frontiers | Frontiers in Neuroscience

ORIGINAL RESEARCH
published: 28 April 2022
doi: 10.3389/fnins.2022.855753

Modeling the Repetition-Based Recovering of Acoustic and Visual Sources With Dendritic Neurons

Giorgia Dellaferrera [1,2*†], *Toshitake Asabuki* [1] *and Tomoki Fukai* [1]

[1] Neural Coding and Brain Computing Unit, Okinawa Institute of Science and Technology, Okinawa, Japan, [2] Institute of Neuroinformatics, University of Zurich and Swiss Federal Institute of Technology Zurich (ETH), Zurich, Switzerland

In natural auditory environments, acoustic signals originate from the temporal superimposition of different sound sources. The problem of inferring individual sources from ambiguous mixtures of sounds is known as blind source decomposition. Experiments on humans have demonstrated that the auditory system can identify sound sources as repeating patterns embedded in the acoustic input. Source repetition produces temporal regularities that can be detected and used for segregation. Specifically, listeners can identify sounds occurring more than once across different mixtures, but not sounds heard only in a single mixture. However, whether such a behavior can be computationally modeled has not yet been explored. Here, we propose a biologically inspired computational model to perform blind source separation on sequences of mixtures of acoustic stimuli. Our method relies on a somatodendritic neuron model trained with a Hebbian-like learning rule which was originally conceived to detect spatio-temporal patterns recurring in synaptic inputs. We show that the segregation capabilities of our model are reminiscent of the features of human performance in a variety of experimental settings involving synthesized sounds with naturalistic properties. Furthermore, we extend the study to investigate the properties of segregation on task settings not yet explored with human subjects, namely natural sounds and images. Overall, our work suggests that somatodendritic neuron models offer a promising neuro-inspired learning strategy to account for the characteristics of the brain segregation capabilities as well as to make predictions on yet untested experimental settings.

Keywords: dendritic neurons, spiking neural networks, blind source separation, sound source repetition, spatio-temporal structure

OPEN ACCESS

Edited by:
Emre O. Neftci,
University of California, Irvine,
United States

Reviewed by:
Lyes Khacef,
University of Groningen, Netherlands
Dylan Richard Muir,
University of Basel, Switzerland
Tom Tetzlaff,
Helmholtz Association of German
Research Centres (HZ), Germany

****Correspondence:***
Giorgia Dellaferrera
gde@zurich.ibm.com

†Present address:
Giorgia Dellaferrera,
Department of Ophthalmology,
Children's Hospital, Harvard Medical
School, Boston, MA, United States

Specialty section:
This article was submitted to
Neuromorphic Engineering,
a section of the journal
Frontiers in Neuroscience

Received: *15 January 2022*
Accepted: *31 March 2022*
Published: *28 April 2022*

Citation:
Dellaferrera G, Asabuki T and Fukai T
(2022) Modeling the Repetition-Based
Recovering of Acoustic and Visual
Sources With Dendritic Neurons.
Front. Neurosci. 16:855753.
doi: 10.3389/fnins.2022.855753

1. INTRODUCTION

Hearing a sound of specific interest in a noisy environment is a fundamental ability of the brain that is necessary for auditory scene analysis. To achieve this, the brain has to unambiguously separate the target auditory signal from other distractor signals. In this vein, a famous example is the "cocktail party effect" (Cherry, 1953), i.e., the ability to distinguish a particular speaker's voice against a multi-talker background (Brown et al., 2001; Mesgarani and Chang, 2012). Many psychophysical and neurobiological studies have been conducted to clarify the psychophysical properties and underlying mechanisms of the segregation of mixed signals (Asari et al., 2006; Bee and Micheyl, 2008; Narayan et al., 2008; McDermott, 2009; McDermott et al., 2011; Schmidt and Römer, 2011; Lewald and Getzmann, 2015; Li et al., 2017; Atilgan et al., 2018), and computational theories and models have also been proposed for this computation (Amari et al., 1995; Bell and Sejnowski, 1995; Sagi et al., 2001; Haykin and Chen, 2005; Elhilali and Shamma, 2009; Thakur et al., 2015; Dong et al., 2016; Kameoka et al., 2018; Karamatli et al., 2018; Sawada et al., 2019). However, how the brain attains its remarkable sound segregation remains elusive. Various properties of auditory cues such as spatial cues in binaural listening (Ding and Simon, 2012) and temporal coherence of sound stimuli (Teki et al., 2013; Krishnan et al., 2014) are known to facilitate the listener's ability to segregate a particular sound from the background. Auditory signals that reached to ears first undergo the analysis of frequency spectrums by cochlea (Oxenham, 2018). Simultaneous initiation and termination of the component signals and the harmonic structure of the frequency spectrums help the brain to identify the components of the target sound (Popham et al., 2018). Prior knowledge about the target sound, such as its familiarity to listeners (Elhilali, 2013; Woods and McDermott, 2018), and top-down attention can also improve their ability to detect the sound (Kerlin et al., 2010; Xiang et al., 2010; Ahveninen et al., 2011; Golumbic et al., 2013; O'Sullivan et al., 2014; Bronkhorst, 2015). Selective attention as the combination of the auditory (sound) and visual (lip movements, visual cues) modalities has also been suggested to be beneficial to solve the cocktail party problem (Yu, 2020; Liu et al., 2021). However, many of these cues are subsidiary and not absolutely required for hearing the target sound. For example, a mixture sound can be separated by monaural hearing (Hawley et al., 2004) or without spatial cues (Middlebrooks and Waters, 2020). Therefore, the crucial mechanisms of sound segregation remain to be explored.

Whether or not biological auditory systems segregate a sound based on principles similar to those invented for artificial systems remains unclear (Bee and Micheyl, 2008; McDermott, 2009). Among such principles, independent component analysis (ICA) (Comon, 1994) and its variants are the conventional mathematical tools used for solving the sound segregation problem, or more generally, the blind source decomposition problem (Amari et al., 1995; Bell and Sejnowski, 1995; Hyvärinen and Oja, 1997; Haykin and Chen, 2005). Owing to its linear algebraic features, the conventional ICA requires as many input channels (e.g., microphones) as the number of signal sources, which does not appear to be a requirement for sound segregation in biological systems. In this context, however, recent works for single-channel source separation based on techniques such as Non-Negative Matrix Factorization (NNMF) have demonstrated that ICA can be applied with a lower number of channels than the number of sources (Krause-Solberg and Iske, 2015; Mika et al., 2020). In addition, NNMF has been shown to extract regular spatio-temporal patterns within the audio and to achieve good performance in applications such as music processing (Smaragdis and Brown, 2003; Cichocki et al., 2006; Santosh and Bharathi, 2017; López-Serrano et al., 2019). It has been suggested as an alternative possibility that human listeners detect latent recurring patterns in the spectro-temporal structure of sound mixtures for separating individual sound sources (McDermott et al., 2011). This was indicated by the finding that listeners could identify a target sound when the sound was repeated in different mixtures in combination with various other sounds but could not do so when the sound was presented in a single mixture.

The finding represents an important piece of information about the computational principles of sound source separation in biological systems. Here, we demonstrate that a computational model implementing a pattern-detection mechanism accounts for the characteristic features of human performance observed in various task settings. To this end, we constructed a simplified model of biological auditory systems by using a two-compartment neuron model recently proposed for learning regularly or irregularly repeated patterns in input spike trains (Asabuki and Fukai, 2020). Importantly, this learning occurs in an unsupervised fashion based on the minimization principle of regularized information loss, showing that the essential computation of sound source segregation can emerge at the single-neuron level without teaching signals. Furthermore, it was previously suggested that a similar repetition-based learning mechanism may also work for the segregation of visual objects (McDermott et al., 2011). To provide a firm computational ground, we extended the tasks of our framework to predictions on visual images.

2. RESULTS

2.1. Learning of Repeated Input Patterns by a Two-Compartment Neuron Model

We used a two-compartment spiking neuron model which learns recurring temporal features in synaptic input, as proposed in Asabuki and Fukai (2020). In short, the dendritic compartment attempts to predict the responses of the soma to given synaptic input by modeling the somatic responses. To this end, the neuron model minimizes information loss within a recent period when the somatic activity is replaced with its model generated by the dendrite. Mathematically, the learning rule minimizes the Kullback–Leibler (KL) divergence between the probability distributions of somatic and dendritic activities. The dendritic membrane potential of a two-compartment neuron obeys $v(t) = \sum_j w_j e_j(t)$, where w_j and e_j stand for the synaptic weight and the unit postsynaptic potential of the j-th presynaptic input, respectively. The somatic activity evolves as

$$\dot{u}(t) = -\frac{1}{\tau}u(t) + g_D[-u(t) + v(t)] - \sum_j G_k \phi^{som}(u_k(t))/\phi_0, \quad (1)$$

where the last term describes lateral inhibition with modifiable synaptic weights G_k (≥ 0), as shown later. The soma generates a Poisson spike train with the instantaneous firing rate $\phi^{som}(u(t))$, where ${\phi_i}^{som}(u_i) = \phi_0[1 + e^{\beta(-u_i+\theta)}]^{-1}$, and the parameters β and θ are modified in an activity-dependent manner in terms of the mean and variance of the membrane potential over a sufficiently long period t_0. To extract the repeated patterns from temporal input, the model compresses the high dimensional data carried by the input sequence onto a low dimensional manifold of neural dynamics. This is performed by modifying the weights of dendritic synapses to minimize the time-averaged mismatch between the somatic and dendritic activities over a certain interval [0,T]. In a stationary state, the somatic membrane potential $u_i(t)$ can be described as an attenuated version $v_i^*(t)$ of the dendritic membrane potential. At each time point, we compare the attenuated dendritic membrane potential with the somatic membrane potential, on the level of the two Poissonian spike distributions with rates $\phi_i^{som}(u(t))$ and $\phi(v_i^*(t))$, respectively, which would be generated if both soma and dendrite were able to emit spikes independently. In practice, the neuron model minimizes the following cost function for synaptic weights w, which represents the averaged KL-divergence between somatic activity and dendritic activity, and in which we explicitly represent the dependency of u_i and v_i^* on X:

$$E(\mathbf{w}) = \int_{\Omega_X} dXP^*(\mathbf{X})$$
$$\int_0^T dt \sum_i D_{KL}[\phi_i^{som}(u_i(t;\mathbf{X}))||\phi^{dend}(v_i^*(t;\mathbf{X}))], \qquad (2)$$

with $P^*(\mathbf{X})$ and Ω_X being the true distribution of input spike trains and the entire space spanned by them, and $\phi^{dend}(x) = \phi_0[1 + e^{\beta_0(-x+\theta_0))}]^{-1}$. To search for the optimal weight matrix, the cost function $E(w)$ is minimized through gradient descent: $\Delta w_{ij} \propto -\partial E/\partial w_{ij}$. Introducing the regularization term $-\gamma \mathbf{w}_i$ and a noise component ξ_i with its intensity g gives the following learning rule (for the derivation see Asabuki and Fukai, 2020):

$$\dot{\mathbf{w}}_i(t) = \eta\{\psi(v_i^*(t))[\{f(\phi_i^{som}+\phi_0 g\xi_i)-\phi^{dend}(v_i^*(t))\}/\phi_0]\mathbf{e}(t)-\gamma\mathbf{w}_i\}, \qquad (3)$$

where $\mathbf{w}_i = [w_{i1,...,w_{iN_{in}}}]$, $\mathbf{e}(t) = [e_1, ...e_{Nin}]$, ξ_i obeys a normal distribution, $\psi(x) = \frac{d}{dx}log(\phi^{dend}(x))$, ϕ^{som} and ϕ^{dend} follow Poisson distributions, η is the learning rate, and

$$f(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \le x < \phi_0, \\ \phi_0 & \text{if } x \ge \phi_0 \end{cases}$$

Finally, if a pair of presynaptic and postsynaptic spikes occur at the times t_{pre} and t_{post}, respectively, lateral inhibitory connections between two-compartment neurons i and j are modified through a symmetric anti-Hebbian STDP as

$$\Delta G_{ij} = C_p exp\left(-\frac{t_{pre}-t_{post}}{\tau_p}\right) - C_d exp\left(-\frac{t_{pre}-t_{post}}{\tau_d}\right) \qquad (4)$$

See Section 4 and **Supplementary Note** for additional details. The prediction is learnable when input spike sequences from presynaptic neurons are non-random and contain recurring temporal patterns. In such a case, the minimization of information loss induces a consistency check between the dendrite and soma, eventually enforcing both compartments to respond selectively to one of the patterns. Mathematically, the somatic response serves as a teaching signal to supervise synaptic learning in the dendrite. Biologically, backpropagating action potentials may provide the supervising signal (Larkum et al., 1999; Larkum, 2013).

