITDs are independent of frequency. Preliminary studies found that using frequency-dependent ITD

values, calculated as described by Fischer et al. (2009) or the ones described by Aaronson and Hartmann (2014), provided no benefit in terms of objective measures of algorithm performance. On the other hand, ΔE is frequency-dependent, and was derived by calculating the ILD of a narrow band noise placed at various azimuths. Directionality of the narrow band noise was imparted by convolving with Head Related Transfer Functions (HRTFs) of the Knowles Electronic Manikin for Acoustic Research (KEMAR) (Burkhard and Sachs, 1975; Algazi et al., 2001).

The responses of model neurons were then calculated as follows. If the stimulus energy envelope difference was within a preset range of the neuron’s preferred ΔE , then that energy-envelope difference was weighted by the energy envelope of either $u_L(t; f_k)$ or $u_R(t; f_k)$. The ITD and ILD components were combined additively at the subthreshold level and then transformed via a sigmoidal input-output non-linearity (i.e., an activation function) to obtain an instantaneous firing rate. Finally, a Poisson spiking generator was used to generate spike trains for each neuron [$r_{\theta}(t; f_k)$, Figure 1]. This sequence of operations is expected to produce a multiplicative spiking response to ITD and ILD in each model neuron as explained in Fischer et al. (2009). These steps, including the activation function, were kept identical for all frequency channels. Parameters for the input-output nonlinearity were modified from a step-function to a sigmoidal function to increase the dynamic range of the model neurons’ firing rates.

The model can be implemented with any number and configuration of STNs. For illustrations of spatial tuning curves in Figure 2A, nine sets of STNs were constructed where $\theta \in \{0^\circ, \pm 30^\circ, \pm 45^\circ, \pm 60^\circ, \pm 90^\circ\}$. The ILDs used in generating the neuron spatial tuning curves are shown in

Figure 2B, where each line represents $\Delta E(\theta, f_k)$ for a set of STN_θ . All other results were obtained by constructing five sets of STNs, where $\theta \in \{0^\circ, \pm 30^\circ, \pm 60^\circ\}$.

Stimulus reconstruction

The stimulus reconstruction module decodes ensembles of neural spikes into audible waveforms, using an approach similar to ideal time-frequency mask estimation (Wang, 2005). The concept of time-frequency masks can be summarized as follows: for a time-frequency representation of an audio mixture (e.g., spectrogram) consisting of a target and interferers, one can evaluate each element (i.e., time-frequency tile) of such a representation and determine whether the energy present is dominated by the target or the masker. If the target sound dominates, a value of unity (1) is assigned to that time-frequency tile, and zero (0) otherwise. This process creates an ideal binary mask. Alternatively, assigning the ratio of energies of the target to total energies in a time-frequency tile yields the ideal ratio mask (Srinivasan et al., 2006). One can then estimate the target sound by applying the mask to the sound mixture *via* element-wise multiplication. This process has been shown to recover the target with high fidelity in various types of noise (Wang, 2005). A key idea to both binary and ratio masks is the application of a gain factor to each time-frequency tile of a signal. In the proposed BOSSA algorithm we adopt a similar approach but calculate the gain factor for each time-frequency tile based solely on user-defined knowledge of the target location, as explained below.

The spiking responses from the spatially tuned neurons, $r(t; f_k)$, were convolved with a kernel, $h(t)$, to calculate a smoothed, firing-rate-like measure. We set the kernel to be an alpha function: $h(t) = te^{-t/\tau_h}$, a common function involved in modeling neural dynamics. We used a value of $\tau_h = 20$ ms (see section Model Parameters) and the kernel was restricted to a length of 100 ms.

The same kernel was convolved with the spike trains of each frequency channel independently. The resulting firing rates of each set of STNs were treated as a non-binary time-frequency mask:

$$FRMask(t; f_k) = r(t; f_k) * h(t)$$

where $*$ denotes convolution. We note that the FRMask is akin to a smoothed version of the firing rate. Thus, in theory, FRMask could be directly derived from the firing rate (without the need for spikes). However, the midbrain model can be used as a front-end to spiking network models, where the calculation of spikes is necessary (Chou et al., 2019). Thus, we kept this more versatile implementation.

The mask was then applied (i.e., point-multiplied) to the left and right channels of the original sound mixture. Then, we summed (without weighting) each frequency channel of

the FRMask-filtered signal to obtain an audible, segregated waveform. We designated this result as \hat{S} .

$$\hat{S}_j = \sum_k FRMask(t; f_k) \cdot u_j(t; f_k), j \in \{L, R\}$$

This procedure resulted in a binaural signal and retained the natural spatial cues of the sound sources.

To reduce spatial leakage, we calculated a DiffMask by calculating FRMasks for each STN_θ , then subtracting scaled versions of the off-center STN_θ from STN_0 , followed by rectification:

$$DiffMask = Max(FRMask_0 - a \sum FRMask_\theta, 0)$$

where $\theta \in [\pm 30^\circ, \pm 60^\circ]$ corresponds to the location of maskers in our experimental stimuli (see section “Psychophysical Experiment”). In this operation, each FRMask was first scaled to [0,1]. The scaling factor a was chosen to be 0.5 (see section “Model Parameters”) and was fixed across all frequencies and spatial channels to reduce the amount of computational complexity in the algorithm.

