IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 11, 2021, accepted September 9, 2021, date of publication September 16, 2021,
date of current version September 29, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3113615*

# Cluster Analysis for the Separation of Auditory Scenes

**MATTHEW S. DALEY, LIA M. BONACCI, DAVID H. GEVER, (Member, IEEE), KRYSTINA DIAZ, AND JEFFREY B. BOLKHOVSKY**

Naval Submarine Medical Research Laboratory, Groton, CT 06340, USA

Corresponding author: Matthew S. Daley (matthew.s.daley.ctr@mail.mil)

This work was supported by Unit F1103—Office of Naval Research under Code 34.

**ABSTRACT** The ''cocktail party problem'' refers to the ability of human listeners to separate the acoustic signal reaching their ears into its individual components, corresponding to individual sound sources in the environment. Despite this phenomenon appearing trivial for humans, solving the cocktail party problem computationally remains an ambitious challenge. The approach used in this paper takes inspiration from human strategies for separating an acoustic environment into distinct perceptual auditory streams. A series of time-frequency-based features, analogous to those thought to emerge at various stages in the human auditory processing pathway, are derived from biaural auditory inputs. These feature vectors are used as inputs to an unsupervised cluster analysis used to group feature values that are assumed to correspond to the same object. Reconstructed auditory streams are then correlated to the original components used to create the auditory scene. Our model is capable of reconstructing streams that correlate to the original components ($r = 0.3$-$0.7$) used to create the complex auditory scene. The success of the reconstructions is largely dependent on the signal-to-noise ratio of the components of the auditory scene.

**INDEX TERMS** Biomedical signal processing, clustering algorithms, machine learning, machine learning algorithms, pattern clustering, signal processing algorithms.

## I. INTRODUCTION

In everyday listening environments, we are often challenged with separating a myriad of sound signals that arrives at our ears into distinct sound sources. This phenomenon, called auditory stream segregation, allows humans to focus on a single voice or sound source from within a noisy environment. The manner in which we are able to perform such a separation is commonly referred to as the cocktail party problem [1], [2], aptly named after the capacity to carry on a conversation with another individual during a noisy cocktail party. Whereas this task seems intuitive and effortless for most human listeners, perfect computational replication of this effect has remained elusive and is the focus of the field of computational auditory scene analysis (CASA) [10].

To address this challenge, multidisciplinary efforts across the fields of audiology, engineering, neuroscience, and psychology have resulted in the development of a two-stage methodology for auditory stream segregation. The strategy begins with processing the acoustic input in a feature analysis stage, which breaks down an auditory waveform into descriptive features. The descriptive features are then input into a cluster analysis stage [3]–[8] that seeks to describe the values of the features with a limited number of distribution sets. Whereas the two-step process is a well-established practice for stream segregation, the specifics of how the features and clusters are defined and analyzed differs considerably across disciplines. The feature analysis stage is dominated by the application of signal processing methods to obtain features used to distinguish auditory objects. In biologically inspired models, the features extracted from the source signal have a physiological and/or psychological basis for the formation of auditory streams [9]. Such features may be based on frequency selectivity of the basilar membrane of the cochlea [10], the spectrotemporal receptive fields (STRFs) of the auditory cortex [4], [11], [12], measurements of pitch and timbre [13], and localization via measuring the interaural time difference (ITD) [14]. These features are fed into a cluster analysis stage that attempts to create groupings of similar features within the space; the goal is to form groups that correspond to individual auditory objects. This clustering has been performed via linear combination of feature

The associate editor coordinating the review of this manuscript and approving it for publication was Yue Ivan Wu.

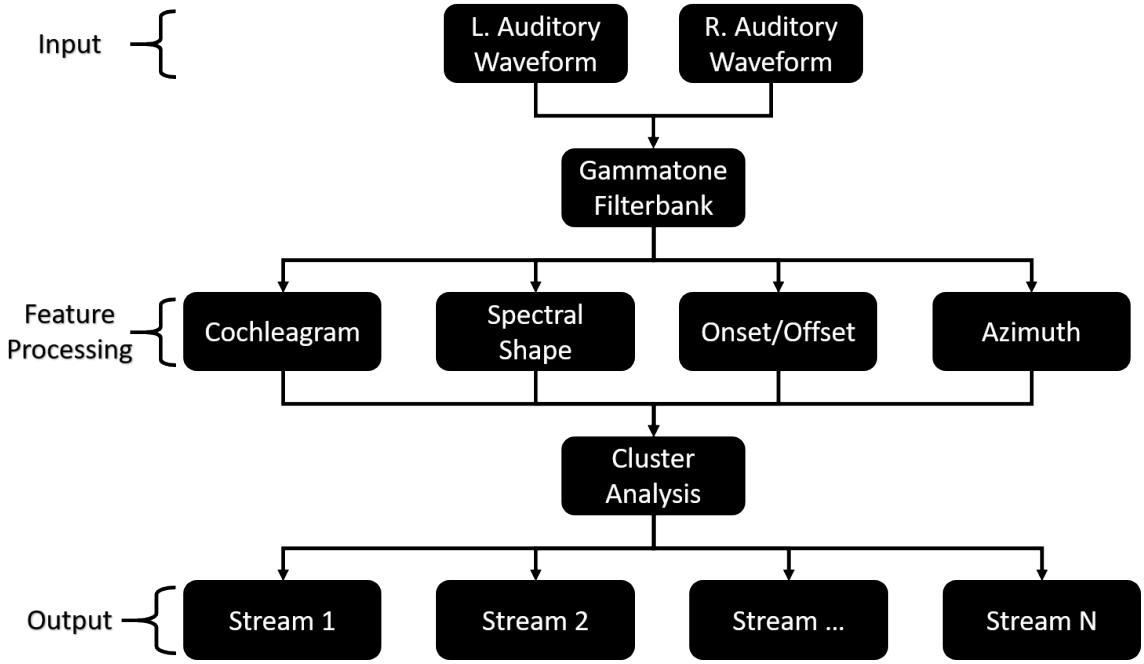


**FIGURE 1.** Model Outline. Beginning with left and right auditory waveforms, the inputs are filtered through a gammatone filterbank to obtain a time-frequency representation that is then processed into a variety of features. These features are input into a cluster analysis stage that results in the segregated auditory streams that comprise the auditory scene.

spaces [5], [7], Kalman-filter estimation [6], and artificial neural networks [8], [15].