We constructed an artificial neural network based on the somatodendritic consistency check model and trained the network to perform the task of source recovering from embedded repetition. The network consisted of two layers of neurons. The input layer encoded the spectrogram of acoustic stimuli into spike trains of Poisson neurons. For each sound, the spike train was generated through a sequence of 400 time steps, where each time step corresponds to a "fire" or "non-fire" event. The output layer was a competitive network of the two-compartment models that received synaptic input from the input layer and learned recurring patterns in the input (**Figure 1**). We designed the output layer and the learning process similarly to the network used previously (Asabuki and Fukai, 2020) for the blind signal separation (BSS) within mixtures of multiple mutually correlated signals. In particular, lateral inhibitory connections between the output neurons underwent spike-timing-dependent plasticity for self-organizing an array of feature-selective output neurons (Section 4). In the spike encoding stage, the spectrogram is flattened into a one-dimensional array where the intensity of each element is proportional to the Poisson firing probability of the associated input neuron. This operation disconnects the signal's temporal features from the temporal dynamics of the neurons. Although this signal manipulation is not biologically plausible and introduces additional latency as the whole sample needs to be buffered, it allows the input layer to encode simultaneously all the time points of the audio signal. Thanks to this strategy, the length of the input spike trains does not depend on the duration of the audio signal, and a sufficiently large population of input neurons can encode arbitrarily long sounds, possibly with some redundancy in the encoding for short sounds. We remark that, while the somatodendritic mismatch learning rule was conceived to capture temporal information in an online fashion, in our framework it is applied to a flattened spectrogram, thus to a static pattern. Furthermore, in order to relate the signal intensity with the encoding firing rate, we normalized the spectrogram values to the interval [0,1]. This strategy is suited to our aim of reproducing the experiments with synthetic sounds and custom naturalistic stimuli. However, in a real-world application any instantaneous outlier in signal intensity would destroy other temporal features of an input signal. Nonetheless, the normalization is performed independently for each mixture, so if the outlier affects a masker sound and not a target, and the target is presented in at least two other mixtures, we expect that the normalization does not affect the ability of the network of identifying sounds presented in different mixtures.

2.2. Synthesized and Natural Auditory Stimuli

We examined whether the results of our computational model are consistent with the outcomes of the experiments on human listeners on artificially synthesized sounds described previously (McDermott et al., 2011). To provide a meaningful comparison with the human responses, we adopted for our simulations settings as close as possible to the experiments, both in terms of dataset generation and performance evaluation (Section 4). In McDermott et al. (2011), the generation of synthetic sounds is performed by first measuring the correlations between pairs of spectrograms cells of natural sounds (spoken words and

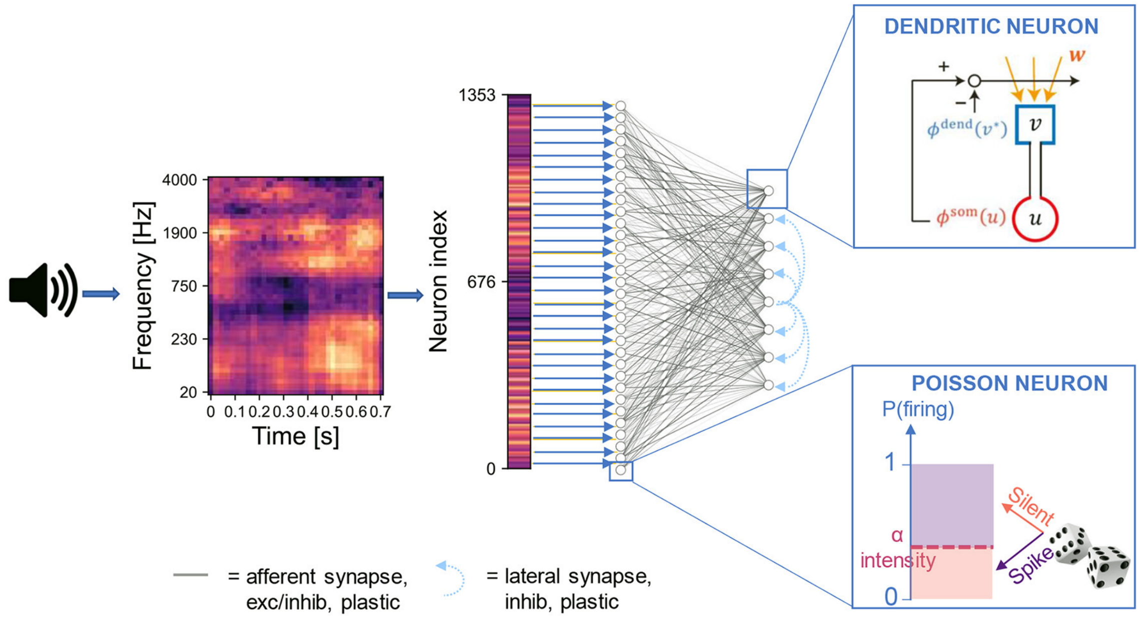

FIGURE 1 | Network architecture. The input signal is pre-processed into a two-dimensional image (i.e., the spectrogram) with values normalized in the range [0,1]. The image is flattened into a one-dimensional array where the intensity of each element is proportional to the Poisson firing probability of the associated input neuron. The neurons in the input layer are connected to those in the output layer through either full connectivity or random connectivity with connection probability $p = 0.3$. The output neurons are trained following the artificial dendritic neuron learning scheme (Asabuki and Fukai, 2020).

animal vocalizations). Then such correlations are averaged across different pairs to obtain temporal correlation functions. The correlation functions in turn are used to generate covariance matrices, in which each element is the covariance between two spectrogram cells. Finally, spectrograms are drawn from the resulting Gaussian distribution and applied to samples of white noise, leading to the synthesis of novel sounds. In our experiments we synthesized the sounds using the toolbox provided at https://mcdermottlab.mit.edu/downloads.html. In the human experiments, a dataset containing novel sounds was generated such that listeners' performance in sound source segregation was not influenced by familiarity with previously experienced sounds. To closely reproduce the experiment, we created a database of synthesized sounds according to the same method as described in McDermott et al. (2011) (Section 4). The synthesized stimuli retained similarity to real-world sounds except that they lacked grouping cues related to temporal onset and harmonic spectral structures. Furthermore, unlike human listeners, our neural network was trained and built from scratch, and had no previous knowledge of natural sounds that could bias the task execution. We exploited this advantage to investigate whether and how the sound segregation performance was affected by the presence of grouping cues in real sounds. To this goal we also built a database composed of natural sounds (Section 4).

To build the sequence of input stimuli, we randomly chose a set of sounds from the database of synthesized or natural sounds, and we generated various mixtures by superimposing them—i.e., we summed element-wise the spectrograms of the original sounds and then normalized the sum to the interval [0,1]. We refer to the main sound, which is always part of mixtures, as the *target*, and to all the other sounds, which were either presented as mixing sounds with the target (i.e., masker sounds) or presented alone, as *distractors*. The target sound is shown in red in the training protocols. Following the protocol in McDermott et al. (2011), we concatenated the mixtures of target and distractors into input sequences. For certain experiments, we also included unmixed distractor sounds. We presented the network with the input sequence for a fixed number of repetitions. As each input signal—both unmixed sounds and mixtures—is flattened into one input vector, each input signal is one element of the input sequence. During the input presentation, the network's parameters evolved following the learning rule described in Asabuki and Fukai (2020). Then, we examined the ability of the trained network to identify the target sound by using probe sounds, which were either the target or distractor sound composing the mixtures presented during training (*correct probe*) or a different sound (*incorrect probe*). Incorrect probes for synthesized target sounds were generated similarly as described in McDermott et al. (2011). Specifically, we synthesized the incorrect probe by using the same covariance structure of the target sound, and then we set a randomly selected time slice of the incorrect probe (1/8 of the sound's duration) to be equal to a time slice of the target of the same duration. Examples of target sounds, distractor sounds and incorrect probes are shown in **Figures 2A–C**, respectively. A further beneficial aspect of our model is the possibility of freezing plasticity during the inference stage, so that the synaptic

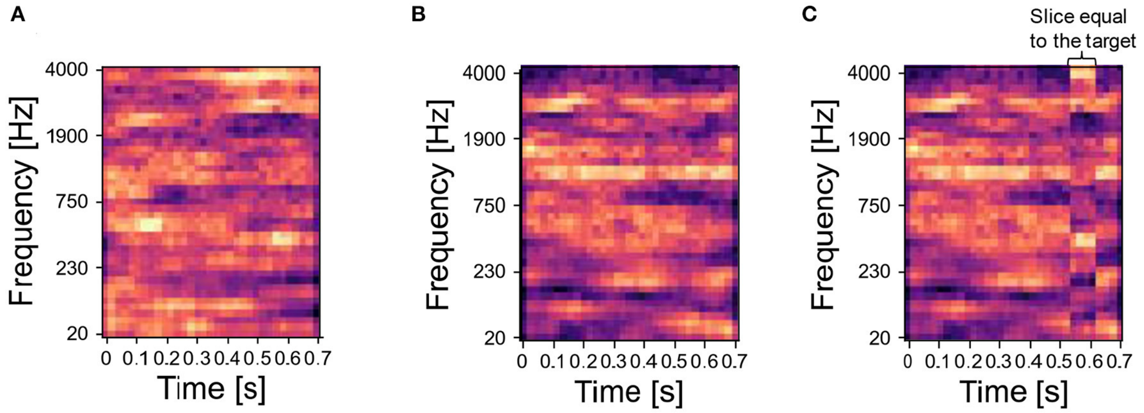

FIGURE 2 | Synthesized sounds—target and associated distractor. **(A)** Spectrogram of one target sound. **(B)** Step 1 to build the spectrogram of an incorrect probe related to the target in **(A)**: a sound is randomly selected from the same Gaussian distribution generating the target. **(C)** Step 2 to build the incorrect probe: after the sampling, a randomly selected time slice equal to 1/8 of the sound duration is set to be equal to the target. In the figure, the temporal slice is the vertical stripe around time 0.5 s.

connections do not change during the probe presentation. This allows us to investigate whether the trained network can identify not only the target but also the masker sounds.

2.3. Learning of Mixture Sounds in the Network Model

Our network model contained various hyperparameters such as number of output neurons, number of mixtures and connectivity pattern. A grid search was performed to find the best combination of hyperparameters. **Figures 3A,B** report the learning curves obtained on synthesized and natural sounds, respectively, for random initial weights and different combinations of hyperparameters. For both types of sounds, synaptic weights changed rapidly in the initial phase of learning. The changes were somewhat faster for synthesized sounds than for natural sounds, but the learning curves behaved similarly for both sound types. The number of output neurons little affected the learning curves, while they behaved differently for different connectivity patterns or different numbers of mixtures. Because familiarity to sounds enhances auditory perception in humans (Jacobsen et al., 2005), we investigated whether pretraining with a sequence containing target and distractors improves learning in our model for various lengths of pretraining. Neither the training speed nor the final accuracy were significantly improved by the pretraining (**Figures 3C–E**). This suggests that the model was "forgetting" about the pretraining stage and learning the mixture sounds from scratch, not exploiting any familiarity with previously seen sounds. We suspect that this behavior is related to the well know limitation of ANNs of lack of continual learning (French, 1999) rather than to a specific feature of our model. Furthermore, we cannot provide a comparison in the learning curve between the model and the psychophysical data, since the model was trained for multiple epochs, while the human listeners were presented with the training sequence only once and then tested on the probe immediately after.