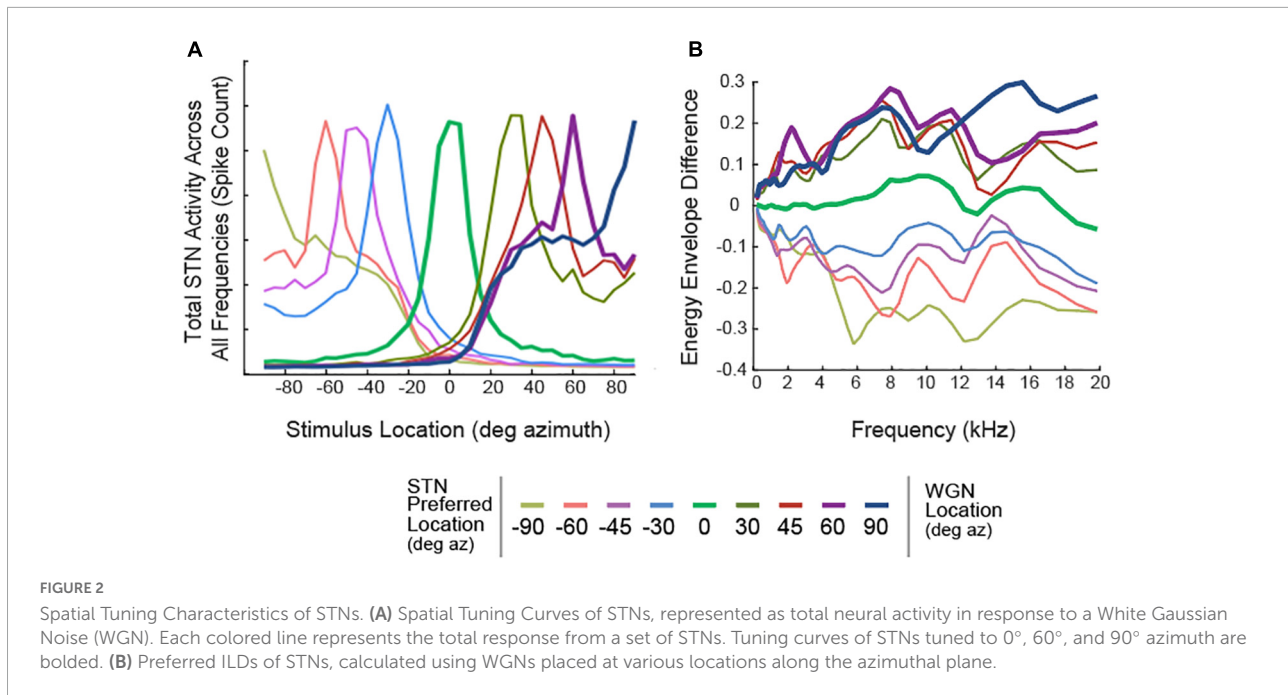
Model parameters

Although a behavioral measure of algorithm performance using human psychophysics is the gold standard, such experiments are too time-consuming to explore model parameter variations. For practical reasons, most model parameters were fixed to biologically plausible values. We explored variations in the time-constant of the alpha function kernel (τ_h), and the scaling factor for DiffMask (a). We chose the specific values of these parameters using an iterative process by trying a range of values, quantifying algorithm performance using the Short Time Objective Intelligibility (STOI) measure (Taal et al., 2010), and choosing parameters that produced the highest average STOI. STOI is an approximation of speech intelligibility, and ranges between 0 and 1. We do not claim that this approach produces an optimal set of parameters for reconstruction. However, objective measures combined with our behavioral results indicate that the parameter values we chose generated good reconstructions.

Algorithm performance

Spatial tuning characteristics

Spatial tuning responses of STNs were important predictors of the model’s segregation performance. We define “spatial tuning curves” as the spiking activity of STNs as a function of stimulus location. To construct spatial tuning curves, white Gaussian noise was convolved with anechoic KEMAR HRTFs, then presented to the algorithm. **Figure 2** shows the responses of



STNs combined across frequency channels. Ideally, STNs would only respond to stimuli from one specific direction. However, **Figure 2** shows that all STNs also respond to off-target locations. For example, STNs tuned to 0° azimuth (**Figure 2A**, green curve) respond to stimuli at $\pm 30^\circ$ azimuth and even have a non-zero response to stimuli at $\pm 90^\circ$ azimuth. We refer to this property as “spatial leakage,” which occurs due to overlap in the bandpass filters as well as the fact that a given binaural cue can occur for stimuli from multiple locations (**Figure 2B**) and thus contain some ambiguity (Brainard et al., 1992).