Many approaches tend to focus on a single aspect of scene analysis, for example, using a single type of feature (e.g., STRFs, spectrotemporal contrasts) as the input for the cluster analysis stage and utilize some form of supervised clustering method to group separate auditory streams within a scene. In this way, current computational models are fragmented in their approach to mimicking the human auditory pathway [16]. The approach used in this study is novel in that we extract many different types of psychoacoustic inspired features and sort them into clusters using unsupervised statistical clustering methods such as Gaussian Mixture Models or K-Means, which seek to separate clusters based on feature coherence. We base this on the assumption that, in humans, different features of an acoustic signal are computed at various stages along the auditory pathway and that coherence between these features is a fundamental basis for stream segregation. We hypothesize that the proposed model will be capable of segregating auditory streams from noisy backgrounds using unsupervised cluster analysis on features thought to be analogous to information used by the auditory pathway.

## II. METHODS

An overview of our stream segregation procedure is described in detail throughout the methods section and is illustrated in Fig. 1. In short, a Gammatone filter bank is used to create a time-frequency approximation of the cochlear response to the auditory stimulus. From the time-frequency representations various features, analogous to those which are thought to emerge at different stages of auditory processing in the brain, are extracted via signal processing methods. Lastly, these features are used as inputs in an unsupervised cluster analysis stage that sorts the auditory scene into individual components. The result is N output auditory objects, where N must be defined a priori.

### A. GAMMATONE FILTERBANK

The model input is a binaural set of waveforms, corresponding to left and right auditory inputs. The left and right waveforms are each filtered through a fourth order gammatone filterbank to model the impulse response function of auditory nerve fibers [17]. For this implementation, a gammatone filter is calculated as follows:

$$g\left(f_0, t\right) = t^{N_f - 1} e^{-2\pi * b(f_0)} \cos\left(2\pi f_0 + \phi\right) u\left(t\right) \tag{1}$$

where $t$ is time, $N_f$ is the filter order, $f_0$ is the filter center frequency in Hz, $\phi$ is the phase, and $u(t)$ is the unit step function. The quantity $b(f_0)$ is obtained by (3) and determines the bandwidth for a given center frequency, based on its equivalent rectangular bandwidth, $ERB(f_0)$, value which is obtained by (2) [10].

$$ERB\left(f_0\right) = 21.4 \log_{10}\left(0.00437 f_0 + 1\right) \tag{2}$$

$$b\left(f_0\right) = 1.019\, ERB\left(f_0\right) \tag{3}$$

The Gammatone filterbank output, $cf$, $t$, is obtained by convolving, with respect to time $*t$, a given input, $x(t)$, with $g(f0, t)$, as shown by (4), where $f0$ is the center frequency of the gammatone filter. For this work, 64 center frequencies with equal distances on the Equivalent Rectangular Bandwidth (ERB) scale were selected in the range of 40 Hz

to 16 kHz, to lie within the human hearing range (20 Hz to 20 kHz; [2]). Digital implementation of this filterbank was completed following the procedure proposed by Patterson *et al.* (1992), and Slaney (1993).

$$c\left(f_0, t\right) = x\left(t\right) *_t g\left(f_0, t\right) \tag{4}$$

### B. FEATURE PROCESSING

Gammatone filtering results in a 2D matrix, c(f,t) with dimensions corresponding to the number of time points and ERB channels—for each of the two auditory input channels. These outputs are processed in the subsequent stage to extract specific features such as a cochleagram or azimuth. The features chosen are those commonly cited as important for stream segregation in auditory perception [6], [8], [18], [19]. For each feature, a map of the input source is created, to provide a time-frequency representation of that feature. The following section details the process of computing each of these feature maps.

#### 1) COCHLEAGRAM

The cochleagram is a time-frequency representation of the auditory signal, similar to a spectrogram, which reflects the tonotopic activation of the basilar membrane in response to sound. To compute the cochleagram, we rectify the output of the gammatone filterbank, c(f,t) (5), convert it to decibels for estimation of the auditory nerve response [20]–[22], and downsample such that $\Delta t = 10ms$ [5].

$$c_{rect}\left(f_0, t\right) = \begin{cases} 0 & if\ c\left(f_0, t\right) < 0 \\ c\left(f_0, t\right) & if\ c\left(f_0, t\right) \geq 0 \end{cases} \tag{5}$$

$$C\left(f_0, t_0\right) = 20\log\left\{\sqrt{\frac{1}{W}\sum\nolimits_{k=t_0-\Delta t}^{t_0+\Delta t} c_{rect}\left(f_0, k\right)^2}\right\} \tag{6}$$

where $f_0$ is a given center frequency of the two-dimmatrix, $t_0$ is a given time point in the cochlegram such that $t_1 - t_0 = \Delta t$, and $W$ is the number of samples in the window.