To reliably compare the performance of our model with human listeners, we designed a similar assessment strategy to that adopted in the experiment. In McDermott et al. (2011), listeners were presented with mixtures of sounds followed by a probe which could be either a correct probe (i.e., the target sound present in the training mixtures) or an incorrect probe (i.e., sounds unseen during the training). The subjects had to say whether they believed the probe was present in the training mixture by using one of the four responses "sure no," "no," "yes," and "sure yes." The responses were used to build a receiver operating characteristics (ROC) as described in Wickens (2002), and the area under the curve (AUC) was used as performance measure, with AUC = 0.5 and 1 corresponding to chance and perfect correct, respectively. In our algorithm, we mimicked this protocol for reporting by using the likelihood as a measure of performance. To this goal, first, for each tested probe, we projected the response of the N output neurons (**Figures 4A,D**) to a two-dimensional PCA projection plane. We defined the PCA space based on the response to the correct probes and later projected on it the datapoints related to the incorrect probes (**Figures 4B,E**). We remark that other clustering approaches such as K-means and self-organizing maps could be used instead of PCA without reducing the output dimension. Second, we clustered the datapoints related to the correct probes through a Gaussian Mixture Model (GMM) with as many classes as the number of correct probes (**Figures 4C,F**). Third, for each datapoint we computed the likelihood that it belonged to one of the clusters. The target likelihood values are fixed to 1 and 0 for datapoints related to correct and incorrect probes respectively. We highlight that the labels introduced in this post-processing phase are not specific for each sound, but rather depend on the role of the sound in the tasks, i.e., if sound X is presented during training as a target or masker sound it is associated to label 1, while if, in another simulation, the same sound X is used to build an incorrect probe (not used during training) then it is associated with label 0. We binned the likelihood range into

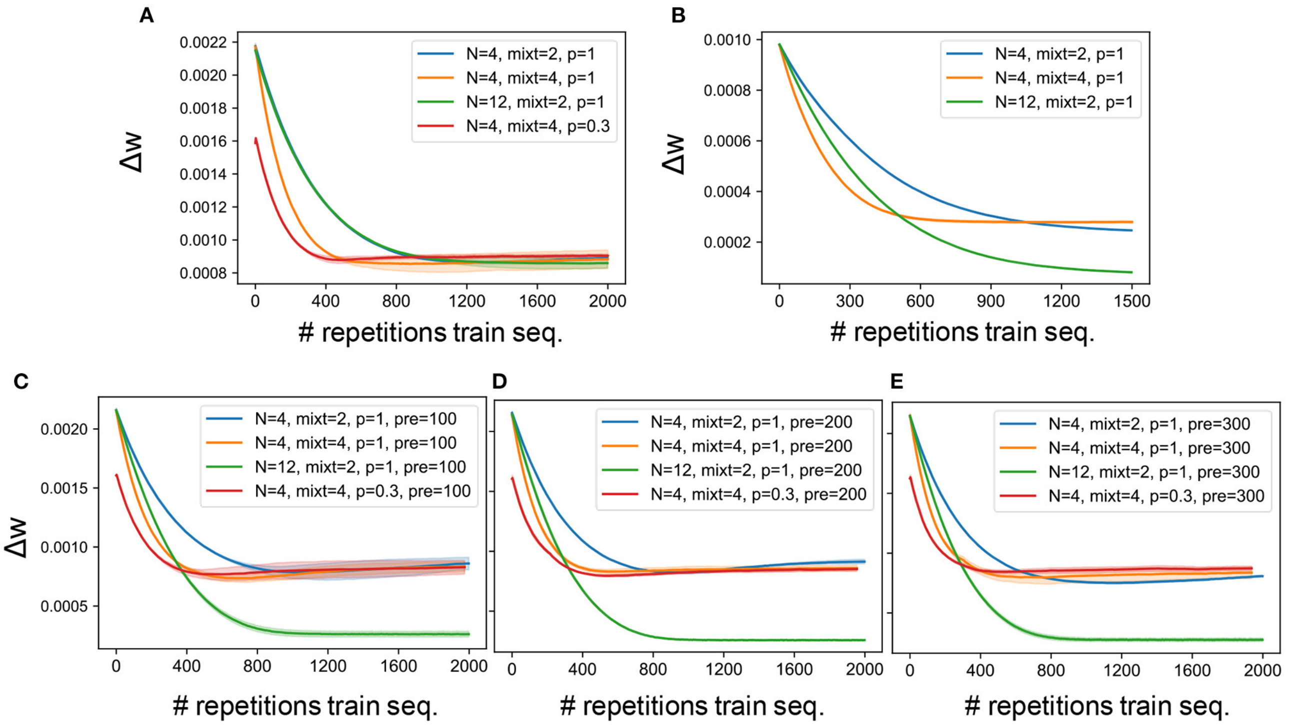

FIGURE 3 | Learning curves. **(A)** Average synaptic weight change for the experiments carried out on the synthetized sounds, the network being initialized with random values. **(B)** Average synaptic weight change for the experiments carried out on the natural sounds, the network being initialized with random values. **(C)** Average synaptic weight change for the experiments carried out on the synthetized sounds, the network being pretrained on the targets set presented for 100 epochs. **(D)** Average synaptic weight change for the experiments carried out on the synthetized sounds, the network being pretrained on the targets set presented for 200 epochs. **(E)** Average synaptic weight change for the experiments carried out on the synthetized sounds, the network being pretrained on the targets set presented for 300 epochs. The solid line and the shaded area represent the mean and standard deviation over 3 independent runs, respectively. Without pretraining, when the number of output neurons is varied no significant change is found, while with pretraining when a larger number of neurons is used, the weight change curve saturates at a lower value, as shown by the blue ($N = 4$) and green ($N = 12$) curves. Furthermore, the figures show that both when a larger number of training mixtures is presented (yellow curves) and when only 30% of the connections are kept (red curves) the slope of the learning curve is steeper. The weight change is computed by storing the weights values every 2,000 time steps (i.e., "fire" or "non-fire" events) and computing the standard deviation over the last 100 recorded values. The standard deviation is then averaged across all connections from input to output neurons. Therefore, each point on the curve reports the average weight change over the past 2000×100 time steps. Note that each sound/mixture is presented for 400 time steps. Finally, the x-axis shows the number of repetitions of the training mixture sequence (2,000 for synthetic sounds and 1,500 for naturalistic sounds).

four intervals corresponding, in an ascending order, to the four responses "sure no," "no," "yes," and "sure yes." Finally, based on the four responses, we built the receiver operating characteristic (ROC) curve: the datapoints falling in the interval (i) $L > 0$ (sure yes) were assigned the probability value $p = 1.0$, those in (ii) $-5 < L < 0$ (yes) $p = 0.66$, those in (iii) $-15 < L < -5$ (no) $p = 0.33$, and those in (iv) $L < -15$ (sure no) $p = 0.0$. The AUC of the ROC is used as the "accuracy" metric to evaluate the performance of the model. For additional details see Section 4. Now, we are ready to examine the performance of the model in a series of experiments. We show examples of the different behavior of the network trained on single (**Figures 4A–C**) or four mixtures (**Figures 4D–F**). As expected, the ability of the model to learn and distinguish the targets from the distractors depended crucially on the number of mixtures.

The algorithm was implemented in Python and a sample code used to simulate Experiment 1 is available at the repository https://github.com/GiorgiaD/dendritic-neuron-BSS.

2.4. Experiment 1: Sound Segregation With Single and Multiple Mixtures of Synthesized Sounds

To begin with, we compared how the number of mixtures influences the learning performance between human subjects and the model. The number of mixtures presented during training was varied from 1, where no learning was expected, to 2 or more, where the model was expected to distinguish the target sounds from their respective distractors. The simulation protocol is shown in **Figure 5A** (bottom). As reported in **Figure 5A** (top), we obtained that, when one mixture only was shown, neither the target nor the mixing sound was learnt, and performance was close to chance. An immediate boost in the performance was observed when the number of mixtures was raised to two. The network managed to distinguish the learnt targets from the incorrect probes with an accuracy greater than 90%. As the number of mixtures increased up to six, the accuracy worsened slightly, remaining above 80%.

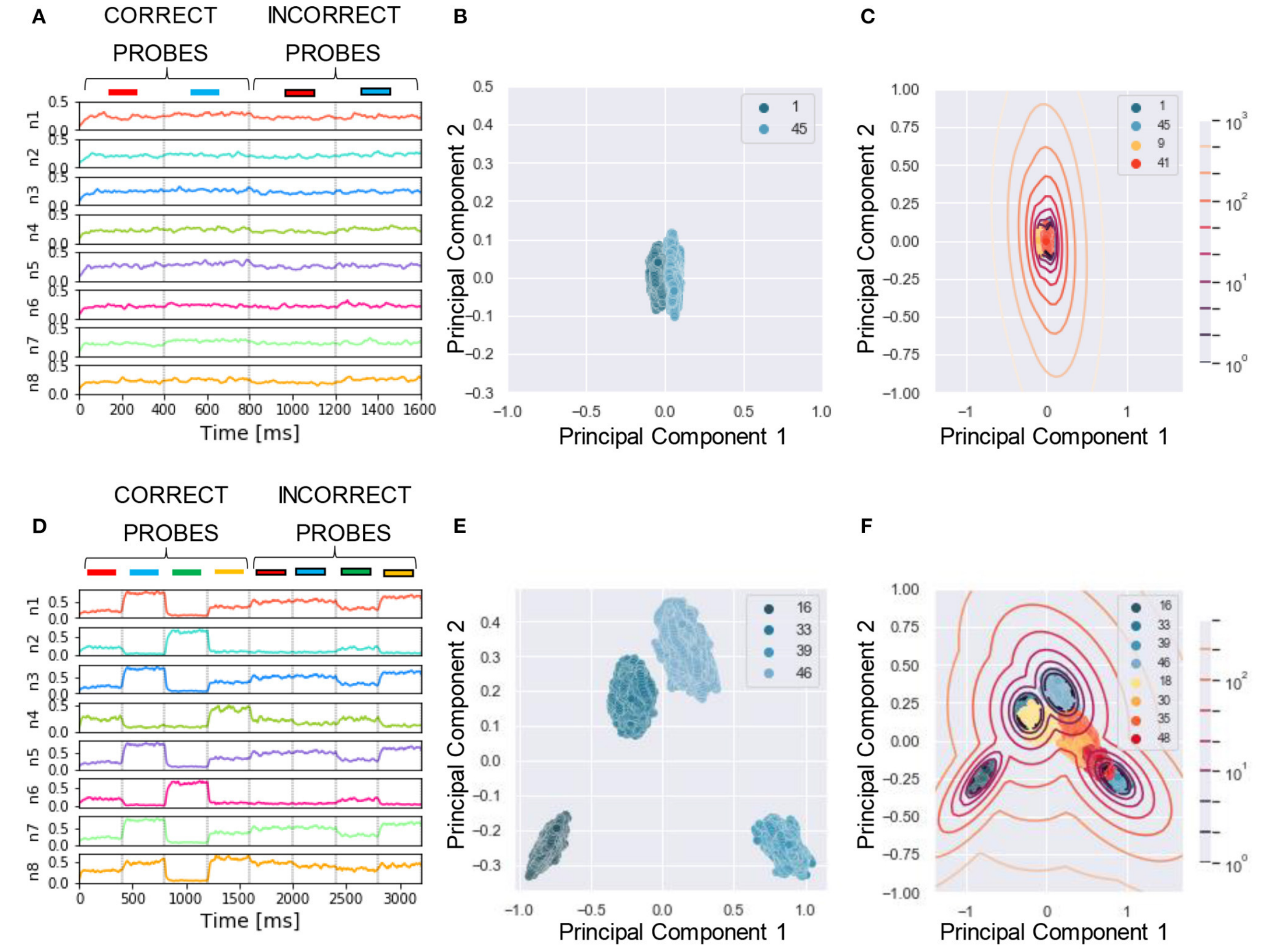

FIGURE 4 | Experiment 1—output dynamics and clustering. **(A–C)** refer to the results of Experiment 1 on synthesized sounds with a single mixture presented during training. **(D–F)** refer to the results of Experiment 1 on synthesized sounds with three mixtures presented during training. The "correct probes" are the target and the distractor sounds composing the mixtures presented during training, while the "incorrect probes" are sounds not presented during training. The numbers in the legends indicate the sound IDs. **(A)** Voltage dynamics of the 8 output neurons during inference, when the target, the distractor and the two associated incorrect probes are tested. The neuron population is not able to respond with different dynamics to the four sounds, and the voltage of all the output neurons fluctuates randomly throughout the whole testing sequence. **(B)** The PCA projection of the datapoints belonging to the two targets (in blue) shows that the clusters are collapsed into a single cluster. **(C)** When GMM is applied, all the datapoints representing both the correct probes (in blue) and the incorrect probes (in orange and red) fall within the same regions, making it impossible to distinguish the different sounds based on the population dynamics. **(D)** Voltage dynamics of the 8 output neurons during inference, when the four targets and the associated distractors are tested. As expected, the neuron population has learnt the feature of the different sounds and responds with different dynamics to the eight sounds. Each output neuron exhibits an enhanced response to one or few sounds. **(E)** The PCA projection of the datapoints belonging to the four correct probes (in blue) shows that the clusters are compact and spatially distant one from the other. **(F)** When GMM is applied, the model shows that the network is, most of the times, able to distinguish the target and distractors (in blue) from the incorrect probes (in yellow, orange and red). The correct probes are never overlapped. Three of the four distractors fall far from the targets' region, while the fourth (in yellow) overlaps with one of the targets. These results are overall coherent with the human performance. In **(C,F)**, the contour lines represent the landscape of the log-likelihood that a point belongs to one of the clusters associated to the correct probes.

