LETTER **Communicated by Sam Neymotin**

Inferring Mechanisms of Auditory Attentional Modulation with Deep Neural Networks

Ting-Yu Kuo
kvq941@alumni.ku.dk
Department of Computer Science and Technology, State Key Laboratory of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China; and Faculty of Humanities, Copenhagen University, Copenhagen 2300, Denmark

Yuanda Liao
liaoyd20@mails.tsinghua.edu.cn
Department of Computer Science and Technology, State Key Laboratory of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China, and Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing 100084, China

Kai Li
qhu.kaili@gmail.com
Department of Computer Science and Technology, State Key Laboratory of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

Bo Hong
hongbo@tsinghua.edu.cn
Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing 100084, China; IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing 100084, China; and Department of Biomedical Engineering, Tsinghua University, Beijing 100084, China

Xiaolin Hu
xlhu@tsinghua.edu.cn
Department of Computer Science and Technology, State Key Laboratory of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China; Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing 100084, China; IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing 100084, China; and Chinese Institute for Brain Research, Beijing 100010, China

Bo Hong and Xiaolin Hu are the corresponding authors.

Neural Computation **34, 1–21** (2022)
https://doi.org/10.1162/neco_a_01537

Humans have an exceptional ability to extract specific audio streams of interest in a noisy environment; this is known as the cocktail party effect. It is widely accepted that this ability is related to selective attention, a mental process that enables individuals to focus on a particular object. Evidence suggests that sensory neurons can be modulated by top-down signals transmitted from the prefrontal cortex. However, exactly how the projection of attention signals to the cortex and subcortex influences the cocktail effect is unclear. We constructed computational models to study whether attentional modulation is more effective at earlier or later stages for solving the cocktail party problem along the auditory pathway. We modeled the auditory pathway using deep neural networks (DNNs), which can generate representational neural patterns that resemble the human brain. We constructed a series of DNN models in which the main structures were autoencoders. We then trained these DNNs on a speech separation task derived from the dichotic listening paradigm, a common paradigm to investigate the cocktail party effect. We next analyzed the modulation effects of attention signals during all stages. Our results showed that the attentional modulation effect is more effective at the lower stages of the DNNs. This suggests that the projection of attention signals to lower stages within the auditory pathway plays a more significant role than the higher stages in solving the cocktail party problem. This prediction could be tested using neurophysiological experiments.

1 Introduction

The ability of humans to separate audio streams of interest under an acoustically challenging environment is important for daily life and is termed the cocktail party effect (Cherry, 1953). The overlapping nature of the various sounds in spectrotemporal space makes the speech separation task computationally challenging (Brungart, Simpson, Ericson, & Scott, 2001; Zion Golumbic et al., 2013). Although it is widely accepted that selective attention enables animals to segregate and regroup overlapping audio streams (Alain, 2000; Bregman & McAdams, 1994), scientists have not reached a consensus regarding the impact of attention on different processing stages. Several studies have suggested that attentional modulation reaches the early stages when information is processed within the sensory cortex. One study found that the impact of attention can be traced along the neuronal activity of the cochlea (Maison, Micheyl, & Collet, 2001). Consistent with this study, animal research has revealed that separation of the sound stream occurs before the cortical stage (Nakamoto, Jones, & Palmer, 2008; Slee & David, 2015). Furthermore, studies on human speech processing have indicated that attention enhances neural processing prior to the cortex stage (Price & Bidelman, 2021; Rinne et al., 2008).

However, numerous studies have also indicated that both the primary and secondary auditory cortices play a major role in the speech separation task (Ding & Simon, 2012; Fritz, Shamma, Elhilali, & Klein, 2003; Mesgarani & Chang, 2012; Zion Golumbic et al., 2013). Therefore, these findings suggest that cortical and subcortical areas carry distinct responsibilities during the speech separation task (Deutsch & Deutsch, 1963; Wittekindt, Kaiser, & Abel, 2014). Exactly how attention affects different auditory processing stages remains to be explored.

Many studies have compared artificial neural networks and real brains and found that well-trained deep neural networks (DNNs) excel at predicting the neural responses of living creatures (Banino et al., 2018; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Kietzmann, McClure, & Kriegeskorte, 2019; Prokott et al., 2021; Yamins & DiCarlo, 2016). A previous study (Khaligh-Razavi & Kriegeskorte, 2014) showed that DNNs trained on object recognition tasks demonstrate rivaling neural representations in the inferior temporal cortex. Moreover, a study on mice navigation (Banino et al., 2018) found that a neural network trained on data collected from living mice showed a similar pattern as that of the mouse hippocampus. Some recent studies have extended these studies by inferring real brain functions through the dissection of neural networks that show similar task performance as that of humans (Prokott, Tamura, & Fleming, 2021; Saiz-Alía & Reichenbach, 2020).

In our study, we used a series of DNN models to investigate attentional modulation mechanisms. The backbone of the DNN was a multilayered autoencoder. The task is to separate two individual speeches from their mixture with selective attention (see Figure 1A). An autoencoder consists of an encoder and a decoder connected by shortcut pathways. This structure was inspired by the Unet (Ronneberger, Fischer, & Brox, 2015). Within the encoder and decoder, the basic building blocks were ResBlocks (He, Zhang, Ren, & Sun, 2016), ResTranspose, and downsampling convolutional layers. Using these DNNs, we aimed to determine which stages in the auditory pathway play a more significant role in solving the cocktail party effect. Our supervised deep convolutional neural networks were optimized toward solving the dichotic listening task, a classical experimental paradigm that is widely used for studying selective attention. In the first stage, we trained the model to demonstrate performance on a speech separation task that rivals that of electrophysiological experiments (Mesgarani & Chang, 2012). Subsequently, we compared the effects of attentional modulation on different stages. To exclude the effects of confounding factors, we trained a series of DNN models using varied metaparameters.