Evidence indicates that auditory attention influences perception of simultaneously occurring auditory events [2], [25]. One way this has been shown to occur is through attenuation of center frequencies outside the current focus of attention. In particular, center frequencies of focused auditory events, and frequencies within their critical band are unaffected, while center frequencies outside this band are increasingly attenuated with distance from the focal center frequency by as much as 15 dB [23]–[26]. To account for this attenuation due to exogenous draw of attention, we define the focused center frequency $f_{peak}$ of time point $t_0$ as the frequency with the greatest magnitude in the cochleagram. We then add a Gaussian window ($G_{atten}\left(f_{peak}, f\right)$) with width $\sigma = 3ERB$ with a maximum value of 0 dB and minimum value of −15 dB centered on $f_{peak}$ to the spectral dimension of the cochleagram at $t_0$ to obtain an attenuated version of the cochleagram: $C_{atten}\left(f, t_0\right)$.

$$G_{atten}\left(f_{peak}, f\right) = 15e^{\frac{1}{2}\left(\frac{f-f_{peak}}{\sigma}\right)^2} - 15 \tag{7}$$

$$C_{atten}\left(f, t_0\right) = C\left(f, t_0\right) + G_{atten}\left(f_{peak}, f\right) \tag{8}$$

#### 2) SPECTRAL SHAPE ANALYSIS

Next, we applied spectral shape filtering, or "scale" filtering, onto the cochleagram across the frequency axis for each time point, using narrow and broad bandwidths [4], [6], [8]. This analysis is based on the premise that spectral shape is an effective physical description of timbre [6]. In our model, a second derivative Gaussian function (9), that has been scaled for the desired bandwidth $\varphi$ (10), is convolved across the spectral dimension of the cochleagram at each time point, $t_0$ (11).

$$h_{bw}\left(f\right) = \left(1 - f^2\right)\exp\left(\frac{f^2}{2}\right) \tag{9}$$

$$H_{bw}\left(f, \varphi\right) = \frac{1}{\varphi} hilbert\left(h_{bw}\left(\frac{1}{\varphi} f\right)\right) \tag{10}$$

$$C_{bw}\left(f, t_0, \varphi\right) = C_{atten}\left(f, t_0\right) *_f H_{bw}\left(f, \varphi\right) \tag{11}$$

In (9-11), $f$ is the frequency along the tonotopic axis in units of *ERB* scale. For the analyses presented here, bandwidths of 1/2 and 2*ERBs* were used for a total of two individual bandwidth feature maps. Lastly, for each of these bandwidth feature maps, we subtracted the map of all broader bandwidth maps. This action is performed because events present in the broad bandwidth maps will have some amount of bleed over into the narrower bandwidth maps; subtracting broader maps from narrower maps prevents this redundancy.

Lastly, bandwidth masks are modified with bandwidth intensity masks. Bandwidth intensity masks are driven by the idea that objects present within an auditory scene will be characterized by some measurable cohesiveness in the spectral domain, and that a binary mask can be made to exclude any events that cannot be characterized by a bandwidth filter. This is computed by summing the results of the spectral analysis across the bandwidth ($\boldsymbol{\varphi}$) dimension and equating any value above a given threshold, $\boldsymbol{\tau}$, to 1, and all others zero (10). These masks are then multiplied by each of the following feature maps to mitigate interference from noisy events.

$$M\left(f, t\right)\begin{cases} = 1, & where \sum\limits_{k=\varphi_1}^{\varphi_n} C_{bw}(f, t, k) > \tau \\ = 0, & otherwise \end{cases} \tag{12}$$

#### 3) ONSET/OFFSET DETECTION

The onset of an auditory event is characterized by a general increase in intensity, whereas a general decrease in intensity characterizes offset. Onset and offset detection attempts to generate auditory segments by matching corresponding onset and offset fronts [10], [18], [27]. From a computational perspective, this process is similar to image segmentation, which seeks to identify boundaries of visual objects, where the "image" is a cochleagram of the original auditory signal.

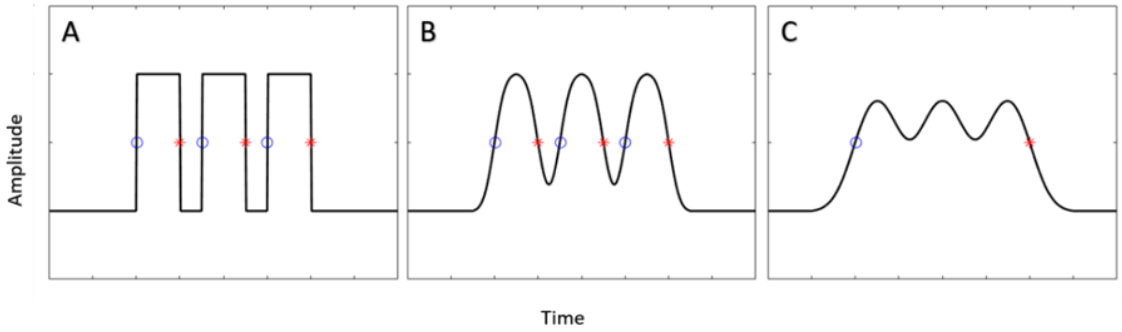


**FIGURE 2.** Effects of gaussian convolution on a single channel of the cochleagram for onsets (blue circles) and offsets (red asterisks) detection. A) No convolution. B) Gaussian window = 500ms. C) Gaussian window = 1000ms.

To find these sudden changes in intensity, we first compute the first-order derivative of the power of the cochleagram with respect to time and then identify the local maximums and minimums of this derivative. Due to background noise, however, there are many extrema values that do not correspond to actual event onsets/offsets. To mitigate this effect, we low-pass filter the signal via convolution with a Gaussian kernel (13 & 14) before calculating the first order derivative and identifying local extrema.

$$G(t,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{13}$$

$$C_{blur} = C(f,t) *_t G(t,\boldsymbol{\sigma}) \tag{14}$$

In (13-14), $\boldsymbol{\sigma}$ is the width of the Gaussian kernel, and $*_t$ denotes convolution in the time domain. The width of the Gaussian kernel has considerable effect on the performance of the peak detection algorithm used to find extrema of the first-order derivative. In particular, convolution with a very wide Gaussian kernel is effective for detecting large event onsets/offsets but not for detecting many smaller events. Similarly, a very narrow kernel is effective for detecting many small events but will miss larger events that should be grouped (see Fig. 2 for example). Because it is difficult to accurately identify events at a single Gaussian width, it is useful to detect onset/offset at multiple resolutions and incorporate some form of multiscale integration [18]. For the computation of this feature space, a set of Gaussian widths ranging from 60 to 1200 ms was used prior to peak detection, and the median onset/offset event value of each time-frequency point throughout all resolutions was used as the final onset/offset event value.