A significant drop in the performance was observed for a greater number of mixtures. From a comparison with the results shown in **Figure 5B**, which were replicated for human subjects (McDermott et al., 2011), it emerged that our model was able to partially reproduce human performance: the success rate was at chance levels when training consists of a single mixture only; the target sounds could be distinguished to a certain accuracy if more than a mixture was learnt. We also verified that the model performance was robust for variations of the network architecture, both in terms of the number of output neurons *N* and the connection probability *p* (**Supplementary Figure 1**). Furthermore we observe that, while none of the output neurons exhibits an enhanced high firing rate when presented with the target sound, the overall population response to the target is substantially different from the response to the masker sounds and to the incorrect probes.

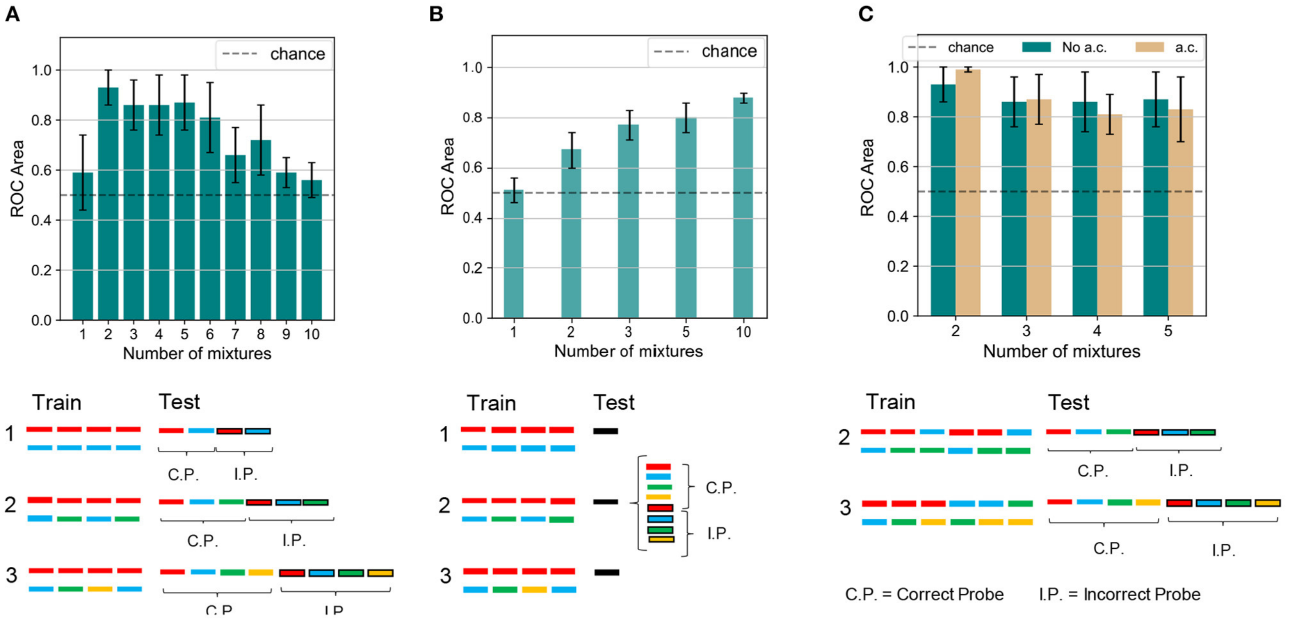

FIGURE 5 | Experiment 1 and 1 a.c.—results and comparison with human performance. **(A)** Results and schematics for Experiment 1 on the dendritic network model. The number of mixtures is varied from 1 to 10. Performance is close to chance for a single training mixture. The performance is boosted as two mixtures are presented. As the number of mixtures is further increased, the clustering accuracy slowly decreases toward chance values. The protocol shown at the bottom of the panel illustrates that (i) in the training phase we feed the network only with the mixture(s), i.e., target+masker sound(s). (ii) in the inference phase we feed the network only with the unmixed sounds (target, distractor separately) and with the incorrect probes (also unmixed sounds). We remark that in the case of one mixture (condition 1) the target and the masker sounds play the same role, while in the case of multiple mixtures (conditions 2 and 3) the target has a different role in the protocol as it is present in more than one mixture while the masker sounds are presented in one mixture only in the training sequence. **(B)** Results and schematics for Experiment 1 on the human experiment. The number of mixtures presented are 1, 2, 3, 5, and 10. For a single mixture the performance is close to chance. As the number of mixtures increases, the classification accuracy improves steadily. Figure reproduced based on data acquired by McDermott et al. (2011). **(C)** Results and schematics for Experiment 1 a.c. on the dendritic network model. The number of mixtures is varied from 2 to 5. Combining all the mixing sounds in mixtures slightly improves the mean performance for two mixing sounds, while it slightly worsens it for a larger number of mixtures. The height of the bars and the error bars show, respectively, mean and standard deviation of the AUC over 10 independent runs.

Our model and human subjects also exhibited interesting differences. When the mixture number was increased to two, performance improved greatly in our model but only modestly in human subjects. Unlike human subjects, our model showed a decreasing accuracy as the number of mixtures further increased. We consider that such discrepancies may arise from a capacity limitation of the network. Indeed, the network architecture is very simple and consists of two layers only, whose size is limited by the spectrogram dimensions for the input layer and by the number of output neurons for the last layer. Therefore the amount of information that the network can learn and store is limited with respect to the significantly more complex structure of the human auditory system. We also suspect that the two-dimensional PCA projection might limit the model performance when a large number of distractors is used. Indeed the PCA space becomes very crowded and although the datapoints are grouped in distinct clusters, the probability that such a cluster lie close to each other is high. To verify this hypothesis, we tested a modification of the inference protocol of the algorithm. During test, we presented the network only with the target sound and one incorrect probe, and performed BSS on the PCA space containing the two sounds. Under this configuration, the model performance is above chance level for two or more different mixtures, and the accuracy does not significantly decrease for large number of mixtures (**Supplementary Figure 2**).

We may use our model for predicting performance of human subjects in auditory perception tasks not yet tested experimentally. To this end, we propose an extension of the paradigm tested previously: for set-ups with the number of mixtures between two and five, we investigated whether presenting all possible combinations of the mixing sounds among themselves, rather than only the distractors with the target, affects the performance. The experiment is labeled "Experiment 1 a.c.," where a.c. stands for "all combinations," and its training scheme is reported in **Figure 5C**. Because all sounds are in principle learnable in the new paradigm, we expect an enhanced ability of distinguishing the correct probes from the incorrect ones. Somewhat unexpectedly, however, our model indicated no drastic changes in the performance when the mixture sequence presented during training contained all possible combinations of the mixing sounds. Such a scheme resulted in a minor improvement in the accuracy only for the experiments with two mixing sounds. Indeed, in the "all

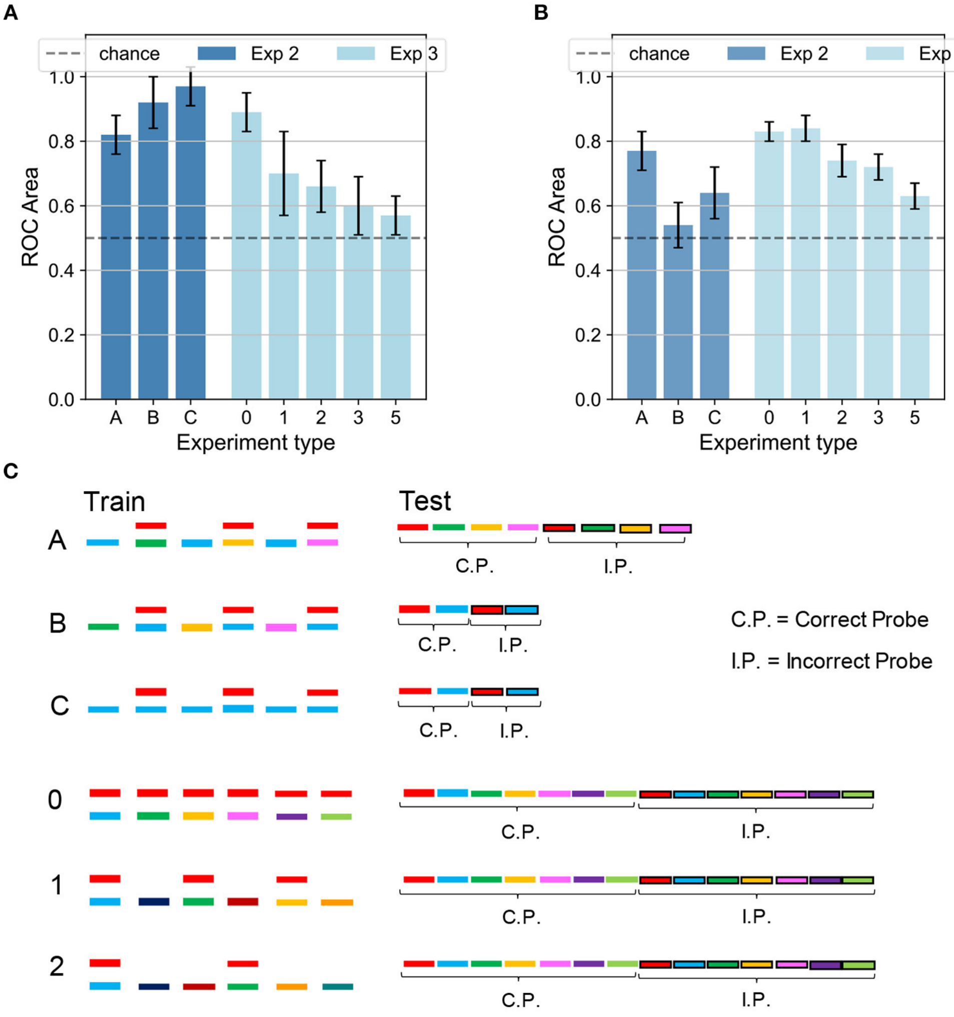

FIGURE 6 | Experiments 2 and 3—results and comparison with human performance. **(A)** Results for Experiments 2 (dark blue) and 3 (light blue) on the dendritic network model. In Experiment 2 the performance is above chance for the three conditions. In Experiment 3 the accuracy decreases as the number of isolated sounds alternating with the mixtures increases. **(B)** Results for Experiments 2 (dark blue) and 3 (light blue) on the human experiment. In Experiment 2 the performance is above chance in the conditions A and C, while it is random for condition B. In Experiment 3 the accuracy decreases as the target presentation is more delayed. Figure reproduced based on data acquired by McDermott et al. (2011). **(C)** Schematics for Experiments 2 and 3. The training is the same for both the dendritic network model and the human experiment. The schematics is omitted for delays 3 and 5. The testing refers to the dendritic network model, while the testing for the human experiment (same as in **Figure 5B**) is omitted. In **(A,B)**, the height of the bars and the error bars show respectively mean and standard deviation of the AUC over 10 independent runs.

combinations" protocol, during training the distractor was presented in more than one different mixture, while in the original task setting only the target was combined with different sounds. We hypothesize that the "all combinations" protocol makes it easier for the network to better distinguish the distractor sound. For four or five mixing sounds, instead, the performance slightly worsened. It is likely that this behavior is related to the already mentioned capacity restraints of the network. Indeed, the length of the training sequence grows as the binomial coefficient $\binom{n}{k}$ where $k = 2$, therefore for four and five targets (i.e., for $n = 4$ or 5) the number of mixtures is increased to 6 and 10, respectively.

2.5. Experiment 2: Sound Segregation With Alternating Multiple Mixtures of Synthesized Sounds

Next, we investigate the model's performance when the training sequence alternated mixtures of sounds with isolated sounds. An analogous protocol was tested in a psychophysical experiment (see experiment 3 in McDermott et al., 2011). **Figures 6A,B** show the network accuracy and human performance, respectively, for the protocols A,B,C in **Figure 6C**. Only the target and the masker sounds were later tested since recognizing the sounds presented individually during training would have been trivial (see conditions B, 1, and 2 in **Figure 6C**). In the alternating task, the network was only partially able to reproduce the human results, displaying an interesting contrast to human behavior. In condition A, in which the sounds mixed with the main target (in red) changed during training, the listeners were able to learn the targets with an accuracy of about 80%, and so did our model. In contrast, our network behaved radically differently with respect to human performance under condition B, in which the training sequence consisted of the same mixture alternating with different sounds. As reported in **Figure 5B**, the listeners were generally not able to identify the single sounds composing the mixture. Our model, instead, unexpectedly achieved a performance well above chance. The output dynamics could distinguish the distractors from the two targets with accuracy surprisingly above 90%. The behavioral discrepancy under condition B could be explained by considering that in the training scheme the network is presented with three different sounds besides the mixture. With respect to Experiment 1 with a single mixture, in this protocol the network could learn the supplementary features of the isolated sounds and could exploit them during inference to respond differently to the distractors. From the spectrograms shown in **Figure 2**, it is evident that some regions of overlap exist between the higher-intensity areas of different sounds. Therefore, the network presented during training with isolated sounds in addition to the single mixture, could detect some similarities between the training sounds and the tested distractors and respond with a more defined output dynamics than in Experiment 1. Finally, under condition C, both human subjects and our model performed above chance. While human performance was slightly above 60%, the network achieved more than 90% accuracy. This result should be interpreted considering that during inference also the isolated sound (blue) was tested together with the associated distractor, which was a trivial task for the nature of our network and thus boosted its overall performance.