2 Method

2.1 Model Structure. Our model consisted of three components: a deep autoencoder (ResUnet), a speaker feature extraction network (FeatureNet),

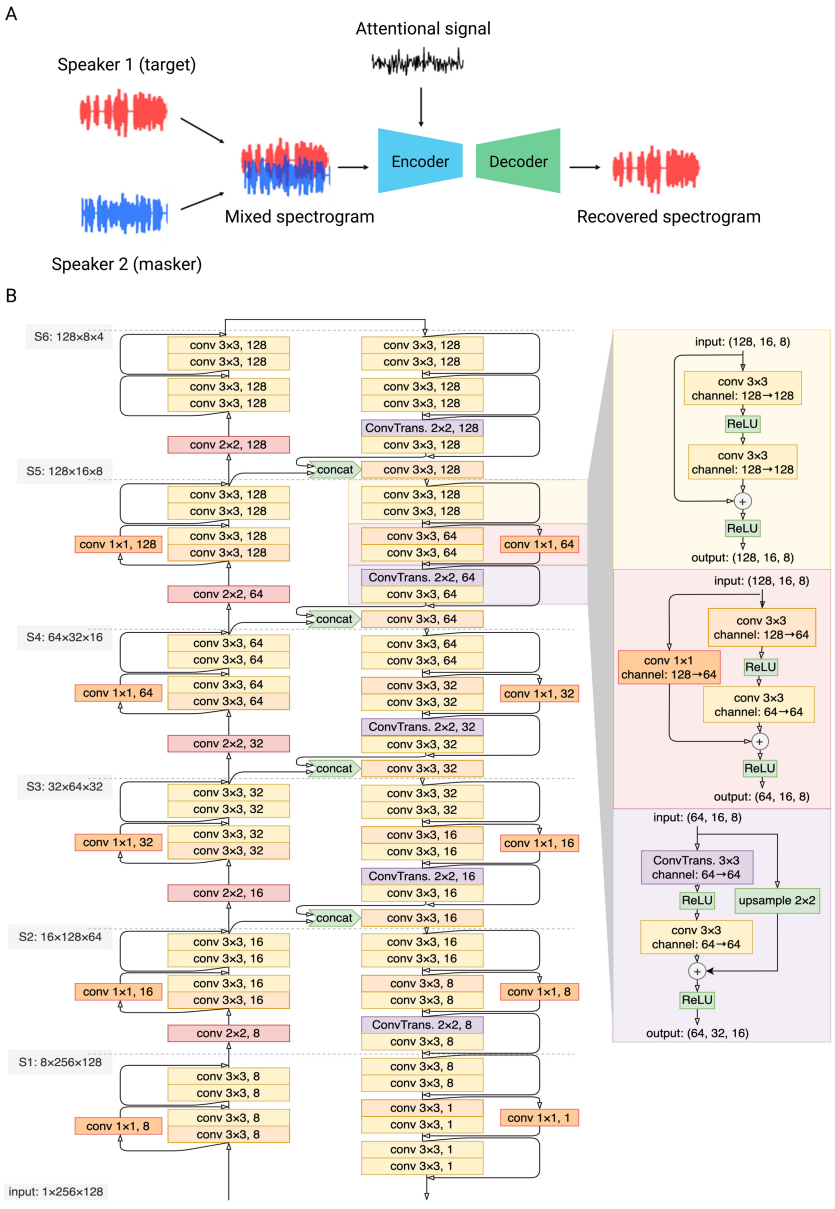

Figure 1: Description of the task and model structure. (A) Illustration of the task. (B) Model structure of a six-stage ResUnet example. S1 to S6: stages in the encoder. S1′ to S6′: the corresponding stages in the decoder. The shape of each stage's output is labeled next to each stage and is presented as channel × height × width. The right part of the model illustrates the three basic

and a network simulating attentional modulation (AttentionNet). Figure 1B illustrates the model structure and basic modules. The encoder was designed to correspond to the ascending auditory pathway and was aimed at learning the hierarchical representation of speech. Inputs were downsampled and reshaped repeatedly and then passed on to the decoder. The structure of the decoder was similar to the encoder and was designed to decode the target spectrogram, as described in a previous human physiology study (Mesgarani & Chang, 2012). It did not necessarily correspond to any biological structures. Although the decoder is not the focus of this letter, we considered the possibility that the effect of speech recovery might vary depending on the decoder structure. We designed multiple decoder variants to make the experiments more compact and used the clean target spectrogram as the ground truth for recovery.

2.2 ResUnet. ResUnet was the main structure in our model and is an encoder-decoder model adapted from ResNet and Unet (Ronneberger et al., 2015). The stages of the encoder were denoted as S_i, $i = 1, 2, \ldots, n$, and the stages of the decoder were denoted as S'_i, $i = 1, 2, \ldots, n$. The encoder and decoder were designed to be structurally symmetrical. Each of the residual stages in the encoder consisted of a downsampling 2D convolutional layer (except S_1, which does not downsample), k ResBlocks ($k = 1, 2, 3$), and a pooling layer.

A decoder is designed to decode feature maps at various encoder stages and recover input spectrograms. Each decoder stage includes multiple ResBlocks (denoted as m, where $m = 0, k$) and one transposed convolutional layer (except S'_1) to increase the size of the feature maps.

Inspired by the previous work of ResNet (He et al., 2016), we defined ResBlock and ResTranspose modules as the two basic building blocks of our model. The ResBlock module has a convolutional pathway and a residual pathway. For the convolutional pathway, we first applied a convolutional layer, which was followed by batch normalization and a rectified linear unit (ReLU) operation. The second convolutional layer and batch normalization were subsequently applied.

The output of the residual pathway could be a copy of the feature maps or an input tensor with a 1×1 convolution filter. If the output has the same

components applied to the model: ResBlock, ResBlock (preserved channel), and ResTranspose. The "conv 3×3, [ch]" flag denotes a convolutional layer with kernel size = (3, 3), stride = 1, padding = 1, output channel = ch; the "conv 2×2, [ch]" flag denotes a downsize convolution with kernel size = (2, 2), stride = 2, padding = 0, output channel = ch; the "transpose 2×2 [ch]" flag denotes a transpose layer with kernel size = (2, 2), stride = 2, padding = 0, output channel = ch; the "conv 1×1 [ch]" flag denotes a convolutional layer with kernel size = (1, 1), stride = 1, padding = 0, output channel = ch.