#### 4) AZIMUTHAL LOCATION

The percept of a sound's azimuthal location has been shown to arise through computation of interaural time differences (ITDs) [14], [28], [29], that is, the location of a sound is based on the difference in its arrival time between two ears. To compute the ITD for a given time point, $\boldsymbol{t_0}$, and center frequency, $\boldsymbol{f_0}$, we first cross-correlate the gammatone-filtered auditory signals of the right and left channel inputs, $\boldsymbol{c_R(f_0,t_0)}$ and $\boldsymbol{c_L(f_0,t_0)}$ respectively, in a window of time, $\boldsymbol{W}$, centered around $\boldsymbol{t_0}$ (15). We then find the peak of the resulting cross-correlation and its associated lag, $\boldsymbol{\tau}$, which corresponds to the time delay (16).

$$CCF(f_0;t_0,\tau) = \frac{1}{W}\sum_{k=t_0-\frac{W}{2}}^{t_0+\frac{W}{2}} c_R(f_0,k)\cdot c_L(f_0,k+\tau) \tag{15}$$

$$ITD(f_0,t_0) = \tau\,(argmax\,[CCF(f_0;t_0,\tau)]) \tag{16}$$

### C. CLUSTER ANALYSIS

Multiple considerations informed our selection of a clustering algorithm. Specifically, our decisions were guided by a broad survey of unsupervised and semi-supervised clustering algorithms [30]. First, the nature of the observed data points is that of continuous numerical data, therefore any algorithms that emphasize categorical data were excluded from consideration. Additionally, it is assumed that a single data point can potentially belong to more than one cluster/stream since the input auditory signal is an amalgamation of individual signals. Therefore, only algorithms capable of providing "fuzzy clusters" (clusters with overlapping borders) were considered. Lastly, due to the multidimensional nature of the feature space, it is difficult to ascertain the precise shape and density of clusters. Thus, we considered both a distributive mixture model approach (e.g., Gaussian Mixture Models [31])—which makes assumptions about the shape of the clusters but not the density—and density-based approaches (e.g., DBSCAN, OPTICS [32], [33])—which make assumptions about the density of the clusters but not the shape. Ultimately, we selected Gaussian Mixture Modeling as our preferred clustering approach because, compared to density-based methods, it is not as sensitive to initial parameters. We evaluated model performance by comparing the separated auditory streams and the original input components used to create the auditory scene. This comparison was carried out via 2-dimensional correlation [34]–[36] where the original source cochleagram is compared pixel-by-pixel to the cochleagram of the reconstructed stream (section 2.3.3).

### D. MODEL EVALUATION

For initial testing of our model, we constructed scenes from a set of 13 auditory events (single pure tone, harmonic tones,

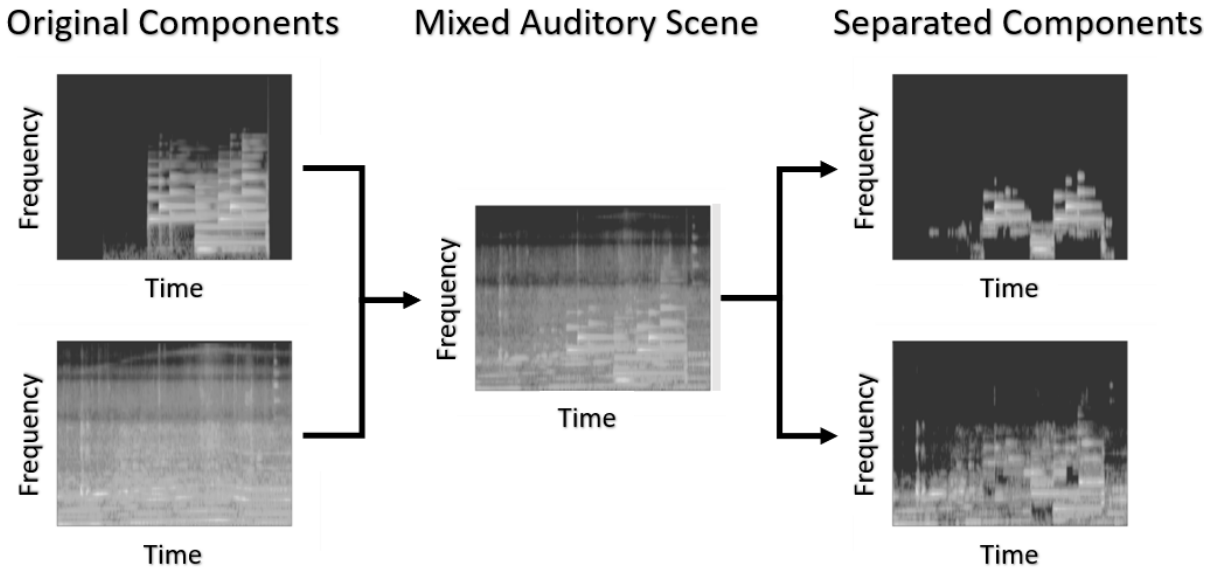


**FIGURE 3.** Cochleagrams of the individual components of an auditory scene involving a piano (top – original components) and a noisy bus ride background (bottom - original components), the cochleagram of the auditory scene (mixed auditory scene), and the results of the separations (separated components) when the mixed auditory scene (SNR = 0 dB) is analyzed with our model.

male voice, female voice, drums, saxophone, piano, reciprocating saw, pig grunt, pneumatic drill, ratchet tool, a passing train, and the ''Wilhelm scream'') over three backgrounds (Gaussian noise, a bus ride, and a busy restaurant); taking every combination of event and background resulted in 39 different scenes. Because the ideal number of clusters for any scene assessed in this section is two – one auditory event and the background - we manually set the number of clusters to two. To test the limits of our model's stream separation capabilities, we adjusted the signal-to-noise ratio (SNR) between event and background. For each of the 39 scenes, we tested 8 SNRs: −3, −6, −9, 0, 3, 6, 9, and 12 dB.