2.6. Experiment 3: Effect of Temporal Delay in Target Presentation With Synthesized Sounds

Temporal delay in the presentation of mixtures containing the target degraded performance similarly in the model and human subjects. We presented the network with a training sequence of six mixtures containing the same target mixed each time with a different distractor (**Figure 6C**, protocols 0,1,2: c.f. experiment 4 in McDermott et al., 2011). The mixtures alternated with an increasing number of isolated sounds, hence increasing the interval between successive presentations of the target. The human ability to extract single sounds from mixtures was previously shown to worsen as the interval between target presentations increased, as replicated in **Figure 6B**. The network presented a similar decreasing trend, as reported in **Figure 6A**. An interesting difference, however, is that the performance of our model drastically dropped even with one isolated sound every other mixture while the human performance was affected when at least two isolated sounds separated the target-containing mixtures. The discrepant behavior indicates that the insertion of isolated sounds between the target-containing mixtures more strongly interferes the learning of the target sound in the model compared to human subjects. This stronger performance degradation may partly be due to the capacity constraint of our simple neural model, which uses a larger amount of memory resource as the number of isolated sounds increases. In contrast, such a constraint may be less tight in the human auditory systems.

Also for Experiments 2 and 3, we tested a modification of the inference protocol, by presenting the network only with the target sound and one incorrect probe. Under this configuration, the model performance of Experiment 2 improves compared to the original protocol, while no substantial changes are noted for Experiment 3 (**Supplementary Figure 3**).

2.7. Experiment 4: Sound Segregation With Single and Multiple Mixtures of Real-World Sounds

We applied the same protocol of Experiments 1 to the dataset of natural sounds. Although such experiments were previously not attempted on human subjects, it is intriguing to investigate whether the model can segregate target natural sounds by the same strategy. The spectrograms of two isolated sounds and of their mixtures are shown in **Figures 7A–C**, together with the respective sound waves (**Figures 7D–F**). The qualitative performance was very similar to that obtained with the synthesized sounds. Specifically, the output dynamics learned from the repetition of a single mixture was randomly fluctuating for both seen and randomly chosen unseen sounds (**Figure 8A**), whereas the network responses to targets and unseen sounds were clearly distinct if multiple mixtures were presented during training (**Figure 8D**). The output dynamics were not quantitatively evaluated because it was not possible to rigorously generate incorrect probes associated with the learnt targets and distractors. Therefore, we qualitatively assessed the performance of the model by observing the clustering of network responses to the learnt targets vs. unseen natural sounds (**Figures 8B–F**). We observed that, in the case of multiple mixtures, the clusters related to natural sounds (**Figures 8E,F**) were more compact than those of synthetic sounds (**Figures 4E,F**). Furthermore, these clusters were more widely spaced on the PCA projection plane: the intraclass correlation in the response to the same target was greater while the interclass similarity in the response to different targets or distractors was lower. These results indicate that grouping cues, such as harmonic structure and temporal onset, improve the performance of the model.

2.8. Experiment 5: Image Segregation With Single and Multiple Mixtures of Real-World Images

Finally, we examined whether the source segregation through repetition scheme can also extend to vision-related tasks, as previously suggested (McDermott et al., 2011). To this end, we employed the same method as developed for sound sources and performed the recovery of visual sources with the protocol of Experiment 1. The mixtures were obtained by overlapping black-and-white images sampled from our visual dataset (Section 4), as shown in **Figure 9**. Similarly to Experiment 4, the performance of the model was assessed only qualitatively in the

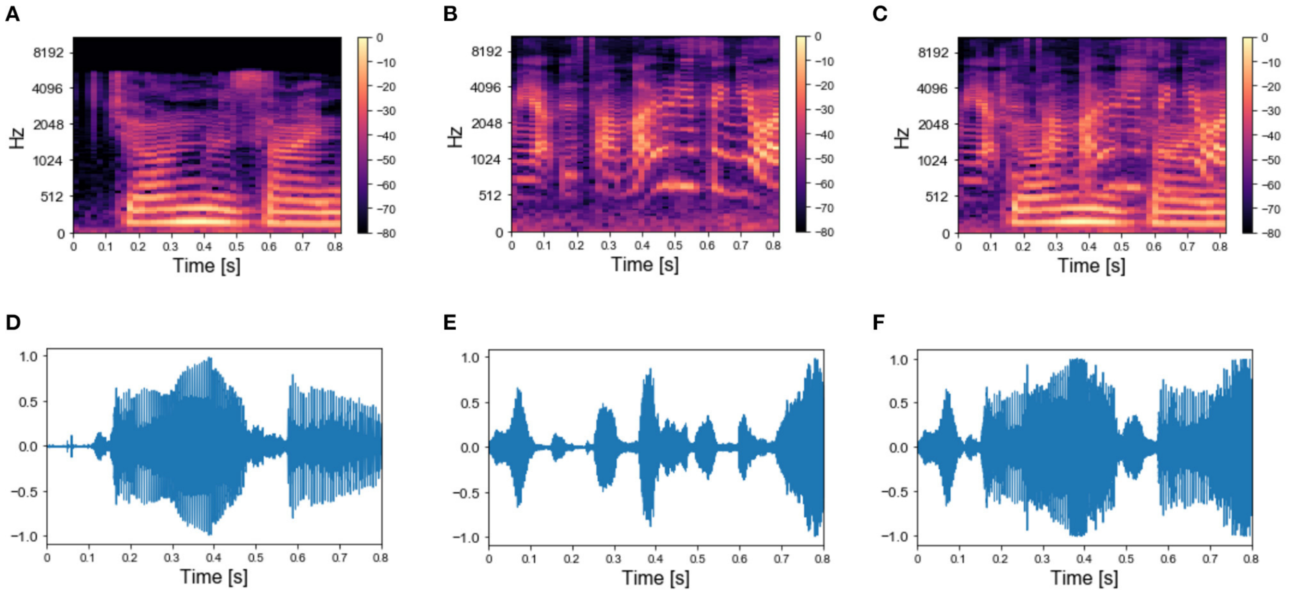

FIGURE 7 | Real-world sounds—targets and mixture. **(A)** Spectrogram of a spoken sentence 800 ms-long. **(B)** Spectrogram of 800 ms-long recording of chimes sounds. **(C)** Spectrogram of the mixture of the sounds in **(A,B)**. **(D)** Sound wave associated with the spectrogram in **(A)**. **(E)** Sound wave associated with the spectrogram in **(B)**. **(F)** Sound wave associated with the spectrogram in **(C)**.

visual tasks. As in the acoustic tasks, the clustering of network responses showed that the model was able to retrieve the single images only when more than one mixture was presented during training. The network responses are shown in **Figure 10**. We remark that the model is presented with the visual stimuli following the same computational steps as for sounds. Indeed, as previously described, the acoustic stimuli are first pre-processed into spectrograms and then encoded by the input layer. While it is not unexpected that similar computational steps lead to consistent results, we remark that the nature of the "audio images," i.e., the spectrograms, is substantially different to that of the naturalistic images, leading to very different distributions of the encoding spike patterns. Therefore, successful signal discrimination in the visual task strengthens our results, proving that our model is robust with respect to different arrangements of signal intensity.

3. DISCUSSION

The recovery of individual sound sources from mixtures of multiple sounds is a central challenge of hearing. Based on experiments on human listeners, sound segregation has been postulated to arise from prior knowledge of sound characteristics or detection of repeating spectro-temporal structure. The results of McDermott et al. (2011) show that a sound source can be recovered from a sequence of mixtures if it occurs more than once and is mixed with more than one masker sound. This supports the hypothesis that the auditory system detects repeating spectro-temporal structure embedded in mixtures, and interprets this structure as a sound source. We investigated whether a biologically inspired computational model of the auditory system can account for the characteristic performance of human subjects. To this end, we implemented a one-layer neural network with dendritic neurons followed by a readout layer based on GMM to classify probe sounds as seen or unseen in the training mixtures. The results in McDermott et al. (2011) show that source repetition can be detected by integrating information over time and that the auditory system can perform sound segregation when it is able to recover the target sound's latent structure. Motivated by these findings, we trained our dendritic model with a learning rule that was previously demonstrated to detect and analyze the temporal structure of a stream of signals. In particular, we relied on the learning rule described by Asabuki and Fukai (2020), which is based on the minimization of regularized information loss. Specifically, such a principle enables the self-supervised learning of recurring temporal features in information streams using a family of competitive networks of somatodendritic neurons. However, while the learning rule has been designed to capture temporal information in an online fashion, in our framework we flatten the spectrogram before encoding it, making the spike pattern static during the stimulus presentation. Therefore, the temporal fluctuations are determined by the stochastic processes in the rate encoding step.

We presented the network with temporally overlapping sounds following the same task protocols as described in McDermott et al. (2011). First, we carried out the segregation task with the same dataset of synthesized sounds presented to human listeners in McDermott et al. (2011). We found that the model was able to segregate sounds only when one of the masker sounds varied, not when both sounds of the mixture were repeated.