Table 1. Structure of the FeatureNet.

LSTM Layer	FC Layer
$\begin{pmatrix} LSTM \\ 256, 768 \end{pmatrix} \times 3$	$\begin{pmatrix} FC \\ 768, 256 \end{pmatrix}$

The structure of the FeatureNet has three long short-term memory (LSTM) modules followed by one fully connected (FC) layer. The numbers specify the number of input and output channels.

channel number as that of the input, the residual pathway goes with the former condition. However, if the output has a different channel number from that of the input, the residual pathway goes with the latter condition. The outputs of the two pathways are then added and rectified using ReLU, which results in the final output of the ResBlock.

Similar to the ResBlock, the ResTranspose module has dual pathways. In the convolutional pathway, we replaced the first convolution with a transposed convolutional layer to upsample the input tensor. The transposed convolution was followed by batch normalization and a ReLU operation. In the residual pathway, the input feature map was upsampled by a scale factor of 2. Channels remained equal for both the input and output of the ResTranspose module. The structures of the ResBlock and ResTranspose are described in Figure 1B.

2.3 FeatureNet. The structure of the FeatureNet was inspired by previous work (Wan, Wang, Papir, & Moreno, 2018). It included three long short-term memory network (LSTM; Sundermeyer, Schlüter, & Ney, 2012) modules and a fully connected (FC layer), which converted the output from LSTM into speech embeddings (256 in our experiment). Inputs for the FeatureNet first underwent a mel-frequency transformation to obtain better audio feature representations. The FeatureNet was trained using the cross-entropy loss function. Parameters of an example FeatureNet are provided in Table 1.

2.4 AttentionNet. The purpose of the AttentionNet was to convert the audio features of the target speaker into attention signals and project the converted signals to the encoder stages in ResUnet. The AttentionNet receives a single vector input from the FeatureNet and outputs n attention signal vectors. The n outputs were obtained using n FC layers. All FC layers took the speech feature vector as the input. For instance, in a model with $n = 6$, six independent FC layers would all receive the same embedded feature from the FeatureNet and output n attention signals. Each attention

vector matches the output feature map at each corresponding stage along the channel and frequency dimensions. For example, if the S_2 encoder stage outputs $128 \times 64 \times 16$-dimensional (using a spectral $\times$ temporal $\times$ channel-notation manner) feature maps, it would receive signals with a dimension of $128 \times 1 \times 16$.

2.5 Training Procedures. The model was changed with three procedures. First, we trained the FeatureNet to extract auditory features from the target speaker. The FeatureNet was trained on a binary classification task that was aimed at differentiating between the target speaker and the masker speaker. After training, we extracted the last hidden layer as a vectorized representation of the target's auditory features. Second, we trained the ResUnet on the clean speech data set. Here, the goal of the ResUnet was to learn to recover an input spectrogram with minimal distortion by minimizing the loss function $\sum_x \left|\left|x - \text{decoder}(\text{encoder}(x))\right|\right|_2^2$, where x denotes a training speech sample represented in the form of a spectrogram. Finally, we trained the full model with all three components, with the mixed input, $x = x_1 + x_2$, where x_1 and x_2 represent the speech of the two speakers. If the attention signal a_i was applied to speaker i, we minimized the loss function $\sum_x \left|\left|x_i - \text{decoder}(\text{encoder}_{a_i}(x))\right|\right|_2^2$, where $\text{encoder}_{a_i}(x)$ stands for the output of the encoder with attention a_i.

3 Experimental Settings

Our experimental setting was the dichotic listening paradigm, a commonly adapted experimental framework for cocktail party effect studies (Cherry, 1953; Ding & Simon, 2012; Mesgarani & Chang, 2012; O'Sullivan et al., 2019; Woldorff et al., 1993; Zion Golumbic et al., 2013). The experimenter broadcast voices from two separate channels: speaker 1 (SP1) and speaker 2 (SP2). Usually one channel is assigned as the target and the other as the masker, and participants are required to attend to one of the speech channels while simultaneously receiving input from both audio channels. We generated our data set by overlapping the voices from SP1 and SP2. The goals for our models were to separate the voices between the target and the masker and reconstruct a clean spectrogram without the masker's voice.

The source code is publicly available at https://github.com/liaoyd16/cocktail_lk.

3.1 Data Set Generation. We generated a clean speech data set from multiple recordings of a selected male and female. The recordings were segmented into 4 second clips and transformed into 256×128 sized spectrograms as the model input. To obtain the spectrogram, we applied a short-time Fourier transform and a base10 logarithm to the spectrogram. The frequency domain of 200 Hz to 3000 Hz was retained to maintain focus

on the human listening domain. Approximately 92.6% of the data (3239 samples) was used for training, and the remaining data were used for testing (240 samples). One clip from SP1 and another from SP2 were left out for the generation of speaker embeddings, which were the inputs for the FeatureNet.

For the dichotic listening paradigm, we created an additional data set of mixed speech. Each data entry was obtained by directly combining two spectrograms. The spectrograms were randomly sampled from the pool of audio clips.

All speech materials were taken from the LibriVox recording (LibriVox, 2014).

3.2 Performance Evaluation. To evaluate recovery performance, we used the mean squared error distance between the original and reconstructed spectrograms as the loss function of the models. We also calculated the attentional modulation index (AMI), as proposed in a previous study (Mesgarani & Chang, 2012), to quantify the effects of attention. AMI was calculated using the correlations of the target, masker, and reconstructed spectrograms:

$$\begin{aligned} AMI_{spec} = {} & Corr\big(SP1_{spec}, SP1_{att}\big) - Corr\big(SP1_{spec}, SP2_{att}\big) \\ & + Corr\big(SP2_{spec}, SP2_{att}\big) - Corr\big(SP2_{spec}, SP1_{att}\big), \end{aligned} \tag{3.1}$$

where $SP1_{spec}$ and $SP2_{spec}$ represent the original acoustic spectrograms of SP1 and SP2, respectively. The reconstructed spectrograms are denoted by $SP1_{att}$ and $SP2_{att}$. A positive AMI value indicates that the recovered spectrogram is more similar to the target than to the masker. A higher AMI value indicates a greater similarity between the recovered spectrogram and the target.