Spatial leakage

Leakage across spatial channels limits the performance of the algorithm, especially when multiple sound sources are present. To demonstrate, two randomly selected sentences were presented individually to the model from 0° azimuth, 90° azimuth, or simultaneously from both locations. The responses of three set of STNs, tuned to 90°, 45°, and 0°, are shown as spike-rasters in **Figure 3**. Each row within a raster plot represents the spiking response from the neuron tuned to that particular frequency channel. Due to spatial leakage, all STNs respond to the single sentence placed at 0° or 90° (**Figures 3A,B**). When both sentences are present, ITDs and ILDs interact to produce complicated STN response patterns (**Figure 3C**). Spatial leakage limits the ability of STNs to respond to a single talker, since any one spatial channel contains information from other spatial channels. Lateral inhibition was designed to address the issue of spatial leakage by suppressing neural activation by off-target sound streams.

DiffMask

The DiffMask operation was inspired by lateral inhibition observed in biological networks. This operation was applied to the spatial tuning curves of 0° STNs to illustrate its sharpening effect on spatial tuning. **Figure 4A** shows the tuning curves prior to the DiffMask operation. Some neurons within the 0° STNs were activated by stimuli from as far away as 90° (see side peaks). After the DiffMask operation, spiking activity elicited by far-away stimuli was silenced, and side-peaks were suppressed considerably (**Figure 4B**). Using a subset of STNs during the DiffMask operation, such as those tuned to $\pm 30^\circ$ (**Figure 4C**) or $\pm 60^\circ$ (**Figure 4D**), did not suppress side-peaks as effectively as if both $\pm 30^\circ$ and $\pm 60^\circ$ were used.

Psychophysical experiment

A psychophysical experiment was conducted to quantify the perceptual benefit provided by the algorithm for listeners with normal hearing. The performance of FRMask and DiffMask was compared against a 16-microphone super-directional beamformer, called BEAMAR (Kidd et al., 2015; Best et al., 2017). BEAMAR attenuates off-center sounds by combining the weighted output of 16 omni-directional microphones into a single channel, using an optimal-directivity algorithm (Stadler and Rabinowitz, 1993). BEAMAR does not process frequencies below 1 kHz in order to retain natural spatial cues in that frequency region. The combination of beamforming at high frequencies and natural binaural signals at low frequencies has

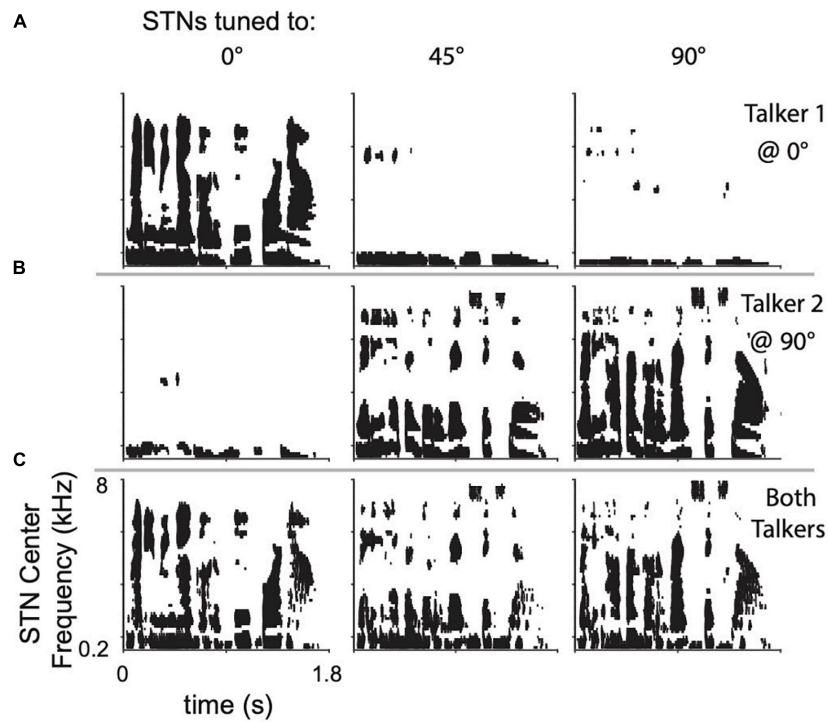


FIGURE 3
 Raster plots of STN responses to (A) top row, a single sentence placed at 0° azimuth, (B) center row, a different sentence placed at 90° azimuth, and (C) bottom row, both sentences present at their respective locations. Columns show the STN responses when tuned to the location indicated.

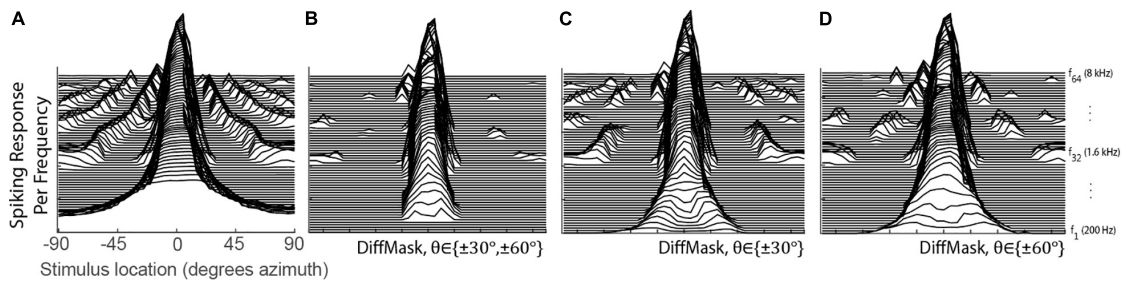


FIGURE 4
 Spatial Tuning of the 0° STNs for before (A) and after (B–D) the DiffMask operation. Each line represents the spatial tuning curve of a single frequency-specific neuron within the set of STNs ranging from 200 to 8 kHz with ERB spacing. STN of the neuron tuned to the lowest frequency is placed on the bottom of the plots. STNs involved in the DiffMask operation are denoted in each subplot.

been shown to provide a significant benefit to both normal-hearing and hearing-impaired listeners attending to a target speech sentence in a multi-talker mixture (Best et al., 2017).