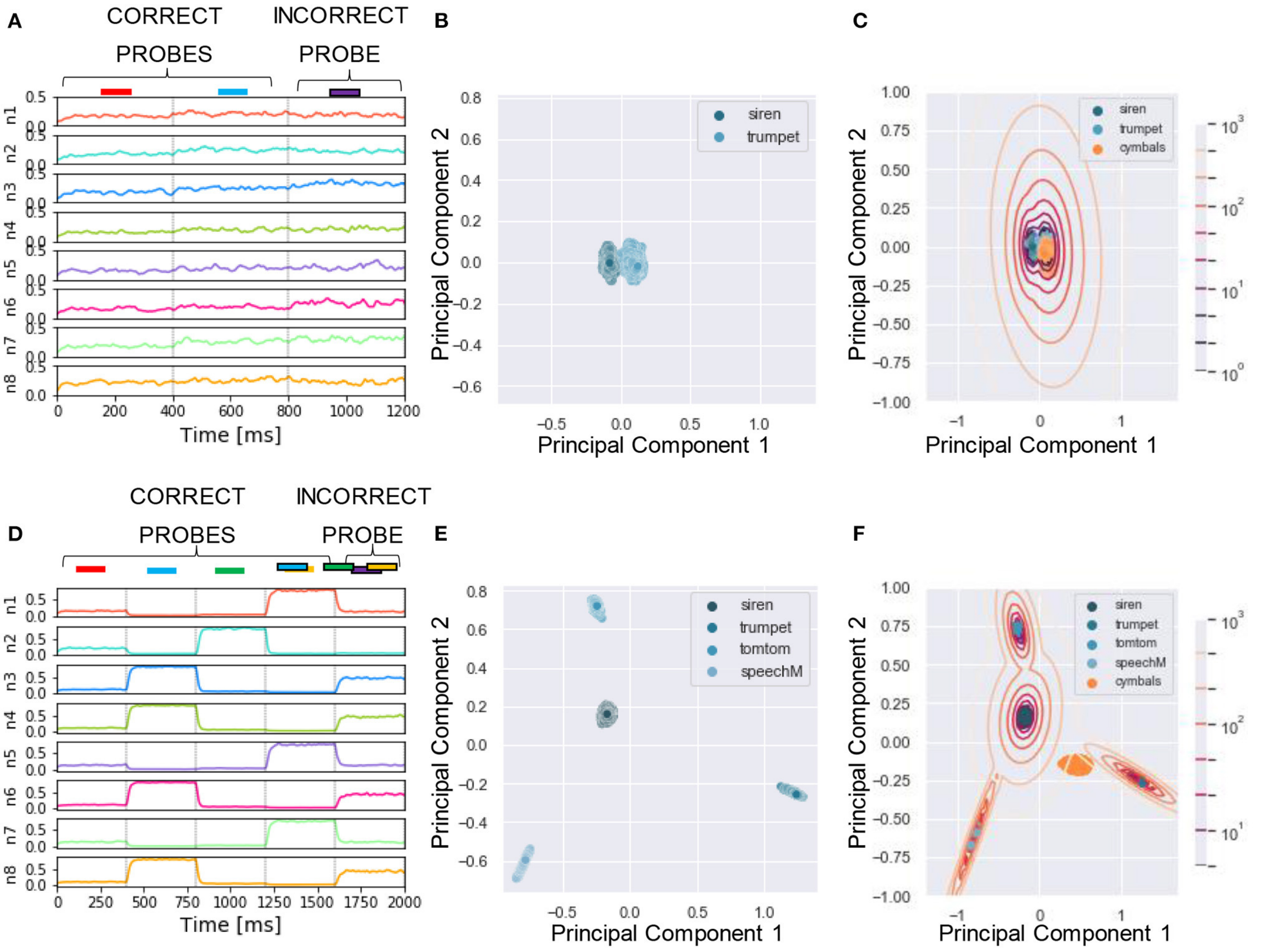

FIGURE 8 | Experiment 4—output dynamics and clustering. **(A–C)** refer to the results of Experiment 4 on real-world sounds with a single mixture presented during training. **(D–F)** refer to the results of Experiment 4 on real-world sounds with three mixtures presented during training. **(A)** Voltage dynamics of the 8 output neurons during inference, when the target, the distractor and one unseen sound are tested. As expected, the neuron population is not able to respond with different dynamics to the three sounds, and the voltage of all the output neurons fluctuates randomly throughout the whole testing sequence. **(B)** The PCA projection of the datapoints belonging to the target and distractor (in blue) shows that the clusters are collapsed into a single cluster. **(C)** When GMM is applied, all the datapoints representing both the learnt sounds (in blue) and the unseen sound (in orange) fall within the same regions, making it impossible to distinguish the different sounds based on the population dynamics. **(D)** Voltage dynamics of the 8 output neurons during inference, when the target, the three distractors, and one unseen sound are tested. As expected, the neuron population has learnt the feature of the different sounds and responds with different dynamics to the five sounds. Each output neuron has an enhanced response to one or few sounds. **(E)** The PCA projection of the datapoints belonging to the four correct probes (in blue) shows that the clusters are more compact and more spatially distant one from the other with respect to the result obtained with the synthetized sounds. **(F)** When GMM is applied, the model shows that the network clearly distinguished the learnt sounds (in blue) from the unseen sound (in orange). These results show that the grouping cues improve the model accuracy with respect to the synthesized dataset.

Our findings bear a closer resemblance to the experimental findings of human listeners over a variety of task settings. Earlier works have proposed biologically inspired networks to perform BSS (Pehlevan et al., 2017; Isomura and Toyoizumi, 2019; Bahroun et al., 2021). However, to our knowledge, this is the first attempt to reproduce the experimental results of recovering sound sources through embedded repetition. For this reason, we could not compare our results with previous works. Additionally, we demonstrated that our network can be a powerful tool for predicting the dynamics of brain segregation capabilities under settings difficult to test on humans. In particular, the recovery of natural sounds is expected to be a trivial task for humans given their familiarity with the sounds, whereas our model is built from scratch and has no prior knowledge about natural sounds. We find that the hallmarks of natural sounds make the task easier for the network when the target is mixed with different sounds, but, as for the synthetic dataset, the sounds cannot be detected if presented always in the same mixture. Furthermore, we extended the study to investigate BSS of visual stimuli and observed a similar qualitative performance as in the auditory settings. This is not surprising from a computational perspective as the computational steps

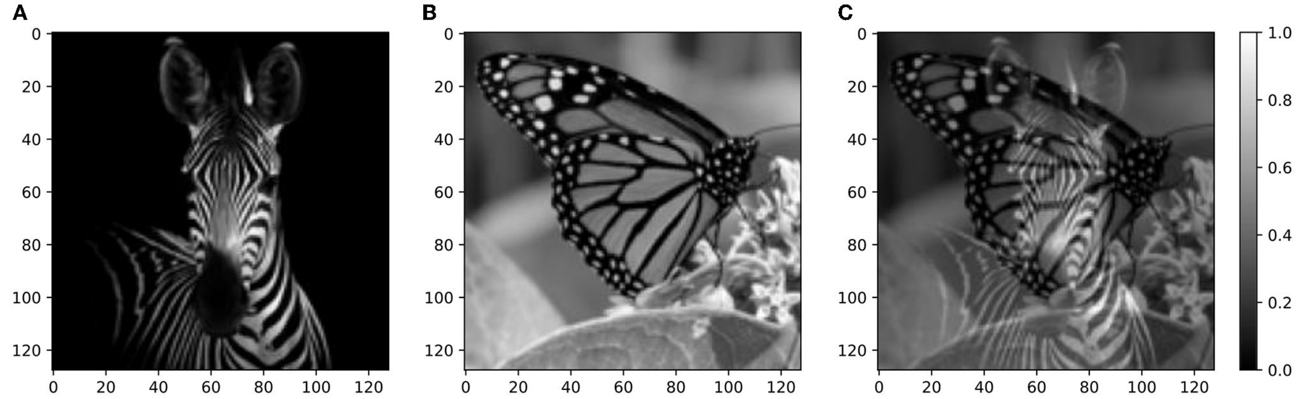

FIGURE 9 | Real-world images—targets and mixture. **(A)** Squared 128 × 128 target image of a zebra. **(B)** Squared 128 × 128 distractor image of a butterfly. **(C)** Mixture of the target and distractor images shown in **(A,B)**. Source: Shutterstock.

of the visual experiment are the same as for the acoustic experiment: there, the sounds are first preprocessed into images, the spectrograms, and then presented to the network in a visual form. From the biological point of view, the neural computational primitives used in the visual and the auditory cortex may be similar, as evidenced by anatomical similarity and by developmental experiments where auditory cortex neurons acquire V1-like receptive fields when visual inputs are redirected there (Sharma et al., 2000; Bahroun et al., 2021). We point out, however, that such a similarity is valid only at high level as there are some substantial differences between visual and auditory processing. For instance, the mechanisms to encode the input signal into spikes rely on different principles: in the retina the spike of a neuron indicates a change in light in the space it represents, while in the cochlea the rate of a neurons represents the amplitude of the frequency it is associated to, like a mechanical FFT. Motivated by these reasons, we suggest extending the experiments of source repetition to vision to verify experimentally whether our computational results provide a correct prediction of the source separation dynamics of the visual system.

Although the dynamics of our model under many aspects matches the theory of repetition-based BSS, the proposed scheme presents a few limitations. The major limitation concerns the discrepancy of the results in experiment 2B. In such a setting, the model performance is well above chance, although the target sound always occurs in the same mixture. We speculate that, in this task settings, the output neurons learn the temporal structure of the distractor sounds presented outside the mixture and that they recognize some similarities in the latent structure of the probes. We note that the degree of similarity among distractors is the same as in the psychophysics experiment. This pushes the neurons to respond differently to the correct and incorrect probes, thereby allowing the output classifier to distinguish the sounds. In contrast, we speculate that human auditory perception relies also on the outcome of the later integration of features detected at early processing stages. This will prevent the misperception of sounds based on unimportant latent features. A second limitation of the selected encoding method consists in the difficulty to model the experiments relying on the asynchronous overlapping of signals and on reversed probe sounds presented by McDermott et al. (2011). Indeed, in our approach, because of the flattening of the spectrogram in the encoding phase, each input neuron responds to one specific time frame, and the output neurons are trained uniquely on this configuration. Hence, temporal shifts or inverting operations are not possible. Third, we observed that in Experiment 1, as the number of mixtures increased over a certain threshold, the model's accuracy degraded. We speculate that, in such settings, substituting PCA with a clustering algorithm not relying on dimensionality reduction, such as K-means, may help mitigate the issue. In addition, an interesting variation of our framework would be replacing the clustering step of the model with an another layer of spiking neurons. Fourth, the flattening of the spectrogram in the spike encoding stage is not biologically plausible and introduces high latency as the entire input signal needs to be buffered before the encoding starts. This strategy exhibits the advantage of making the length of the spike train fixed for any sound length, though modifications of the encoding scheme that preserves the signal's temporal structure might be more suitable for applications tailored for real-world devices. Furthermore, an instantaneous identity coding approach, either from raw signal or *via* a spectrogram, would not be affected by the previously described issues related to the spectrogram normalization in the presence of outliers in signal intensity. Motivated by these points, in a follow up work we intend to explore an extension of the presented framework combining time frame-dependent encoding and spike-based post-processing clustering, which would allow us to integrate the model in embedded neuromorphic applications for sound source separation with reduced response latency. In this context, for further lowering the temporal latency, as well as for reducing the model's energy consumption in neuromorphic devices, the time-to-first-spike

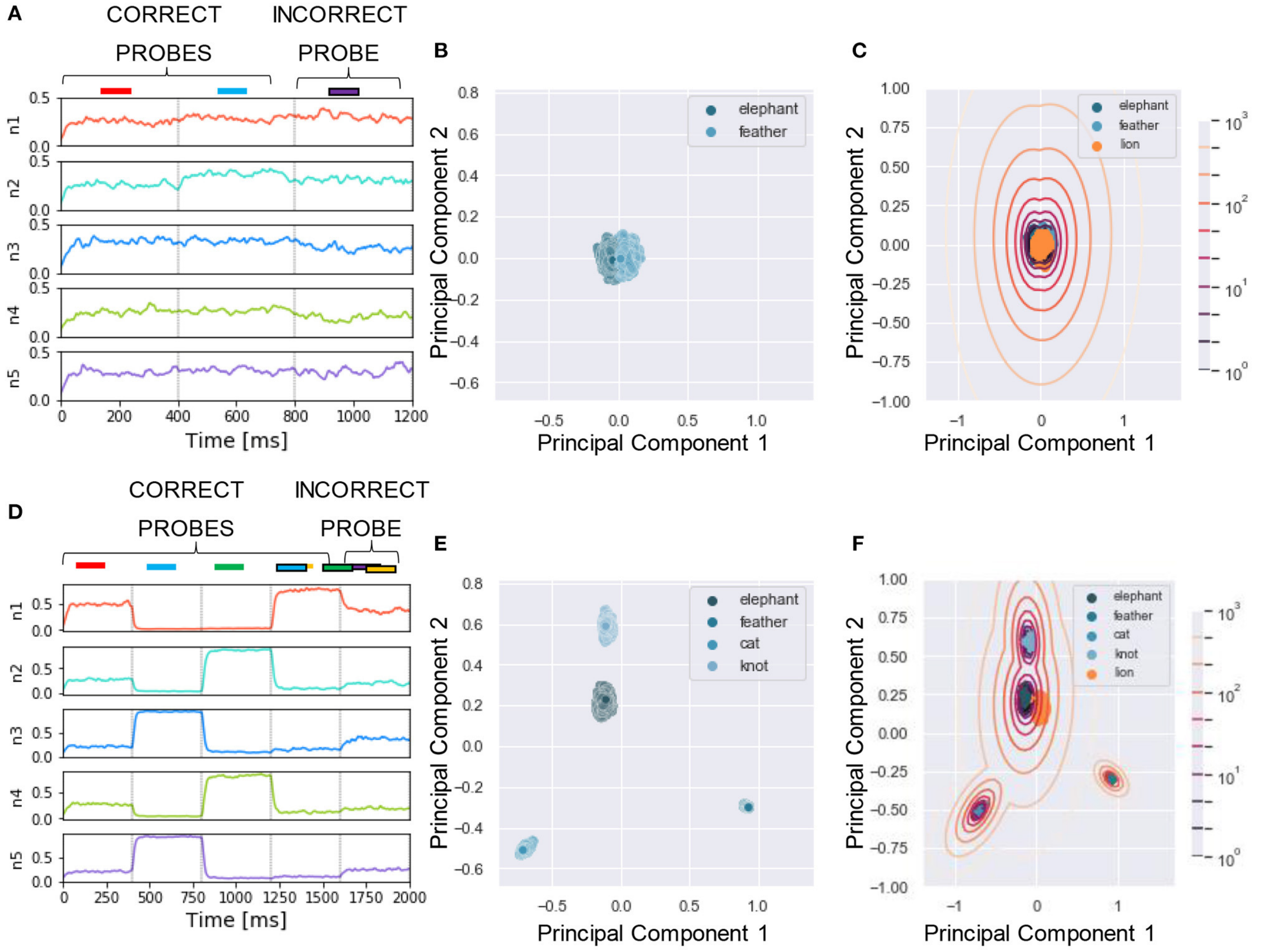

FIGURE 10 | Experiment 5—output dynamics and clustering. **(A–C)** refer to the results of Experiment 5 on real-world images with a single mixture presented during training. **(D–F)** refer to the results of Experiment 5 on real-world images with three mixtures presented during training. **(A)** Voltage dynamics of the 5 output neurons during inference, when the two training images and one unseen image are tested. As expected, the neuron population is not able to respond with different dynamics to the three images, and the voltage of all the output neurons fluctuates randomly throughout the whole testing sequence. **(B)** The PCA projection of the datapoints belonging to the two seen images (in blue) shows that the clusters are collapsed into a single cluster. **(C)** When GMM is applied, all the datapoints representing both the targets (in blue) and the unseen image (in orange) fall within the same regions, making it impossible to distinguish the different images based on the population dynamics. **(D)** Voltage dynamics of the 5 output neurons during inference, when the four targets and one unseen image are tested. As expected, the neuron population has learnt the features of the different images and responds with different dynamics to the five images. Each output neuron has an enhanced response to one or few inputs. **(E)** The PCA projection of the datapoints belonging to the four learnt images (in blue) shows that the clusters are compact and spatially distant one from the other. **(F)** When GMM is applied, the model shows that the network clearly distinguished the target and distractors (in blue) from the unseen image (in orange). These results suggest that humans would be able to distinguish single visual targets if previously seen in different mixtures.

encoding method could be explored as an alternative to the current rate coding approach.