4 Results

We trained multiple DNNs to perform the dichotic listening task, similarly to the electrophysiological study in humans (Mesgarani & Chang, 2012) (see section 2). The DNNs had autoencoder structures that encoded the inputs, which generated outputs of the same dimensions as those of the inputs. The input data contained mixed spectrograms of the two speakers, SP1 and SP2 (see Figure 1A). The goal of the task was to differentiate the masker speakers from the target speakers. The key to the success of the models was the selective attention signals that were applied to the encoder stages. We then compared the contributions of the attention projected onto different encoder stages and investigated the influence of model architectures on the results.

4.1 Selective Recovery of Spectrograms. We conducted audio separation experiments on our trained model, the scheme of which resembled the dichotic listening tasks designed for human participants (Cherry, 1953; Mesgarani & Chang, 2012). In this task, when attention toward the target speaker was imposed, the output of the model (reconstructed spectrograms) was expected to approximate the original spectrograms of the target speaker.

In each trial, the input was a mixed spectrogram that was generated by overlapping SP1 and SP2 soundtracks and converting them into a spectrogram. SP1 and SP2 took turns being the target and masker speakers, and the corresponding attention signal was provided accordingly. We also collected the recovered spectrograms when there was no attention target assigned. In summary, each trial contained six spectrograms: two original spectrograms (SP1 and SP2), a mixed spectrogram (input to model), a reconstructed spectrogram without attention, and two reconstructed spectrograms while attending to SP1 and SP2. The original spectrograms were generated by 4 second sound clips.

Among the spectrograms of each trial, comparisons were made between the original and the reconstructed SP1-SP2 pairs. A sample (sample ID 9) containing the original, mixed, and reconstructed spectrograms is displayed in Figure 2. The original spectrograms of SP1 and SP2 are displayed in Figures 2A and 2B. The mixed spectrogram (see Figure 2C) demonstrates that the experimental setting was acoustically challenging since the spectrotemporal features from the two audio sources are highly overlapping. The superposed energy contours of the two spectrograms (see Figure 2D) confirmed the difficulty in separating overlapping spectrograms. For the nonattended conditions, the reconstructed spectrogram resembled the mixed spectrogram (see Figure 2E). By contrast, for the attended conditions, the reconstructed spectrograms of SP1 and SP2 (see Figures 2F and 2G) resembled the original spectrograms (see Figures 2A and 2B).

For the reconstruction, we calculated the correlation between the reconstructed spectrograms and the target and masker spectrograms for SP1 and SP2, respectively. Correlations were calculated for both attending and nonattending conditions. For the attending trials, the reconstructed spectrograms (of both SP1 and SP2) demonstrated a higher correlation with targets and a lower correlation with maskers, whereas those of the nonattending trials showed no such preference in general (see Figure 3A). In fact, depending on whether attention was focused on SP1 or SP2, the correlation was scattered either above or below the diagonal line, which indicated that the correlation between the targets and the reconstructed spectrograms was significantly higher than that between the maskers and the reconstructed spectrograms. In addition, the recovered spectrograms of the nonattending trials failed to show a preference for either SP1 or SP2. For SP1-attending trials, 85.42% of the data was spread underneath the diagonal line, which indicated a higher similarity to the original SP1 spectrograms.

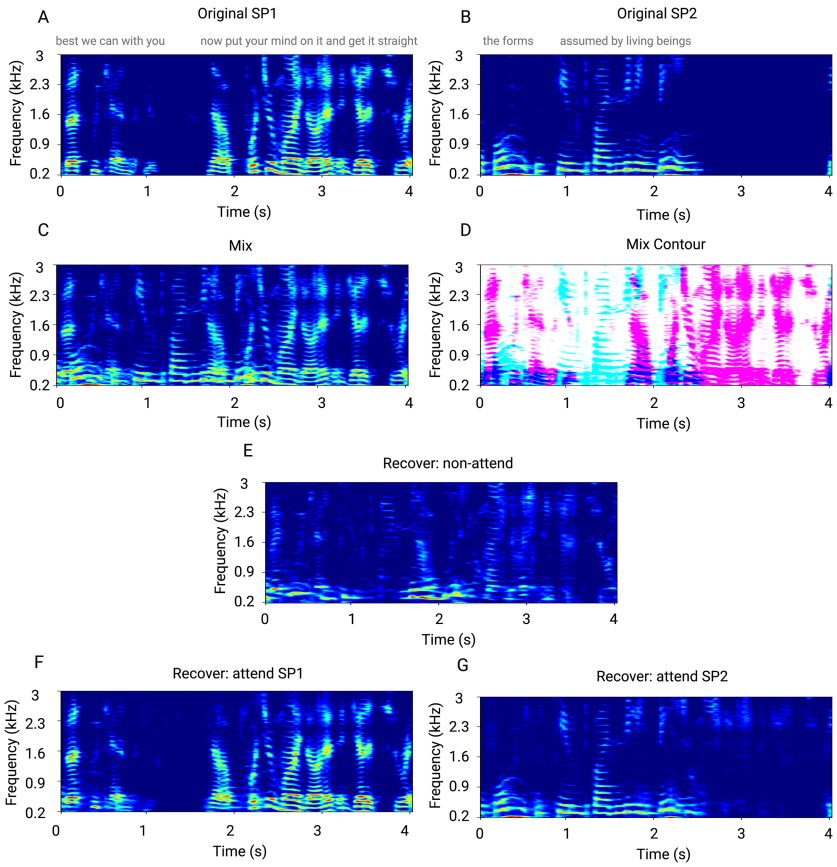

Figure 2: A sample of selective recovery of spectrograms. Each experimental trial included original spectrograms from target and masker, a mixed spectrogram, and three recovered spectrograms under different attention situations. The graph illustrates sample 9. We labeled the text in corresponding locations on spectrograms of SP1 and SP2. The contour illustration (D) was used only to demonstrate the overlap of energy contours and was not used in the experiments.