Participants

Participants in this study were eleven young normal-hearing listeners, ages 18–32. All listeners had symmetrical audiogram measurements between 0.25 and 8 kHz with hearing thresholds within 20 dB HL. Participants were paid for their

participation and gave written informed consent. All procedures were approved by the Boston University Institutional Review Board (protocol 1301E).

Stimuli

Five-word sentences were constructed from a corpus of monosyllabic words (Kidd et al., 2008), 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.

Each word in each sentence was spoken by a different female talker, randomly chosen from a set of eight female talkers, without repetition. During each trial, a target sentence was mixed with four masker sentences, all constructed in the same manner. Words from the target and masker sentences were time-aligned, so that the words from each category shared the same onset. The design of these stimuli was intended to reduce the availability of voice and timing-related cues, and as such increase the listener's use of spatial information to solve the task.

The five sentences were simulated to originate from five spatial locations: 0° , $\pm 30^\circ$, and $\pm 60^\circ$ azimuth, by convolving each sentence with anechoic KEMAR HRTFs. The target sentence was always located at 0° azimuth. The four maskers were presented at 55 dB SPL from $\pm 30^\circ$, and $\pm 60^\circ$ azimuth. The level of the target was varied to achieve target-to-masker ratios (TMRs) of -5 , 0 , and 5 dB.

Stimuli were processed using one of three methods: BEAMAR, FRMask, and DiffMask. A control condition was also included, in which stimuli were spatialized using KEMAR HRTFs to convey "natural" cues but were otherwise unprocessed.

Procedures

Three blocks were presented for each of the four conditions, with each block containing five trials at each of the three TMRs (15 total trials per block). This resulted in 15 trials per TMR for each of the four processing conditions, and a total of 180 trials across all conditions. The order of presentation of TMRs within a block, and the order of blocks for each participant, were chosen at random. The experiment took approximately 1 h to complete.

Stimuli were controlled in MATLAB and presented *via* a real time processor and headphone driver (RP2.1 & HB7, Tucker Davis Technologies, Alachua, FL, United States) through a pair of headphones (Sennheiser HD265 Linear). The sound system was calibrated at the headphones with a sound 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 position of the five word sentence). For each trial, participants were presented a sentence mixture and were instructed to listen for the target sentence located directly in front. They responded with a mouse by choosing one word from each column on the grid.

Analysis

Each participant's performance was evaluated by calculating the percentage of correctly answered keywords across all trials for a given condition. Psychometric functions were generated

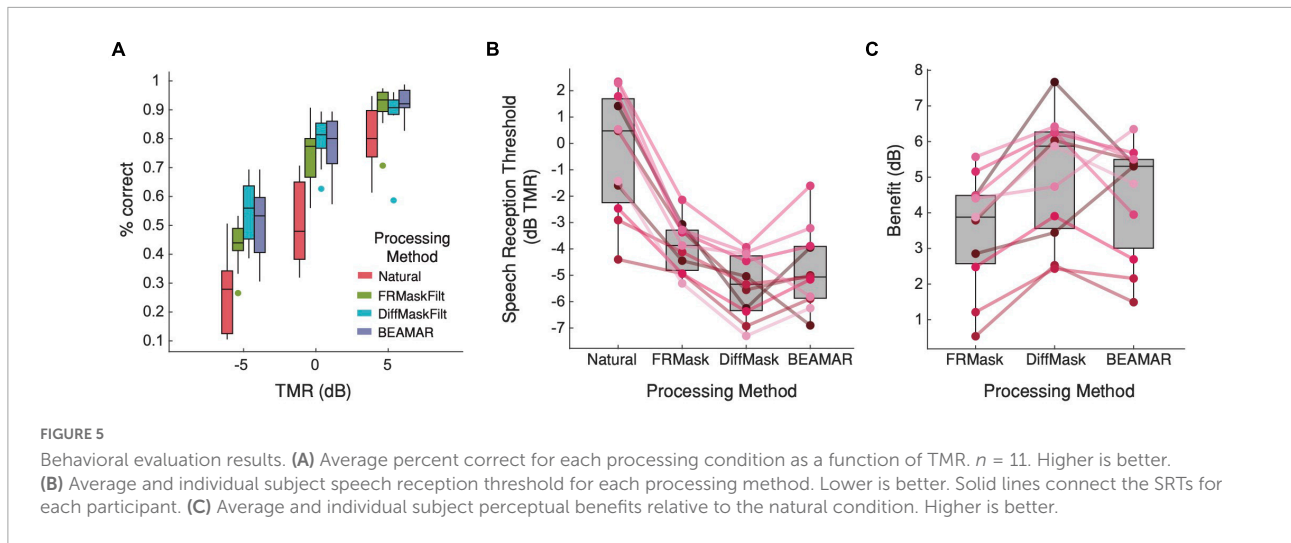
by plotting the percent correct as a function of TMR and fitting a logistic function to those data. Speech reception thresholds (SRTs), which are the TMRs corresponding to 50% correct, were extracted from each function using the `psignifit` toolbox (Schütt et al., 2016). Differences in SRTs between the natural condition and each of the processing conditions was taken to be the "benefit" provided by that processing method. Statistical analysis was done in Python using the `statsmodels` package (Seabold and Perktold, 2010).