Furthermore, as previously mentioned, the training scheme in Asabuki and Fukai (2020) has proven to be able to learn temporal structures in a variety of tasks. In particular, the model was shown to perform chunking as well as to achieve BSS from mixtures of mutually correlated signals. We underline that our computational model and experiments differ in fundamental ways from the BSS task described by Asabuki and Fukai (2020). First, the two experiments diverge in their primary scope. The BSS task aims at using the average firing rate of the single neurons responding to sound mixtures to decode separately the original sounds. In our work, instead, sound mixtures are included only in the training sequence and, during inference, only individual sounds are presented to the network. Our goal is to verify from the population activity whether the neurons have effectively learned the sounds and can distinguish them from unseen distractors. Furthermore, in Asabuki and Fukai (2020) the stimulus was encoded into spike patterns using one Poisson process proportional to the amplitude of the sound waveform at each time step, disregarding the signal intensity at different frequencies. This method was not suitable for the source segregation through repetition task, where the sound mixtures retain important information on the frequency features of the original sounds at each time frame. Furthermore, we flatten the audio signal spectrogram before encoding it, unlike in the BSS task described by Asabuki and Fukai (2020).

In summary, we have shown that a network of dendritic neurons trained in an unsupervised fashion is able to learn the features of overlapping sounds and, once the training is completed, can perform blind source separation if the individual sounds have been presented in different mixtures. These results account for the experimental performance of human listeners tested on the same task setting. Our study has demonstrated that a biologically inspired simple model of the auditory system can capture the intrinsic neural mechanisms underlying the brain's capability of recovering individual sound sources based on repetition protocols. Furthermore, as the adopted learning scheme in our model is local and unsupervised, the network is self-organizing. Therefore, the proposed framework opens up new computational paradigms with properties specifically suited for embedded implementations of audio and speech processing tasks in neuromorphic hardware.

4. MATERIALS AND METHODS

4.1. Datasets

A dataset of synthesized sounds were created in the form of spectrogram, which shows how signal strength evolves over time at various frequencies, according to the method described previously (McDermott et al., 2011). In short, the novel spectrograms were built as Gaussian distributions based on correlation functions analogous to those of real-world sounds. White noise was later applied to the resulting spectrograms. Five Gaussian distributions were employed to generate each of ten different sounds in **Figure 5A**. The corresponding spectrograms featured 41 frequency filters equally spaced on an ERBN (Equivalent Rectangular Bandwidth, with subscript N denoting normal hearing) scale (Glasberg and Moore, 1990) spanning 20–4,000 Hz, and 33 time frames equally dividing the 700 ms sound length. For our simulations, we used the same MATLAB toolbox and parameters used in the previous study (McDermott et al., 2011). For further details on the generative model for sounds, please refer to the SI Materials and Methods therein.

In addition to the dataset of synthesized sounds, we built a database composed of 72 recordings of isolated natural sounds. The database contained 8 recordings of human speech from the EUSTACE (the Edinburgh University Speech Timing Archive and Corpus of English) speech corpus (White and King, 2003), 23 recordings of animal vocalizations from the Animal Sound Archive (Frommolt et al., 2006), 29 recordings of music instruments by Philharmonia Orchestra (Philarmonia Orchestra Instruments, 2019), and 12 sounds produced by inanimate objects from the BBC Sound Effect corpus (BBC, 1991). The sounds were cut into 800 ms extracts. Then the library librosa (McFee et al., 2015) was employed to extract spectrograms with 128 frequency filters spaced following the Mel scale (Stevens et al., 1937) and 10 ms time frames with 50% overlap.

For image source separation, we built a database consisting of 32 black-and-white pictures of various types, both single objects and landscapes. The images were later squared, and their size was reduced to 128 × 128 pixels.

4.2. Neuron Model

In this study we used the same two-compartment neuron model as that developed previously (Asabuki and Fukai, 2020). The mathematical details are found therein. Here, we only briefly outline the mathematical framework of the neuron model. Our two-compartment model learns temporal features of synaptic input given to the dendritic compartment by minimizing a regularized information loss arising in signal transmission from the dendrite to the soma. In other words, the two-compartment neuron extracts the characteristic features of temporal input by compressing the high dimensional data carried by a temporal sequence of presynaptic inputs to the dendrite onto a low dimensional manifold of neural dynamics. The model performs this temporal feature analysis by modifying the weights of dendritic synapses to minimize the time-averaged mismatch between the somatic and dendritic activities over a certain recent interval. In a stationary state, the somatic membrane potential of the two-compartment model could be described as an attenuated version of the dendritic membrane potential with an attenuation factor (Urbanczik and Senn, 2014). Though we deal with time-dependent stimuli in our model, we compare the attenuated dendritic membrane potential with the somatic membrane potential at each time point. This comparison, however, is not drawn directly on the level of the membrane potentials but on the level of the two non-stationary Poissonian spike distributions with time-varying rates, which would be generated if both soma and dendrite were able to emit spikes independently. In addition, the dynamic range of somatic responses needs to be appropriately rescaled (or regularized) for meaningful comparison. An efficient learning algorithm for this comparison can be derived by minimizing the Kullback–Leibler (KL) divergence between the probability distributions of somatic and dendritic activities. Note that the resultant learning rule enables unsupervised learning because the somatic response is fed back to the dendrite to train dendritic synapses. Thus, our model proposes the view that backpropagating action potentials from the soma may provide a supervising signal for training dendritic synapses (Larkum et al., 1999; Larkum, 2013).

4.3. Network Architecture

The network architecture, shown in **Figure 1**, consisted of two layers of neurons, either fully connected or with only 30% of the total connections. The input layer contained as many Poisson neurons as the number of pixels present in the input spectrogram (acoustic stimulus) or input image (visual stimulus). The postsynaptic neurons were modeled according to the two-compartment neuron model proposed previously (Asabuki and Fukai, 2020). Their number was varied from a pair to few tenths, depending on the complexity of the task. Unless specified otherwise, 8 and 5 output neurons were set for acoustic and visual task respectively.

In the first layer, the input was encoded into spikes through a rate coding-based method (Almomani et al., 2019). The strength of the signal at each pixel drove the firing rate of the associated input neuron, i.e., the spike trains were drawn from Poisson point processes with probability proportional to the intensity of the pixel. We imposed that, for each input stimulus, the spike pattern was generated through a sequence of 400 time steps, where each time step corresponds to a "fire" or "non-fire" event.

We designed the output layer and the learning process similarly to the previous network used for the blind signal separation (BSS) within mixtures of multiple mutually correlated signals as well as for other temporal feature analyses (Asabuki and Fukai, 2020). As mentioned previously, the learning rule was modeled as a self-supervising process, which is at a conceptual level similar to Hebbian learning with backpropagating action potentials. The soma generated a supervising signal to learn and detect the recurring spatiotemporal patterns encoded in the dendritic activity. Within the output layer, single neurons learned to respond differently to each input pattern. Competition among neurons was introduced to ensure that different neurons responded to different inputs. With respect to the network used for BSS containing only two output neurons, we rescaled the strength of the mutual inhibition among dendritic neurons by a factor proportional to the inverse of the square root of the number of output neurons. This correction prevented each neuron from being too strongly inhibited when the size of the output layer increased (i.e., exceeds three or four). Furthermore, we adopted the same inhibitory spike timing-dependent plasticity (iSTDP) as employed in the previous model. This rule modified inhibitory connections between two dendritic neurons when they coincidently responded to a certain input. The iSTDP allowed the formation of chunk-specific cell assemblies when the number of output neurons was greater than the number of input patterns.

For all parameters but noise intensity ξ_i during learning, we used the same values as used in the original network model (Asabuki and Fukai, 2020). For bigger values of noise intensity g, the neural responses were subject to more fluctuations and neurons tended to group in only one cell assembly. From the analysis of the learning curves shown in **Figure 3**, we decided to train the network from randomly initialized weights and to expose it, during training, to the mixture sequence 3,000 times for the synthesized sounds and 1500 times for the real-world sounds. The learning rate was kept constant throughout the whole process. During testing, the sequence of target sounds and respective distractors was presented 50 times, and the resulting neural dynamics was averaged over 20 trials. The performance results shown in the section 2 were computed as average over 10 repetitions of the same simulation set-up. In each repetition different target sounds and distractors were randomly sampled from the dataset in order to ensure performance independence of specific sounds.

4.4. Experimental Settings and Performance Measure

The synapses were kept fixed during inference in our network, implying that the responses to probes tested later were not affected by the presentation of other previously tested probes. This allowed us to test the trained network on a sequence of probes, rather than only on one probe as in the studies of the human brain where plasticity cannot be frozen during inference (McDermott et al., 2011). In **Figures 5A**, **6C**, the first half of the sequence contained the target and the distractors, the second half the respective incorrect probes, which were also built by using the same method as in human experiment (McDermott et al., 2011). Each incorrect probe was a sound randomly selected from the same Gaussian distribution generating the associated target. After the sampling, a randomly selected time slice equal to 1/8 of the sound duration was set to be equal to the target.

The possibility of presenting more than one probe allowed us to test the performance of the network for all the sounds present in the mixtures. To ensure a stable neural response against the variability of the encoding, we repeated the sequence 50 times. The response of the network consisted of the ensemble activity of the output neurons. As previously explained, 400 time-steps were devoted to the presentation to each stimulus. The response to each probe, therefore, consisted of 400 data points describing the dynamical activity of each output neuron, each point being a collection of N values, where N is the number of output neurons. An example of one testing epoch output is shown in **Figures 4A,C**. We neglected the first 50 data points, since, during the initial transient time, the membrane potential was still decaying or rising after the previous input presentation. For visualization purpose, we applied the principal component analysis (PCA) to reduce the dimensionality of the data from N to 2. In our settings, the two principal components explain approximately 40% of the variance of the neural response. The PCA transformation was based uniquely on the data points obtained with the presentation of the target and the distractors, as shown in **Figures 4B,E**. The same transformation was later exploited to project the points related to the incorrect probes. Only the target and distractors patterns were presented during the learning process, and the responses to unseen patterns were afterwards projected on the space defined by the training.

The two-dimensional projection of the target-related data points were clustered in an unsupervised manner through GMM. We set the number of Gaussians equal to the number of targets such that the covariance matrices had a full rank. With the defined GMM model at hand, we proceeded with fitting all the PCA data points, related to both correct and incorrect probes. The model tells which cluster each data point belonged to and what was the likelihood (L) that the cluster had generated this data point. **Figures 4C,F** show the datapoints projected on the PCA plane together with the GMM clustering and likelihood curves.

We used the likelihood as a measure of performance. The four intervals of the likelihood range corresponding to the four responses "sure no," "no," "yes," and "sure yes" were (i) $L > 0$ (sure yes), (ii) $-5 < L < 0$ (yes), (iii) $-15 < L < -5$ (no), and (iv) $L < -15$ (sure no). In building the receiver operating characteristic (ROC) curve, the datapoints falling in the interval (i) were assigned the probability value 1.0, those in (ii) 0.66, those in (iii) 0.33, and those in (iv) 0.0.