However, for SP2-attending trials, 97.08% of the data were spread above the diagonal line, which indicated the same pattern as that for SP1-attending trials. Despite the slight preference toward SP2, the overall correlation distribution of both speakers indicated an obvious attentional modulation effect because the reconstructed spectrograms were highly correlated with the original spectrograms. The high correlation between the reconstructed spectrograms and the original target spectrograms (0.794 for SP1 and 0.838

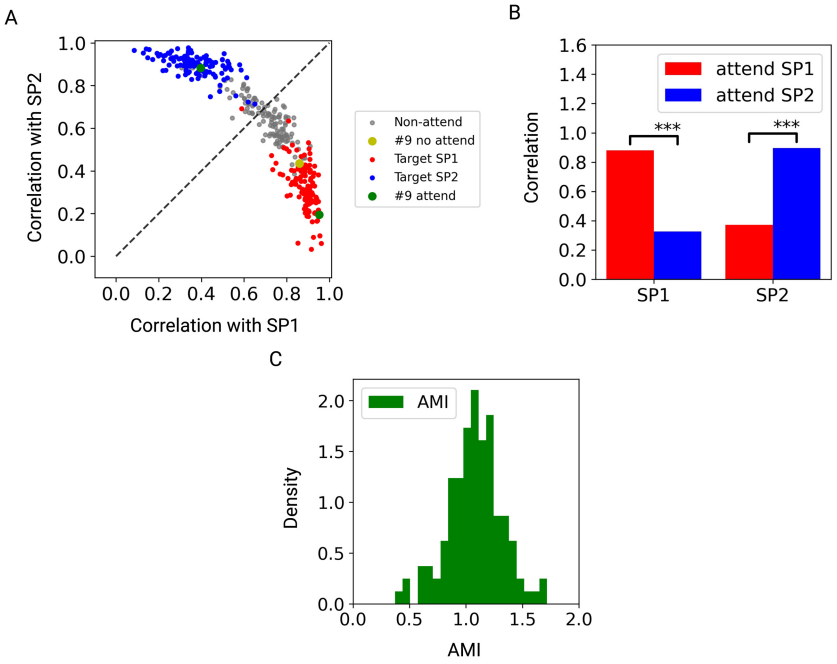

Figure 3: Distributions of spectrogram correlation coefficients and attentional modulation index (AMI). Analysis of 240 testing trials. Each trial contained six spectrograms: original speaker 1 (SP1), original speaker 2 (SP2), mixed, recovery spectrograms for nonattended trials, attended SP1, and attended SP2. (A) Correlation between the recovered spectrogram and the spectrograms of SP1 and SP2 speech, under the attended SP1 (red), attended SP2 (blue), and nonattention (gray) conditions. Each point denotes the correlation between the recovered spectrogram and SP1 speech (horizontal axis) and SP2 speech (vertical axis) of one condition. $N = 240$ for each condition. The yellow and green dots denote different conditions of sample 9 shown in Figure 2. (B) Comparisons of the mean correlations of consistent and inconsistent speakers. Left: Correlations between reconstructed spectrograms and the spectrograms of SP1 in SP1-attending trials and SP2-attending trials. Right: Correlations between reconstructed spectrograms and the spectrograms of SP2 in SP1-attending trials and SP2-attending trials. $^{***}p < 0.01$, Student's t-test. (C) Distribution of the AMI when applying attention to all stages. $N = 240$. Results were obtained from the model with $n = 6$, $k = 3$, and $m = 3$.

for SP2) reinforces the observations described above (see Figure 3B). To further confirm that attentional modulation helped to recover spectrograms that were closer to the target spectrograms, we compared the mean correlations between the reconstructed and masker spectrograms (see Figure 3B).

Results showed that the mean correlation between the reconstructed and target spectrograms was higher than the mean correlation between the reconstructed and masker spectrograms, regardless of the speaker who was being attended to ($p < 0.01$, Student's t-test).

Finally, based on previous electrophysiological research on the cocktail party effect (Mesgarani & Chang, 2012), we used a scalar metric called the AMI to evaluate the selectivity of the attention introduced (see section 2). The AMI is calculated by adding the correlations between the recovered spectrograms and their target spectrograms, then subtracting the correlations between the recovered spectrograms and their masker spectrograms. Thus, an AMI larger than 0 indicates that the recovered spectrogram is more similar to the target spectrogram and less similar to the masker spectrogram. We calculated AMI values for all trials (see Figure 3C) and found that most values were clearly larger than 0 (mean $= 0.663$, SD $= 0.373$), which indicated the emergence of the attentional effect. Our results are in line with a previous study conducted in humans (Mesgarani & Chang, 2012).

4.2 Contributions of Attention Signals at Different Encoder Stages. To examine the contributions of attention signals at each encoder stage, we manipulated the attention projections in two ways: by applying attention signals to multiple stages in a gradual manner and to signals to one stage at a time.

The first condition was aimed at investigating the consecutive change in correlation between the recovered and the original spectrograms. Specifically, we explored whether there was a sudden change at a certain stage or whether any obvious tendencies of change in correlation occurred. We applied attention starting from the highest encoder stage (only this stage receives attention) and finishing at the lowest stage (all stages receive attention). When the highest stage alone received attention, the distribution of correlations between the reconstructed SP1 spectrograms and the original SP1/SP2 spectrograms largely overlapped with the distribution obtained from the reconstructed SP2 spectrogram (see Figure 4A). When the attention signals were applied to additional stages, the two distributions became more separated (see Figures 4B to 4E). We found that the tendency of separation was nonlinear and exhibited a peak at a certain point. The separability between the two distributions remained small when the attention signals were applied only to the higher stages (i.e., stages above S_3). However, the separability changed drastically when attention signals were applied to S_3 and the stages below. This indicated that the lower stages play a major role in selective recovery.

For each attention condition, we also calculated the mean correlations between the recovered spectrograms and the clean SP1/SP2 spectrograms for all trials. The increasing trend of the correlations between the recovered and target spectrograms fit the exponential curves well, and the correlation growth peaked at the lower stages (see Figures 4F and 4G).