Results

Figure 5A shows the percentage of correct responses for each TMR and processing method. A two-way repeated-measures ANOVA found a significant interaction between processing method and TMR on performance [$F_{(6,60)} = 6.97$, $p < 0.001$]. *Post hoc* pairwise comparisons using Tukey's HSD test found significant differences between the natural condition and each of the three processing methods for all three TMRs ($p < 0.001$), suggesting that subjects significantly benefitted from listening to processed speech across all TMRs. At $+5$ -dB TMR, performance was equivalent under all three processing conditions. However, at -5 -dB and 0 -dB TMR, performance was better for DiffMask than FRMask, and similar for DiffMask and BEAMAR. Figure 5B presents the same results in terms of SRTs, and Figure 5C shows the benefit (in dB) of each processing method relative to the natural condition. A one-way repeated measures ANOVA followed by Tukey's multiple pairwise comparison showed that all three algorithms provided significant benefit to listeners ($p < 0.001$). Benefits provided by BEAMAR and DiffMask were not significantly different ($p = 0.66$). Out of the eleven listeners, two achieved the lowest SRT and gained the most benefit from BEAMAR, while nine achieved the lowest SRT and gained the most benefit from DiffMask.

Discussion

Extensive research has been devoted to developing a solution for the CPP [for review, see Qian et al. (2018)], and many approaches benefit from using multiple microphones. For example, the performance of methods using independent component analysis degrades quickly as the number of sources exceeds the number of microphones (Hyvärinen et al., 2001). In acoustic beamforming, performance of the beamformer can be significantly improved by increasing the number of microphones used (Greenberg and Zurek, 2001; Greenberg et al., 2003). Although traditional beamformers produce a single-channel output, which cannot carry binaural information, a variety of spatial-cue preservation strategies have been proposed to overcome this limitation (Doclo et al., 2010;



Best et al., 2017; Wang et al., 2020). Here we demonstrated that equivalent performance to a highly optimized beamformer (such as BEAMAR) may be possible using a biologically inspired algorithm that uses only two microphones placed in the ears. Our biologically oriented sound segregation (BOSSA) model provided a substantial benefit in a challenging cocktail party listening situation, and this benefit was larger than that provided by BEAMAR in the majority of our young, normal hearing participants. While this is a promising result, further work is needed to examine the benefits of BOSSA under a wider variety of scenarios and in other groups of listeners. Comparisons to other two-microphone solutions such as binaural beamformers (Doclo et al., 2010; Best et al., 2015), as well as deep-learning solutions that operate on two or even a single microphone (Roman et al., 2003; Healy et al., 2013), would also be interesting.

Spiking neural networks traditionally do not have applications in audio processing due to the lack of a method that produces intelligible, high-quality reconstructions. The “optimal prior” method of reconstruction is often used to obtain reconstructions from physiologically recorded neural responses (Bialek et al., 1991; Stanley et al., 1999; Mesgarani et al., 2009; Mesgarani and Chang, 2012), but produces single-audio-channel reconstructions of poor quality and intelligibility (Chou et al., 2019). The optimal prior method computes a linear filter between a training stimulus and the response of neuron ensembles, and filter needs to be re-trained if the underlying network changes. In contrast, the mask-based reconstruction method used in this study estimates time-frequency masks from spike trains. It is able to obtain reconstructions with much higher intelligibility and preserves spatial cues, all without the need for training. These properties enable rapid development of spiking neural network models for audio-related applications.

Within the biologically plausible algorithms we tested, the difference in performance between FRMask and DiffMask is noteworthy and interesting. The spatial tuning plots (Figure 2)

quantify the tuning of a given spatial channel to a single sound as it is moved around the lateral spatial field which are reasonably well-tuned. Moreover, Figures 3A,B, for example, illustrate the response of the 0° channel to sounds presented at 0° and 90° . In this case, the 0° channel responded primarily to the frontal sound. By themselves, these plots do not suggest problems with spatial tuning and leakage. However, in our psychophysical experiments, we presented a target sound at 0° with four competing maskers from $\pm 30^\circ$ and $\pm 60^\circ$, a far more challenging scenario. In such a scenario, spatial leakage is more significant, and refining/improving spatial tuning improves sound segregation, as demonstrated in the improvement with DiffMask over FRMask.