The described evaluation metrics was applied only to the experiments carried on the dataset composed of synthesized sounds. For the experiments based on natural sounds and images, the results of clustering were shown only qualitatively for the target-related datapoints. Indeed, due to the real-world nature of signals, it was not possible to simply use Gaussian functions to build physically consistent incorrect probes. On the real-world sound dataset, we performed all the same protocol of Experiment 1 (Experiment 4). On the image dataset we performed an experiment with a protocol analogous to Experiment 1. Here, the mixtures were obtained by overlapping two images, both with transparency 0.5, similarly to the spectrogram overlapping described for the acoustic task. The input images were normalized to the range [0,1] and the intensity of each pixel was encoded through the firing rate of one input neuron. We followed the same procedure and network setting described for the audio stimuli segregation to assess the ability of the network to separate visual stimuli presented in mixtures.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/GiorgiaD/dendritic-neuron-BSS.

AUTHOR CONTRIBUTIONS

TF, GD, and TA conceived the idea. GD designed and performed the simulations, with input from TA. GD and TF wrote the manuscript. TA and GD wrote the **Supplementary Material**. All authors analyzed the results. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partly supported by JSPS KAKENHI no. 19H04994 to TF.

ACKNOWLEDGMENTS

We are grateful to all the colleagues in the Neural Coding and Brain Computing Unit for fruitful interaction.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.855753/full#supplementary-material

REFERENCES

Ahveninen, J., Hämäläinen, M., Jääskeläinen, I. P., Ahlfors, S. P., Huang, S., Lin, F.-H., et al. (2011). Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4182–4187. doi: 10.1073/pnas.1016134108

Almomani, D., Alauthman, M., Alweshah, M., Dorgham, O., and Albalas, F. (2019). A comparative study on spiking neural network encoding schema: implemented with cloud computing. *Cluster Comput.* 22, 419–433. doi: 10.1007/s10586-018-02891-0

Amari, S., Cichocki, A., and Yang, H. (1995). "A new learning algorithm for blind signal separation," in *NIPS'95: Proceedings of the 8th International Conference on Neural Information Processing Systems* (Cambridge, MA), 757–763.

Asabuki, T., and Fukai, T. (2020). Somatodendritic consistency check for temporal feature segmentation. *Nat. Commun.* 11, 1554. doi: 10.1038/s41467-020-15367-w

Asari, H., Pearlmutter, B. A., and Zador, A. M. (2006). Sparse representations for the cocktail party problem. *J. Neurosci.* 26, 7477–7490. doi: 10.1523/JNEUROSCI.1563-06.2006

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K., et al. (2018). Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. *Neuron* 97, 640.e4–655.e4. doi: 10.1101/098798

Bahroun, Y., Chklovskii, D. B., and Sengupta, A. M. (2021). A normative and biologically plausible algorithm for independent component analysis. *arXiv [Preprint]*. arXiv: 2111.08858. doi: 10.48550/arXiv.2111.08858

BBC. (1991). *BBC sound effects library. Compact disc.; Digital and Analog Recordings.; Detailed Contents on Insert in Each Container.;Recorded: 1977–1986.* Princeton, NJ: Films for the Humanities and Sciences.

Bee, M., 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. Psychol.* 122, 235–251. doi: 10.1037/0735-7036.122.3.235

Bell, A., and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* 7, 1129–1159.

Bronkhorst, A. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attent. Percept. Psychophys.* 77, 1465–1487. doi: 10.3758/s13414-015-0882-9

Brown, G., Yamada, S., and Sejnowski, T. (2001). Independent component analysis at neural cocktail party. *Trends Neurosci.* 24, 54–63. doi: 10.1016/S0166-2236(00)01683-0

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25, 975–979.

Cichocki, A., Zdunek, R., and Amari, S. (2006). "New algorithms for non-negative matrix factorization in applications to blind source separation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (Toulouse).

Comon, P. (1994). Independent component analysis, a new concept? *Signal Process.* 36, 287–314.

Ding, N., and Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* 107, 78–89. doi: 10.1152/jn.00297.2011

Dong, J., Colburn, H. S., and Sen, K. (2016). Cortical transformation of spatial processing for solving the cocktail party problem: a computational model. *eNeuro* 3, 1–11. doi: 10.1523/ENEURO.0086-15.2015

Elhilali, M. (2013). "Bayesian inference in auditory scenes," in *Conference Proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Osaka), 2792–2795.

Elhilali, M., and Shamma, S. (2009). A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *J. Acoust. Soc. Am.* 124, 3751–3771. doi: 10.1121/1.3001672

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* 3, 128–135. doi: 10.1016/S1364-6613(99)01294-2

Frommolt, K. -H., Bardeli, R., Kurth, F., and Clausen, M. (2006). *The Animal Sound Archive at the Humboldt-University of Berlin: Current Activities in Conservation and Improving Access for Bioacoustic Research*. Ljubljana: Slovenska akademija znanosti in umetnosti.

Glasberg, B. R., and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* 47, 103–138.

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *J. Neurosci.* 33, 1417–1426. doi: 10.1523/JNEUROSCI.3675-12.2013

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. *J. Acoust. Soc. Am.* 115, 833–843. doi: 10.1121/1.1639908

Haykin, S., and Chen, Z. (2005). The cocktail party problem. *Neural Comput.* 17, 1875–1902. doi: 10.1162/0899766054322964

Hyvärinen, A., and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Comput.* 9, 1483–1492.

Isomura, T., and Toyoizumi, T. (2019). Multi-context blind source separation by error-gated Hebbian rule. *Sci. Rep.* 9, 7127. doi: 10.1038/s41598-019-43423-z

Jacobsen, T., Schröger, E., Winkler, I., and Horváth, J. (2005). Familiarity affects the processing of task-irrelevant auditory deviance. *J. Cogn. Neurosci.* 17, 1704–1713. doi: 10.1162/089892905774589262

Kameoka, H., Li, L., Inoue, S., and Makino, S. (2018). Semi-blind source separation with multichannel variational autoencoder. *arXiv preprint arXiv:1808.00892*. doi: 10.48550/arXiv.1808.00892

Karamatli, E., Cemgil, A. T., and Kirbiz, S. (2018). "Weak label supervision for monaural source separation using non-negative denoising variational autoencoders," in *2019 27th Signal Processing and Communications Applications Conference (SIU)* (Sivas).

Kerlin, J., Shahin, A., and Miller, L. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *J. Neurosci.* 30, 620–628. doi: 10.1523/JNEUROSCI.3631-09.2010

Krause-Solberg, S., and Iske, A. (2015). "Non-negative dimensionality reduction for audio signal separation by NNMF and ICA," in *2015 International Conference on Sampling Theory and Applications, SampTA 2015* (Washington, DC), 377–381.

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

Larkum, M. (2013). A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends Neurosci.* 36, 141–151. doi: 10.1016/j.tins.2012.11.006

Larkum, M., Zhu, J., and Sakmann, B. (1999). A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature* 398, 338–341.

Lewald, J., and Getzmann, S. (2015). Electrophysiological correlates of cocktail-party listening. *Behav. Brain Res.* 292, 157–166. doi: 10.1016/j.bbr.2015.06.025

Li, Y., Wang, F., Chen, Y., Cichocki, A., and Sejnowski, T. (2017). The effects of audiovisual inputs on solving the cocktail party problem in the human brain: an fMRI study. *Cereb. Cortex* 28, 3623–3637. doi: 10.1093/cercor/bhx235

Liu, Q., Huang, Y., Hao, Y., Xu, J., and Xu, B. (2021). LiMuSE: Lightweight multi-modal speaker extraction. *arXiv [Preprint]*. arXiv: 2111.04063.

López-Serrano, P., Dittmar, C., Özer, Y., and Müller, M. (2019). NMF toolbox: music processing applications of nonnegative matrix factorization.

McDermott, J. H. (2009). The cocktail party problem. *Curr. Biol.* 19, R1024–R1027. doi: 10.1016/j.cub.2009.09.005

McDermott, J. H., Wrobleski, D., and Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1188–1193. doi: 10.1073/pnas.1004765108

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., et al. (2015). "librosa: Audio and music signal analysis in Python," in *Proc. of the 14th Python in Science Conf. (SCIPY 2015)* (Austin), 18–24.

Mesgarani, N., and Chang, E. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–236. doi: 10.1038/nature11020

Middlebrooks, J. C., and Waters, M. F. (2020). Spatial mechanisms for segregation of competing sounds, and a breakdown in spatial hearing. *Front. Neurosci.* 14, 571095. 10.3389/fnins.2020.571095

Mika, D., Budzik, G., and Józwik, J. (2020). "ICA-based single channel source separation with time-frequency decomposition," in *2020 IEEE 7th International Workshop on Metrology for AeroSpace (MetroAeroSpace)* (Pisa), 238–243.

Narayan, R., Best, V., Ozmeral, E., McClaine, E., Dent, M., Shinn-Cunningham, B., et al. (2008). Cortical interference effects in the cocktail party problem. *Nat. Neurosci.* 10, 1601–1607. doi: 10.1038/nn2009

O'Sullivan, J., Power, A., Mesgarani, N., Rajaram, S., Foxe, J., Shinn-Cunningham, B., Slaney, M., et al. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Oxenham, A. J. (2018). How we hear: the perception and neural coding of sound. *Annu. Rev. Psychol.* 69, 27–50. doi: 10.1146/annurev-psych-122216-011635

Pehlevan, C., Mohan, S., and Chklovskii, D. B. (2017). Blind nonnegative source separation using biological neural networks. *Neural Comput.* 29, 2925–2954. doi: 10.1162/neco_a_01007

Philarmonia Orchestra Instruments. (2019). Available online at: https://philharmonia.co.uk/resources/instruments/

Popham, S., Boebinger, D., Ellis, D., Kawahara, H., and McDermott, J. (2018). Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nat. Commun.* 9, 2122. doi: 10.1038/s41467-018-04551-8

Sagi, B., Nemat-Nasser, S. C., Kerr, R., Hayek, R., Downing, C., and Hecht-Nielsen, R. (2001). A biologically motivated solution to the cocktail party problem. *Neural Comput.* 13, 1575–1602. doi: 10.1162/089976601750265018

Santosh, K. S., and Bharathi, S. H. (2017). "Non-negative matrix factorization algorithms for blind source sepertion in speech recognition," in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)* (Bangalore), 2242–2246.

Sawada, H., Ono, N., Kameoka, H., Kitamura, D., and Saruwatari, H. (2019). A review of blind source separation methods: two converging routes to ilrma originating from ICA and NMF. *APSIPA Trans. Signal Inform. Process.* 8, 1–14. doi: 10.1017/ATSIP.2019.5

Schmidt, A. K. D., and Römer, H. (2011). Solutions to the cocktail party problem in insects: selective filters, spatial release from masking and gain control in tropical crickets. *PLoS ONE* 6, e28593. doi: 10.1371/journal.pone.0028593

Sharma, J., Angelucci, A., and Sur, M. (2000). Induction of visual orientation modules in auditory cortex. *Nature* 404, 841–847. doi: 10.1038/35009043

Smaragdis, P., and Brown, J. (2003). "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY), 177–180.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *J. Acoust. Soc. Am.* 8, 185–190.

Teki, S., Chait, M., Kumar, S., Shamma, S., and Griffiths, T. D. (2013). Segregation of complex acoustic scenes based on temporal coherence. *eLife* 2, e00699. doi: 10.7554/eLife.00699.009

Thakur, C., Wang, R., Afshar, S., Hamilton, T., Tapson, J., Shamma, S., et al. (2015). Sound stream segregation: a neuromorphic approach to solve the "cocktail party problem" in real-time. *Front. Neurosci.* 9, 309. doi: 10.3389/fnins.2015.00309

Urbanczik, R., and Senn, W. (2014). Learning by the dendritic prediction of somatic spiking. *Neuron* 81, 521–528. doi: 10.1016/j.neuron.2013.11.030

White, L., and King, S. (2003). *The Eustace Speech Corpus*. Centre for Speech Technology Research, University of Edinburgh.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. New York, NY: Oxford University Press.

Woods, K. J. P., and McDermott, J. H. (2018). Schema learning for the cocktail party problem. *Proc. Natl. Acad. Sci. U.S.A.* 115, E3313–E3322. doi: 10.1073/pnas.1801614115

Xiang, J., Simon, J., and Elhilali, M. (2010). Competing streams at the cocktail party: exploring the mechanisms of attention and temporal integration. *J. Neurosci.* 30, 12084–12093. doi: 10.1523/JNEUROSCI.0827-10.2010

Yu, D. (2020). "Solving cocktail party problem–from single modality to multi-modality," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)* (Virtual workshop).

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.