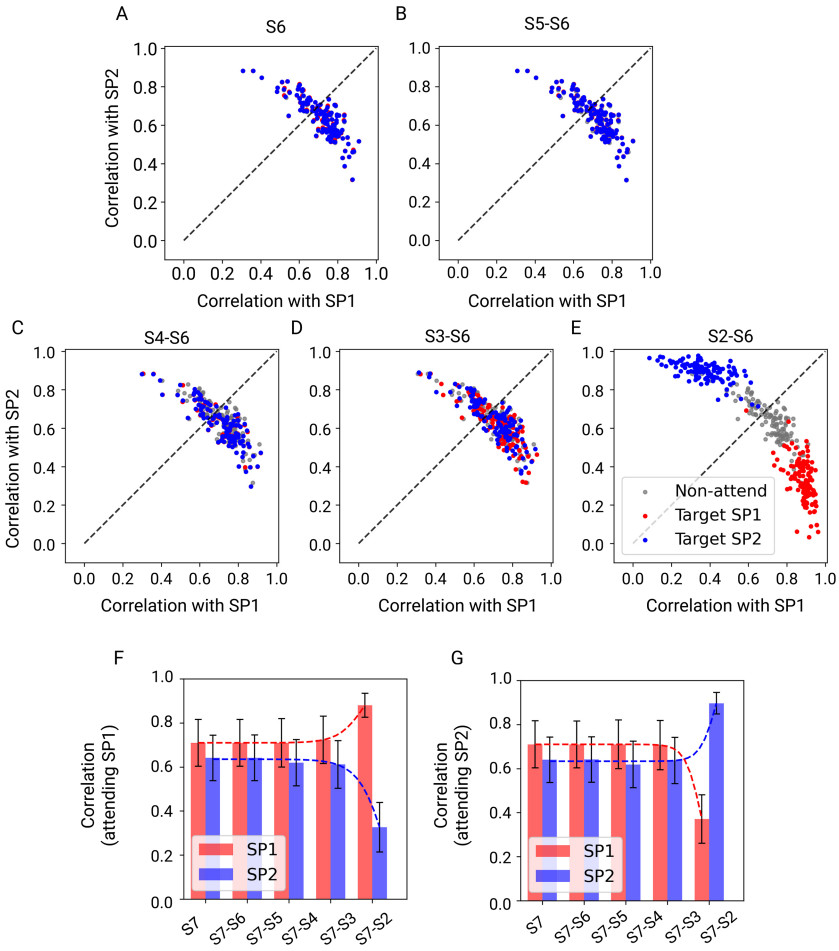

Figure 4: Speech separation results with cumulative attention across stages. (A–E) Scatter plots of the spectrogram correlations for speaker 1 (SP1)-attending and speaker 2 (SP2)-attending conditions. Notations are the same as those in Figure 3A. (F, G) The correlation of recovered spectrograms to original spectrograms when attention was applied to different stages, with the recovery target as SP1 and SP2, respectively. We fit the data using the sigmoid function and plotted it along the dotted lines. The function fitting was measured in R-squared metrics and reached R-squared over 0.999 for all dotted lines.

To eliminate the possibility that these results were caused by the complex interaction between multiple stages when receiving attention signals, we also conducted experiments whereby we applied attention signals to each stage individually. Results showed that the lower stages contributed

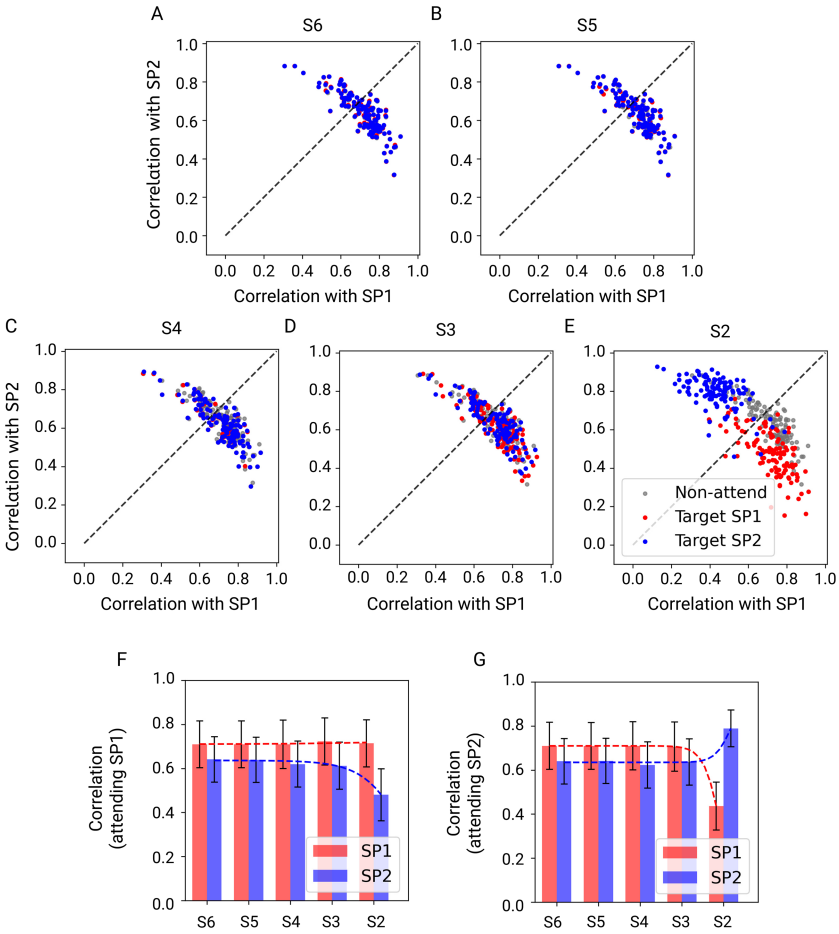

Figure 5: Speech separation results with attention applied to single stages. These figures are plotted in the same fashion as in Figure 4, except that the attention signals were applied to only one stage in each graph.

more to the AMI fluctuation, consistent with the results of the previous experiment (see Figure 5).