It is also worth noting that our algorithms were based on processing in the barn owl midbrain which contains a topographic map of space, whereas, in mammals, no such topographic map has been found. Despite this difference, the spatially tuned responses of neurons in the model could be leveraged to improve speech segregation performance in humans. This demonstrates that brain inspired algorithms based on non-human model systems can improve human perception and performance.

The work presented here represents a preliminary evaluation of the BOSSA model, and it identified a number of issues and limitations that deserve further investigation. While the formulation of DiffMask can sharpen the spatial tuning of the STNs, neurons tuned to frequencies below 300 Hz were completely silenced for the stimuli we tested (Figure 4B). Low spatial acuity in this frequency range results in a similar response at on and off target STNs. The off-target response scaling and summation that forms DiffMask then results in a complete subtraction of on-target activity below 300 Hz. Additionally, some side peaks still persist even after the DiffMask operation, implying that spatial leakage was not fully addressed. Different formulations of the DiffMask may address these shortcomings.

Moreover, our DiffMask implementation used a specific number of off-target STNs at specific locations, which were aligned with the locations of makers in our experimental stimuli. Further works is needed to explore how DiffMask can be optimized to support arbitrary target and masker configurations, and how the resolution of the STNs affects model performance. We have avoided using deep-learning approaches in this study in favor of biological interpretability, but such approaches may help guide the optimization of DiffMask and could be very valuable in that respect. Another potential limitation of the algorithm is that it processes each frequency channel independently. While this design choice reduces both the complexity of the algorithm and its computation time, it excludes the possibility for across-frequency processing that could improve performance (Krishnan et al., 2014; Szabó et al., 2016). Finally, animals have been observed to resolve binaural cue ambiguity by having neurons preferentially tune to more reliable spatial cues in different frequency regions (Cazettes et al., 2014). Inspiration could be taken from these observations to improve spatial tuning and overcome spatial leakage. Again, deep-learning based optimization methods may help identify these reliable cues for human listeners and multitalker mixtures.

Future work with the BOSSA model could include both sound segregation and localization by comparing the response of each spatial tuning curve to predict source azimuth, possibly utilizing a denser array of STNs. Another idea we plan to explore in the future is to apply automatic speech recognition systems to optimize the parameters of the algorithm. This optimization can be performed relatively fast before conducting time-consuming psychophysics experiments. During this optimization process we also plan to investigate the effects of varying sound pressure level and source dynamics on BOSSA performance.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Boston University Institutional Review

Board. The patients/participants provided their written informed consent to participate in this study.

Author contributions

KC designed the algorithm under the supervision of KS and HC, conducted the experiment, analyzed the data, and wrote the first version of the manuscript, with editing by AB, VB, HC, and KS. KC and VB designed the psychophysical experiment. AB assisted with additional data analysis. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by R01 DC000100-42 (HC-PI, KS-Co-I) 12/01/2015–7/31/2020 NIH/NIDCD, Title: Binaural Hearing. VB and AB were supported in part by NIH/NIDCD R01 DC015760.

Acknowledgments

The authors would like to thank Matthew Goupell for reviewing an earlier version of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Aaronson, N. L., and Hartmann, W. M. (2014). Testing, correcting, and extending the Woodworth model for interaural time difference. *J. Acoust. Soc. Am.* 135, 817–823. doi: 10.1121/1.4861243
- Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001). "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the*

Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575). (Piscataway, NJ: IEEE), 99–102.

Bee, M. A., and Micheyl, C. (2008). The cocktail party problem: What is it? How can it be solved? And why should animal behaviorists study it? *J. Comp. Psych.* 122, 235–251. doi: 10.1037/0735-7036.122.3.235