Continued testing involved adding a second auditory event to the scene shortly after the initial event. Scenes were created using the same 13 original components and 3 backgrounds; taking every combination of two events and a background thus resulted in 81 different scenes. Because the ideal number of clusters for any scene assessed in this section is three – two auditory events and a background – we manually set the number of clusters to three. For each scene, we again tested segregation at the 8 aforementioned SNRs.

We evaluated overall performance of the model at each SNR level via 2-dimensional correlation analysis, as in Krishnan *et al.* [8]. Here, we correlated the cochleagrams of the original components used to create each auditory scene with the segregation outputs of the model. This set of correlations was evaluated for each original component of a particular auditory scene. In scenes composed of one auditory event over a background, the ideal result is one high and one low correlation coefficient. We assume the high correlation coefficient corresponds to the correlation of the original component to the stream isolating that component, and the low correlation coefficient corresponds to the correlation of the original component to the stream isolating the other component of the auditory scene. In scenes comprised of two auditory events over a background, the ideal output is a high correlation coefficient and two lower correlation coefficients. Based on this reasoning, we chose the maximum value among the two or three correlations with an original scene component as the metric on which to evaluate segregation performance for each original component in each scene.

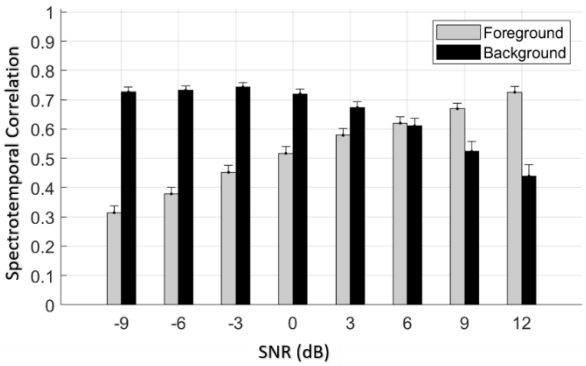


**FIGURE 4.** Two-dimensional correlations at each SNR level. Gray bars indicate max correlation of known foreground component to model outputs. Black bars indicate max correlation of known background component to model outputs. Error bars indicate the standard error of the correlation values.

## III. RESULTS

Fig. 3 shows an example of a scene made of a mixture of a piano and a noisy bus ride as a background; here the SNR between the piano and bus ride background was set to 0 dB. The top separated component appears to correspond to

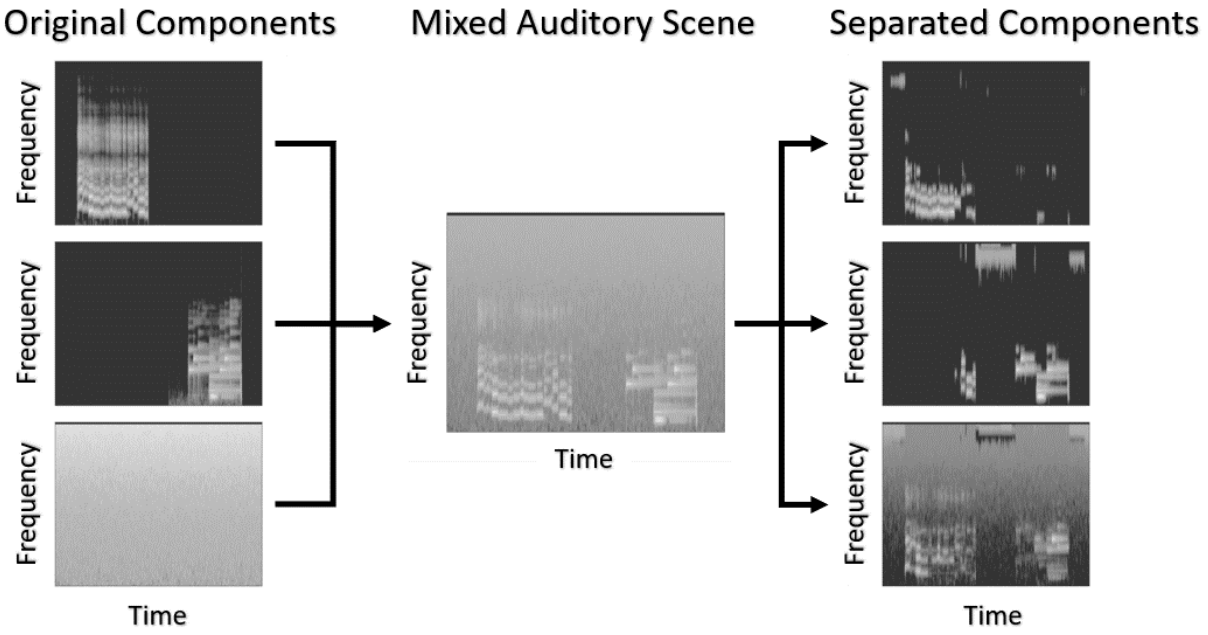


**FIGURE 5.** Separation of a complex auditory scene containing a saxophone (top - original components) and a piano (middle - original components), separated by 60° azimuthally with no temporal overlap, from a noisy bus background (bottom - original components; SNR = 0 dB).

the piano component, and the bottom separated component appears to correspond largely to the background component. When comparing the separated piano stream to the original component it is clear that a substantial portion of information is lost during separation from the mixture; however, enough information is present in the reconstruction to recognize the corresponding original component. Similarly, the reconstructed background component contains less information than the original background component (i.e., the lack of most high frequency energy from the original background component), and also includes some formants of the piano component. However, much of the lower frequency energy remains intact, making the separated component recognizable as corresponding to the background.