4.3 Comparison of Variant Model Structures. To exclude the possibility that the conclusion we have drawn was specific to the choice of model structure, we built variants of the autoencoder, which were trained to perform the same task. Specifically, we altered the number of stages (n) and the number of dimension-preserving ResBlocks (k) in the encoder and the number of ResBlocks (m) in the decoder. We used $n = 4, 6, 9; k = 1, 2, 3$; and

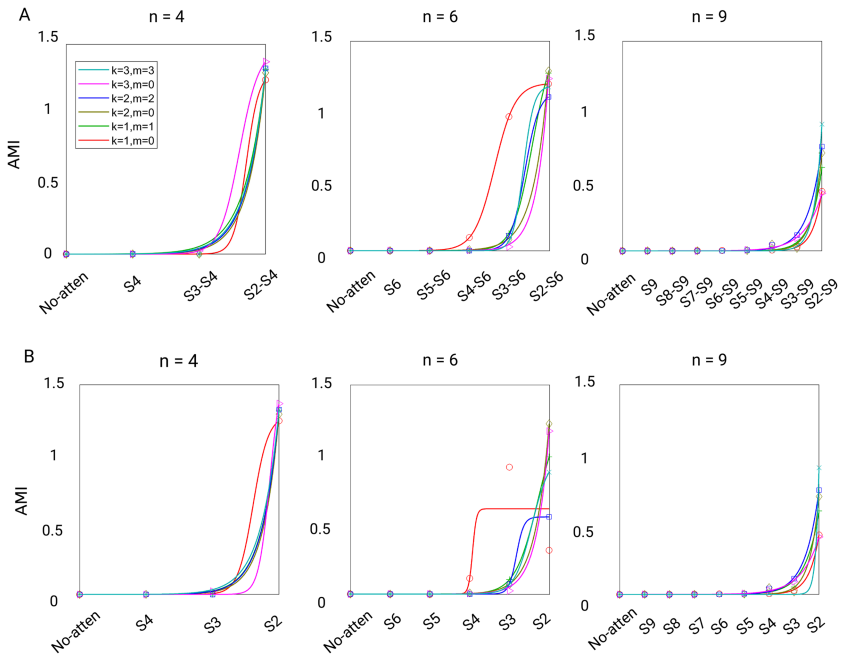

Figure 6: Comparison of attentional modulation index (AMI) changes for different models. The models are parameterized with different m, n, and k. Curves were fit using sigmoid functions. (A) Changes in AMI when applying attention cumulatively in $n = 4, 6$, and 9 models, starting from the highest stage. (B) Changes in AMI when applying attention to a single stage in each experimental trial within $n = 4, 6$, and 9 models, respectively.

$m = 0, k$. For m, we used its two extremes: $m = 0$ (only ResTranspose but no additional ResBlock in any decoder stage) and $m = k$ (ResBlocks in the decoder stage were symmetrical to those in the encoder stage). In total, we trained and tested 18 model structure variants.

We calculated the mean AMI for each model configuration. To examine the incremental contributions to AMI, we applied attention signals gradually, starting from the top. When attention signals were not extended to the lower half of the model, we observed only trivial changes in AMI (see Figure 6A). However, as the attention signals were extended to the lower stages, the mean AMI increased rapidly and reached its peak. In addition, we collected AMIs under conditions where attention was projected onto a single stage (see Figure 6B). Under different model structure variations (with different stage numbers and encoder designs), we found that projections onto the lower stages induced a greater AMI increase. By contrast, attention signals projected to the higher stages had only a trivial effect.

Taken together, the earlier stages in the pathway exhibited stronger attentional modulation effects than the later stages and seemed to dominate the cocktail party effect without joint participation of higher attention signals.

5 Discussion and Conclusion

5.1 Conclusion. In this study, we took advantage of artificial models and drew similarities between the hierarchical nature of the auditory neural pathway and hierarchical DNNs. We then compared the impact of attention on the AMI along the auditory pathway. First, we trained a DNN based on the dichotic listening task paradigm until it reached a similar performance as that of humans (Mesgarani & Chang, 2012). We then extended the original model structures by constructing 18 variants. The attention signals were applied either separately or gradually, starting from the highest stage of the models in our experiments. Results showed that the earlier stages had a stronger attentional modulation effect, whereas the later stages had a much weaker modulation effect on the auditory selection process. Our findings drew a similar conclusion as that of previous studies that reported an effect of selective attention on the early stages of auditory processing (Nakamoto et al., 2008; Price & Bidelman, 2021; Rinne et al., 2008; Slee & David, 2015; Woldorff et al., 1993; Zion Golumbic et al., 2013).

Our results may be regarded as a pilot experiment that offers potential other avenues for investigating the impact of attention at different stages of the auditory pathway. A possible experimental scenario is a noninvasive ablation study of attentional effects; a specific region on the auditory pathway could be inhibited selectively using repetitive transcranial magnetic stimulation while the researchers measure the fluctuations in participants' performance. Our results could offer valuable insight into future physiological experiments and aid the formulation of testable hypotheses.

5.2 Biological Plausibility of the Models. In this study, we attempted to infer actual neural mechanisms by probing DNN models. As in previous studies (Fukushima & Miyake, 1982; Hassabis, Kumaran, Summerfield, & Botvinick, 2017; Tai, Socher, & Manning, 2015), we drew from the concepts of neuroscience and applied them to DNN models. Several studies that compared DNN models and real brains have shown that a well-trained DNN model resembles neuronal activity patterns of real brains (Cadieu et al., 2014; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016). For example, Cadieu et al. (2014) conducted a comparative study between representation patterns of multiple DNN models and the inferior temporal (IT) cortex of the macaque during visual object recognition. Results showed that a well-structured model achieves similar task accuracy; moreover, artificial neurons within the model exhibit representations similar to those of biological

neurons. Numerous other studies have also suggested that DNNs exhibit hierarchical processes: lower DNN stages show better prediction of lower-level neural activity, and higher stages better predict activity of higher cortical regions (Eickenberg et al., 2017; Khaligh-Razavi & Kriegeskorte, 2014). These studies demonstrate the potential for DNNs to simulate the brain to accomplish computationally challenging tasks.