- Bentsen, T., May, T., Kressner, A. A., and Dau, T. (2018). The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility. *PLoS One* 13:e0196924. doi: 10.1371/journal.pone.0196924
- Bernstein, J. G. W., Goupell, M. J., Schuchman, G. I., Rivera, A. L., and Brungart, D. S. (2016). Having two ears facilitates the perceptual separation of concurrent talkers for bilateral and single-sided deaf cochlear implantees. *Ear Hear.* 37, 289–302. doi: 10.1097/AUD.0000000000000284
- Best, V., Mejia, J., Freeston, K., van Hoesel, R. J., and Dillon, H. (2015). An evaluation of the performance of two binaural beamformers in complex and dynamic multitalker environments. *Int. J. Audiol.* 54, 727–735. doi: 10.3109/14992027.2015.1059502
- Best, V., Roverud, E., Mason, C. R., and Kidd, G. Jr. (2017). Examination of a hybrid beamformer that preserves auditory spatial cues. *J. Acoust. Soc. Am.* 142, EL369–EL374. doi: 10.1121/1.5007279
- Bialek, W., Rieke, F., de Ruyter van Steveninck, R., and Warland, D. (1991). Reading a neural code. *Science* 252, 1854–1857. doi: 10.1126/science.2063199
- Brainard, M. S., Knudsen, E. I., and Esterly, S. D. (1992). Neural derivation of sound source location: Resolution of spatial ambiguities in binaural cues. *J. Acoust. Soc. Am.* 91, 1015–1027. doi: 10.1121/1.402627
- Burkhard, M. D., and Sachs, R. M. (1975). Anthropometric manikin for acoustic research. *J. Acoust. Soc. Am.* 58, 214–222. doi: 10.1121/1.380648
- Cazettes, F., Fischer, B. J., and Pena, J. L. (2014). Spatial cue reliability drives frequency tuning in the barn Owl's midbrain. *Elife* 3:e04854. doi: 10.7554/eLife.04854
- Chiariotti, P., Martarelli, M., and Castellini, P. (2019). Acoustic beamforming for noise source localization – Reviews, methodology and applications. *Mech. Syst. Signal. Process.* 120, 422–448. doi: 10.1016/j.ymssp.2018.09.019
- Chou, K. F., Dong, J., Colburn, H. S., and Sen, K. (2019). A physiologically inspired model for solving the cocktail party problem. *J. Assoc. Res. Otolaryngol.* 20, 579–593. doi: 10.1007/s10162-019-00732-4
- Chung, K. (2004). Challenges and recent developments in hearing aids: Part I. speech understanding in noise, microphone technologies and noise reduction algorithms. *Trends Amplif.* 8, 83–124. doi: 10.1177/108471380400800302
- Dhamani, I., Leung, J., Carlile, S., and Sharma, M. (2013). Switch attention to listen. *Sci Rep* 3:1297. doi: 10.1038/srep01297
- Doclo, S., Gannot, S., Moonen, M., and Spriet, A. (2010). “Acoustic beamforming for hearing aid applications,” in *Handbook on Array Processing and Sensor Networks*, eds S. Haykin and K. J. R. Liu (Hoboken, NJ: John Wiley and Sons), 269–302. doi: 10.1002/9780470487068.ch9
- Fischer, B. J., Anderson, C. H., and Peña, J. L. (2009). Multiplicative auditory spatial receptive fields created by a hierarchy of population codes. *PLoS One* 4:e8015. doi: 10.1371/journal.pone.0008015
- Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (2017). A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* 25, 692–730. doi: 10.1109/TASLP.2016.2647702
- Ghosh-Dastidar, S., and Adeli, H. (2009). Spiking neural networks. *Int. J. Neural Syst.* 19, 295–308. doi: 10.1142/S0129065709002002
- Glasberg, B. R., and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear Res.* 47, 103–138. doi: 10.1016/0378-5955(90)90170-T
- Goupell, M. J., Kan, A., and Litovsky, R. Y. (2016). Spatial attention in bilateral cochlear-implant users. *J. Acoust. Soc. Am.* 140, 1652–1662. doi: 10.1121/1.4962378
- Goupell, M. J., Stakhovskaya, O. A., and Bernstein, J. G. W. (2018). Contralateral interference caused by binaurally presented competing speech in adult bilateral cochlear-implant users. *Ear Hear.* 39, 110–123. doi: 10.1097/AUD.0000000000000470
- Greenberg, J. E., and Zurek, P. M. (2001). “Microphone-array hearing aids,” in *Microphone Arrays*, eds M. Brandstein and D. Ward (Berlin: Springer), 229–253. doi: 10.1007/978-3-662-04619-7_11
- Greenberg, J. E., Desloge, J. G., and Zurek, P. M. (2003). Evaluation of array-processing algorithms for a headband hearing aid. *J. Acoust. Soc. Am.* 113:1646. doi: 10.1121/1.1536624
- Haykin, S., and Chen, Z. (2005). The cocktail party problem. *Neural Comput.* 17, 1875–1902. doi: 10.1162/0899766054322964
- Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners. *J. Acoust. Soc. Am.* 134, 3029–3038. doi: 10.1121/1.4820893
- Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. New York, NY: Wiley. doi: 10.1002/0471221317
- Kidd, G., 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. doi: 10.1121/1.2998980
- Kidd, G., Mason, C. R., Best, V., and Swaminathan, J. (2015). Benefits of acoustic beamforming for solving the cocktail party problem. *Trends Hear.* 19:233121651559338. doi: 10.1177/2331216515593385
- Kochkin, S. (2000). MarkeTrak V: “Why my hearing aids are in the drawer”: The consumers’ perspective. *Hear. J.* 53, 34–41. doi: 10.1097/00025572-200002000-00004
- Kochkin, S. (2007). MarkeTrak VII: Obstacles to adult non-user adoption of hearing aids. *Hear. J.* 60, 24–51. doi: 10.1097/01.HJ.0000285745.08599.7f
- Krishnan, L., Elhilali, M., and Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLoS Comput. Biol.* 10:e1003985. doi: 10.1371/journal.pcbi.1003985
- Launer, S., Zakis, J. A., and Moore, B. C. J. (2016). “Hearing aid signal processing,” in *Hearing Aids*, 1st Edn, eds G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper (Berlin: Springer International Publishing), 93–130. doi: 10.1007/978-3-319-33036-5_4
- Litovsky, R. Y. (2012). Spatial release from masking. *Acoust. Today* 8:18. doi: 10.1121/1.4729575
- Litovsky, R. Y., Goupell, M. J., Misurelli, S. M., and Kan, A. (2017). “Hearing with cochlear implants and hearing aids in complex auditory scenes,” in *The Auditory System at the Cocktail Party. Springer Handbook of Auditory Research*, Vol. 60, eds J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay (Cham: Springer), 261–291. doi: 10.1007/978-3-319-51662-2_10
- Mandel, M. I., Weiss, R. J., and Ellis, D. P. W. (2010). Model-based expectation maximization source separation and localization. *IEEE Trans. Audio Speech Lang. Process.* 18, 382–394. doi: 10.1109/TASL.2009.2029711
- McDermott, J. H. (2009). The cocktail party problem. *Curr. Biol.* 19, R1024–R1027. doi: 10.1016/j.cub.2009.09.005
- Mesgarani, N., and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–236. doi: 10.1038/nature11020
- Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *J. Neurophysiol.* 102, 3329–3339. doi: 10.1152/jn.91128.2008
- Parthasarathy, A., Hancock, K. E., Bennett, K., DeGruttola, V., and Polley, D. B. (2019). Neural signatures of disordered multi-talker speech perception in adults with normal hearing. *bioRxiv [Preprint]* doi: 10.1101/744813
- Pichora-Fuller, M. K., Alain, C., and Schneider, B. A. (2017). “Older adults at the cocktail party,” in *The Auditory System at the Cocktail Party*, eds J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay (Cham: Springer), 227–259. doi: 10.1007/978-3-319-51662-2_9
- 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. doi: 10.1097/AUD.0000000000000004
- Qian, Y. M., Weng, C., Chang, X., Wang, S., and Yu, D. (2018). Past review, current progress, and challenges ahead on the cocktail party problem. *Front. Inf. Technol. Electron. Eng.* 19:40–63. doi: 10.1631/FITEE.1700814
- Rennies, J., and Kidd, G. (2018). Benefit of binaural listening as revealed by speech intelligibility and listening effort. *J. Acoust. Soc. Am.* 144, 2147–2159. doi: 10.1121/1.5057114
- Roman, N., Wang, D., and Brown, G. J. (2003). Speech segregation based on sound localization. *J. Acoust. Soc. Am.* 114, 2236–2252. doi: 10.1121/1.1610463
- Roy, K., Jaiswal, A., and Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. *Nature* 575, 607–617. doi: 10.1038/s41586-019-1677-2
- Schütt, H. H., Harmeling, S., Macke, J. H., and Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vis. Res.* 122, 105–123. doi: 10.1016/j.visres.2016.02.002
- Seabold, S., and Perktold, J. (2010). “Statsmodels: Econometric and statistical modeling with Python,” in *Proceedings of the 9th Python in Science Conference*. (Austin, TX: SciPy Society), 57. doi: 10.25080/Majora-92bf1922-011
- Shinn-Cunningham, B. (2017). Cortical and sensory causes of individual differences in selective attention ability among listeners with normal hearing thresholds. *J. Speech Lang. Hear. Res.* 60, 2976–2988. doi: 10.1044/2017_JSLHR-H-17-0080