Fig. 4 summarizes the results of the 2-dimensional correlation analysis. Correlation coefficients were computed between each original component, whose identities are known, and both of the output separated components, whose identities are unknown. We assumed that for a given original component, the higher of the two correlations with model outputs corresponded to the closest ''match'' of the segregated component with its original spectrotemporal representation. Therefore, this correlation coefficient was taken as the metric on which to assess how effectively the model segregated that original component from the mixture. Fig. 4 displays a summary of the maximum correlation coefficients to the original foreground (auditory event) and background components across all SNRs evaluated. At the lowest SNR, the correlation coefficients of the background separated component is highest ($r \geq 0.7$) while the foreground component is lowest ($r \geq 0.3$). At the highest SNR, the correlations coefficients of the foreground separated component is highest ($r \geq 0.7$) while it is lowest for the background ($r \geq 0.45$). At a SNR of 6 dB, the correlation coefficients are nearly equivalent ($r \geq 0.6$).

Fig. 5 shows an example of a scene made of a mixture of a saxophone followed by a piano (separated by 60^o in azimuth), with a noisy bus ride as a background. In this example, the SNR between each event (i.e., piano and saxophone) and the background was 0 dB. The top separated component appears to correspond to the saxophone component and the middle separated component appears to correspond to the piano component. In the saxophone and piano components, much of the high frequency information is lost upon reconstruction, however enough information is retained in this reconstruction to recognize the original identity of the separated components. The bottom separated component appears to largely correspond to the background component. Whereas this component shows a retention of information across much of the frequency spectrum, there is also the inclusion of elements originally belonging to the saxophone and piano streams. An example of separating a saxophone and a piano, separated by 60° azimuthally with no temporal overlap, from a noisy background (SNR = 0 dB) is shown in Fig. 5.

To quantify the model's ability to segregate, we again computed correlation coefficients between each original component, whose identities are known, and the output separated

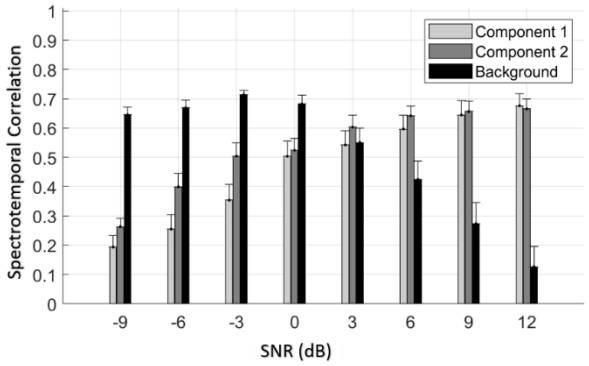


**FIGURE 6.** 2-dimensional correlations at each SNR level of scenes composed of two temporally distinct components and a background component. Light and dark gray bars indicate max correlation of known foreground components to model outputs. Black bars indicate max correlation of known background component to model outputs. Error bars indicate the standard error of the correlation values.

component, whose identities are unknown. Fig. 6 displays a summary of the maximum correlation coefficients to the foregrounds (auditory events) and background across all SNRs evaluated. At the lowest SNR, the correlation coefficients of the background separated component is highest ($r \geq 0.65$) while the foreground component is lowest ($0.19 - 0.28$). At the highest SNR, the correlations coefficients of the foreground separated component is highest $r \geq 0.68$) while it is lowest for the background $r \geq 0.1$). At an SNR of 3dB, the correlation coefficients are nearly equivalent ($r \geq 0.55$).

## IV. DISCUSSION

Most non-hearing impaired listeners are able to solve the cocktail party problem with little conscious effort. Yet, computational solutions to this problem have historically failed to match this ability. Our computational model attempts to fill the gap between human ability and computational recreation by creating a multidimensional space, composed of features noted to be critical for segregation in psychoacoustic literature [3], [10], [13], [16], and identifying clusters within this multidimensional space via Gaussian mixture modeling to extract auditory objects. Our model performs optimally (approximately 0.75 correlation between components and model outputs) at high signal-to-noise ratios but drops below 0.5 at SNRs higher than those seen in human trials [2].

Our evaluation of model performance was based on the correlation between the spectrotemporal representations, or cochleagrams, of the original components of an auditory scene and the segregated outputs of the model. As expected, lower signal-to-noise ratios resulted in lower correlation of the foreground, and higher correlation of the background. Interestingly, the correlation coefficients between $-9$ and $-3$ dB SNR for the background separation appears to have encountered a limit ($\boldsymbol{r \geq 0.7}$). A similar pattern was seen in the segregation of three-component scenes (Fig. 6). This limit is likely the result of the use of ideal binary masks to reconstruct original components of the scene; the use of binary masks results in blank spaces in at least one of the reconstructed streams anywhere there was spectrotemporal overlap between components. The more overlap seen in an auditory scene is directly related to the number of blank spaces present in the reconstructions; an effect seen as which SNR level produces equivalent correlation coefficients between background and foreground components (6 dB in two component scenes, 3 dB in three component scenes). The correlation limit and coefficient equivalence metrics highlight limitations in the model's reconstruction methods. Future work should explore methods of signal reconstruction that can account for, and possibly fill in, these blank spaces.