- Shinn-Cunningham, B. G., and Best, V. (2008). Selective attention in normal and impaired hearing. *Trends Amplif.* 12, 283–299. doi: 10.1177/1084713808325306
- Slaney, M. (1998). Auditory toolbox: A Matlab toolbox for auditory modeling work. *Interval Res. Corp Tech. Rep.* 10:1998.
- Srinivasan, S., Roman, N., and Wang, D. L. (2006). Binary and ratio time-frequency masks for robust speech recognition. *Speech Commun.* 48, 1486–1501. doi: 10.1016/j.specom.2006.09.003
- Stadler, R. W., and Rabinowitz, W. M. (1993). On the potential of fixed arrays for hearing aids. *J. Acoust. Soc. Am.* 94, 1332–1342. doi: 10.1121/1.408161
- Stanley, G. B., Li, F. F., and Dan, Y. (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J. Neurosci.* 19, 8036–8042. doi: 10.1523/JNEUROSCI.19-18-08036.1999
- Szabó, B. T., Denham, S. L., and Winkler, I. (2016). Computational models of auditory scene analysis: A review. *Front. Neurosci.* 10:524. doi: 10.3389/fnins.2016.00524
- Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in 2010 *IEEE International Conference on Acoustics, Speech and Signal Processing*. (Dallas, TX: IEEE), 4214–4217. doi: 10.1109/ICASSP.2010.5495701
- Villard, S., and Kidd, G. Jr. (2019). Effects of acquired aphasia on the recognition of speech under energetic and informational masking conditions. *Trends Hear.* 23:2331216519884480. doi: 10.1177/2331216519884480
- Wang, D. (2005). “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*. (Boston, MA: Kluwer Academic Publishers), 181–197. doi: 10.1007/0-387-22794-6_12
- Wang, D., and Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* 26, 1702–1726. doi: 10.1109/TASLP.2018.2842159
- Wang, L., Best, V., and Shinn-Cunningham, B. G. (2020). Benefits of beamforming with local spatial-cue preservation for speech localization and segregation. *Trends Hear.* 24:233121651989690. doi: 10.1177/2331216519896908
- Wang, Y., Narayanan, A., and Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Trans. Speech Lang. Process.* 22, 1849–1858. doi: 10.1109/TASLP.2014.2352935
- Woodworth, R. S. (1938). *Experimental Psychology*. New York, NY: Holt.