The strength of our model is that it achieves auditory signal segregation in complex scenes without the need for prior training. Human listeners are similarly capable of signal segregation without prior knowledge of objects in the scene, however, they often use expectations to improve this ability. For example, when conversing with a friend in a noisy restaurant, we likely use the known location and/or pitch of the friend's voice to help segregate and select it from among other auditory objects in the scene; this is often referred to as top-down attention. Though our model does not currently account for this aspect of human listening, such attentional mechanisms are essential to perception. Therefore, future implementations of this model should include components that mimic top-down attention. Additionally, our localization method is limited in that is only capable of segregating perceptual objects that are in the fore of the perceiver's "head." Differentiating forward-behind for auditory signals would likely some head related transfer function (HRTF). Future works applying this methodology to specific physical setups should seek to incorporate some HRTF that enables to the system to expand its localization capabilities to include forward and behind locations. The framework presented here can be easily adapted to take in known object features to inform cluster formation and selection within the multidimensional space. This could remove the requirement for users to manually input the number of clusters to extract from the feature space; instead this parameter could always be set to 2 clusters: one for the intended focus of attention and one for everything else (i.e., the background). Furthermore, such an attentional framework could be extended to include auditory object classification such that feature expectations would be naturally associated with the to-be-attended object.

Though successful in unsupervised scene segregation, our model is computationally intensive and currently incapable of working in real time. Improvement of the computational efficiency of our model requires that various trade-offs be made between segregation effectiveness and computational efficiency. For instance, the number of channels used in the gammatone filterbank decomposition at the beginning of the feature analysis stage has a major effect on the computation time needed to run the model (e.g., including fewer channels leads to lower computation times) and final signal segregation (more channels lead to higher quality segregation). Future work should be performed to optimize this trade-off between segregation accuracy and computational efficiency.

## V. CONCLUSION

Our model shows that unsupervised cluster analysis applied to psychoacoustic-inspired auditory features is capable of separating auditory streams within an auditory scene. Using unsupervised cluster analysis in this way is akin to mimicking the automatic responses of the auditory pathway in the brain. However, it is incomplete to say that auditory stream segregation is a fully automatic reaction independent of a priori knowledge and training adaptations. Top-down attention, goal setting, and/or intentions are clear factors in how humans successfully solve the 'cocktail party problem.' Therefore, we conclude that our current model is primarily a model of the automatic auditory stream segregation abilities of the human auditory pathway in the brain, and can serve as a basis for future models that incorporate higher level cognitive processes.

## DISCLAIMER

The views expressed in this article reflect the results of research conducted by the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the United States Government.

I am an employee of the U.S. Government. This work was prepared as part of my official duties. Title 17, U.S.C., §105 provides that copyright protection under this title is not available for any work of the U.S. Government. Title 17, U.S.C., §101 defines a U.S. Government work as a work prepared by a military Service member or employee of the U.S. Government as part of that person's official duties.

## REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.
[2] A. S. Bergman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1964.
[3] F. Alías, J. Socoró, and X. Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," *Appl. Sci.*, vol. 6, no. 5, p. 143, May 2016.
[4] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 887–906, Aug. 2005.
[5] B. De Coensel and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound," in *Proc. 20th Int. Congr. Acoust. (ICA)*, 2010, pp. 1–8.
[6] M. Elhilali and S. A. Shamma, "A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation," *J. Acoust. Soc. Amer.*, vol. 124, no. 6, pp. 3751–3771, Dec. 2008.
[7] O. Kalinli and S. S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proc. Interspeech*, Aug. 2007, pp. 1941–1944.
[8] L. Krishnan, M. Elhilali, and S. Shamma, "Segregating complex sound sources through temporal coherence," *PLoS Comput. Biol.*, vol. 10, no. 12, Dec. 2014, Art. no. e1003985.
[9] B. C. J. Moore and H. E. Gockel, "Properties of auditory stream formation," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 367, no. 1591, pp. 919–931, Apr. 2012.
[10] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.
[11] R. C. deCharms, "Optimizing sound features for cortical neurons," *Science*, vol. 280, no. 5368, pp. 1439–1444, May 1998.
[12] S. A. Shamma, H. Versnel, and N. Kowalski, "Ripple analysis in ferret primary auditory cortex. I. Response characteristics of single units to sinusoidally rippled spectra," *Auditory Neurosci.*, vol. 1, pp. 233–254, Jan. 1995.
[13] J. K. Bizley and Y. E. Cohen, "The what, where and how of auditory-object perception," *Nature Rev. Neurosci.*, vol. 14, no. 10, pp. 693–707, Oct. 2013.
[14] R. Lyon, "A computational model of binaural localization and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 1983, pp. 1148–1151.
[15] M. Boes, D. Oldoni, B. De Coensel, and D. Botteldooren, "A biologically inspired recurrent neural network for sound source recognition incorporating auditory attention," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–8.
[16] B. T. Szabó, S. L. Denham, and I. Winkler, "Computational models of auditory scene analysis: A review," *Frontiers Neurosci.*, vol. 10, p. 524, Nov. 2016.
[17] P. L. M. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," in *Proc. Symp. Hearing Theory, (IPO)*, 1972, pp. 58–69.
[18] G. Hu and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 2, pp. 396–405, Feb. 2007.
[19] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis," *Trends Neurosci.*, vol. 34, no. 3, pp. 114–123, Mar. 2011.
[20] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Amer.*, vol. 79, no. 3, pp. 702–711, 1986.
[21] R. Meddis, "Simulation of auditory–neural transduction: Further studies," *J. Acoust. Soc. Amer.*, vol. 83, no. 3, pp. 1056–1063, Mar. 1988.
[22] R. Meddis, M. J. Hewitt, and T. M. Shackleton, "Implementation details of a computation model of the inner hair-cell auditory-nerve synapse," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1813–1816, Apr. 1990.
[23] G. Z. Greenberg and W. D. Larkin, "Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method," *J. Acoust. Soc. Amer.*, vol. 44, no. 6, pp. 1513–1523, Dec. 1968.
[24] B. Scharf, S. Quigley, C. Aoki, N. Peachey, and A. Reeves, "Focused auditory attention and frequency selectivity," *Perception Psychophys.*, vol. 42, no. 3, pp. 215–223, May 1987.
[25] M. Botte, "Auditory attentional bandwidth: Effect of level and frequency range," *J. Acoust. Soc. Amer.*, vol. 98, no. 5, pp. 2475–2485, Nov. 1995.
[26] M.-C. Botte, C. Drake, R. Brochard, and S. Mcadams, "Perceptual attenuation of nonfocused auditory streams," *Perception Psychophys.*, vol. 59, no. 3, pp. 419–425, Apr. 1997.
[27] G. Hu and D. Wang, "Auditory segmentation based on event detection," in *Proc. ISCA Tutorial Res. Workshop (ITRW) Stat. Perceptual Audio Process.*, 2004, pp. 1–6.
[28] J. E. Rose, N. B. Gross, C. D. Geisler, and J. E. Hind, "Some neural mechanisms in the inferior colliculus of the cat which may be relevant to localization of a sound source," *J. Neurophysiol.*, vol. 29, no. 2, pp. 288–314, Mar. 1966.
[29] B. M. Sayers and E. C. Cherry, "Mechanism of binaural fusion in the hearing of speech," *J. Acoust. Soc. Amer.*, vol. 29, no. 9, pp. 973–987, Sep. 1957.
[30] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: A brief survey," *Rev. Mach. Learn. Techn. Process. Multimedia Content*, vol. 1, pp. 9–16, Jul. 2004.
[31] C. A. Bouman, M. Shapiro, G. W. Cook, C. B. Atkins, and H. Cheng. (1997). *Cluster: An Unsupervised Algorithm for Modeling Gaussian Mixtures*. [Online]. Available: http://dynamo.ecn.purdue.edu/~bouman/software/cluster
[32] B. Borah and D. K. Bhattacharyya, "An improved sampling-based DBSCAN for large spatial databases," in *Proc. Int. Conf. Intell. Sens. Inf. Process.*, Jan. 2004, pp. 92–96.
[33] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, 2017.
[34] D. I. Barnea and H. F. Silverman, "A class of algorithms for fast digital image registration," *IEEE Trans. Comput.*, vol. C-21, no. 2, pp. 179–186, Feb. 1972.

[35] K. Yen, E. K. Yen, and R. G. Johnston, "The ineffectiveness of the correlation coefficient for image comparisons," Los Alamos Nat. Lab., Washington, DC, USA, Tech. Rep. 17, 1996.
[36] E. Hall, L. Rouge, and R. Wong, "Hierarchical search for image matching," in *Proc. IEEE Conf. Decis. Control Including 15th Symp. Adapt. Processes*, Dec. 1976, pp. 791–796.

**MATTHEW S. DALEY** received the B.S. degree in biological sciences and the M.S. degree in biomedical engineering from Arizona State University, Tempe, AZ, USA, in 2013 and 2017, respectively.

From 2018 to 2021, he has worked with Naval Submarine Medical Research Laboratory, Groton, CT, USA. He has published technical reports and peer-reviewed journal articles while working with Naval Submarine Medical Research Laboratory. His research interests include simulation of the human auditory systems, underwater bioeffects of sound on human divers, and the development of physiologically based cognitive performance predictive models.

**LIA M. BONACCI** received the B.S. degree in biomedical engineering from the University of Connecticut, in 2014, and the Ph.D. degree in biomedical engineering from Boston University, in 2019. Her doctoral research focused on characterizing neural correlates of spatial attention in complex auditory scenes.

Since 2019, she has been working as a Research Scientist with Naval Submarine Medical Research Laboratory, Groton, CT, USA. Her current research interests include bio-inspired models of auditory processing and multi-sensory attention.

**DAVID H. GEVER** (Member, IEEE) received the B.Sc. degree in geology from Michigan State University, East Lansing, MI, USA, in 1972, and the M.Sc. degree in information system science from Salve Regina University, Newport, RI, USA, in 2002.

He was formerly an employee of the Woods Hole Oceanographic Institution (WHOI) as a Research Associate, followed by Sierra Geophysics as a Software Engineer, followed by SAIC, and Leidos with Naval Submarine Medical Research Laboratory (NSMRL), Groton, CT, USA. He is currently working as a Contractor and performing duties as a Software Analyst. He has published technical reports while working at WHOI, SAIC, and Leidos/NSMRL. His research interests include scientific application programming and data analysis in the fields of seismic reflection and refraction, ocean acoustics, sonar, and human factors engineering.

**KRYSTINA DIAZ** received the B.A. degree in general psychology and the B.S. degree in criminal justice: forensic psychology from the University of New Haven, West Haven, CT, USA, in 2017.

Shortly afterward, she began working as a Leidos Employee with Naval Submarine Medical Research Laboratory (NSMRL), Groton, CT, USA, conducting human factors research and investigating solutions for psychological resilience in operational environments.

**JEFFREY B. BOLKHOVSKY** received the B.S. and M.S. degrees in biomedical engineering from the Worcester Polytechnic Institute, in 2011 and 2014, respectively, and the Ph.D. degree in biomedical engineering from the University of Connecticut, in 2017. His doctoral research focused on use of cognitive modeling and physiological measures to track cognitive performance decrement due to fatigue.

From 2017 to 2018, he worked at Leidos, as a Research Scientist. Since 2018, he has been working as a Research Physiologist with Naval Submarine Medical Research Laboratory. His research interests include physiological monitoring, cognitive performance changes due to stress, auditory signal processing, and human systems integration with autonomous systems.