In view of the similarity between DNNs and the sensory pathways on the computational level, many works used deep learning models to study the computational principles of the brain. For example, by using an unsupervised deep learning model, Zhang, Hu, Hong, and Zhang (2019) suggested that sparse activity of neurons in different visual regions and nonlinear transformation between these regions can lead to phonetic representation in the superior temporal cortex. For another example, by using a self-supervised deep learning model, Konkle and Alvarez (2022) suggested that the category representation encoded in anterior regions of the ventral visual stream can emerge by learning to represent individual images rather than categories.

In our study, we included features that were determined in previous neurophysiological studies to offer more biologically plausible models. First, experimental evidence has revealed that the auditory pathway is extensively modulated by event-driven attention. Attention signals project onto multiple brain areas in the ascending auditory pathway, from the primary and nonprimary auditory cortices to the cochlea. Second, studies have demonstrated the hierarchical structure of the auditory system, where different functions are employed along the ascending neural pathway, such as auditory parsing, extraction, and transformation (Pérez-González & Malmierca, 2014). Third, we followed the decoding method of electroencephalography signals used in Mesgarani and Chang (2012) and established an encoder-decoder network that resembled the hierarchical encoding process in the actual auditory pathway. According to these findings, we trained the AttentionNet independently from the ResUnet to simulate attentional projections along all encoder stages. It is worth noting that the AttentionNet is merely a simple network with a single FC layer, and thus it may be an oversimplification of attentional signal projections.

Aside from the model structure, our experimental setting was similar to the dichotic listening experiment. The classical dichotic listening task is considered a simplified condition of the multispeaker environment (Cherry, 1953) in which participants receive two separate audio streams in different ears.

5.3 Limitations. Previous studies on auditory attention typically use two types of stimuli: tones and speech (Ding & Simon, 2012; Zion Golumbic et al., 2013). The primary difference between the two is that speech data contain phonemes and semantic context, whereas tones do not. The emergent stage of speech separation may differ depending on the experimental data.

For model training, we used human speech data segmented into 4 second clips and shuffled. The sounds were segmented according to time window instead of phonemes, resulting in phonemes being retained in some speech clips but not in others. Therefore, one might argue that our experimental setting was not practical because the setting was closer to the situation in which participants listen to corrupted speech rather than mixed but intact speech as in typical experimental settings.

Furthermore, biological neurons possess greater complexity than the computational units in deep learning models in terms of structure, such as neural dendrites, lateral connections, and top-down connections. Thus, we used the model solely to infer neuronal representation at the populational level. Future work could include additional features that have been discovered in the actual brain, such as recurrent circuitry structure (Aponte et al., 2021) and cortical microcircuit found in the auditory cortex (Blackwell & Geffen, 2017).

Finally, we employed a supervised learning paradigm rather than an unsupervised learning paradigm. It remains unclear whether the learning behavior of animals is supervised or unsupervised (Hinton & McClelland, 1988; Lillicrap, Santoro, Marris, Akerman, & Hinton, 2020; Whittington & Bogacz, 2019). Previous studies have indicated that supervised learning provides better simulations of neural representations (Banino et al., 2018; Khaligh-Razavi & Kriegeskorte, 2014). Nevertheless, it is worth comparing and validating our results using unsupervised learning models.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (grant 2021ZD0200301), the National Natural Science Foundation of China (grants 62061136001 and 61836014), Brain+X Seed Foundation of Tsinghua-IDG/McGovern Institute, and the Tsinghua-Toyota Joint Research Fund.

References

Alain, C. (2000). Selectively attending to auditory objects. *Frontiers in Bioscience*, *5*(3), 202–212. 10.2741/A505

Aponte, D. A., Handy, G., Kline, A. M., Tsukano, H., Doiron, B., & Kato, H. K. (2021). Recurrent network dynamics shape direction selectivity in primary auditory cortex. *Nature Communications*, *12*(1), 314. 10.1038/s41467-020-20590-6, PubMed: 33436635

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., . . . Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433. 10.1038/s41586-018-0102-6, PubMed: 29743670

Blackwell, J. M., & Geffen, M. N. (2017). Progress and challenges for understanding the function of cortical microcircuits in auditory processing. *Nature Communications*, *8*(1), 2165. 10.1038/s41467-017-01755-2, PubMed: 29255268

Bregman, A. S., & McAdams, S. (1994). Auditory scene analysis: The perceptual organization of sound. *Journal of the Acoustical Society of America*, *95*(2), 1177–1178. 10.1121/1.408434

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America*, *110*(5), 2527–2538. 10.1121/1.1408946

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., . . . DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963. 10.1371/journal.pcbi.1003963, PubMed: 25521294

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*(5), 975–979. 10.1121/1.1907229

Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*(1), 80–90. 10.1037/h0039515, PubMed: 14027390

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. 10.1073/pnas.1205381109

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194. 10.1016/j.neuroimage.2016.10.001, PubMed: 27777172

Fritz, J., Shamma, S., Elhilali, M., & Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, *6*(11), 1216–1223. 10.1038/nn1141, PubMed: 14583754

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In S. Amari & M. A. Arbib (Eds.), *Competition and cooperation in neural nets* (pp. 267–285). Berlin: Springer. 10.1007/978-3-642-46466-9_18

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired artificial intelligence. *Neuron*, *95*(2), 245–258. 10.1016/j.neuron.2017.06.011, PubMed: 28728020

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for Image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). Piscataway, NJ: IEEE. 10.1109/CVPR.2016.90

Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In D. Anderson (Ed.), *Advances in neural information processing systems* (pp. 358–366). College Park, MD: American Institute of Physics.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915. 10.1371/journal.pcbi.1003915, PubMed: 25375136

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. In T. C. Kietzmann, P. McClure, & N. Kriegeskorte (Eds.), *Oxford research encyclopedia of neuroscience*. New York: Oxford University Press. 10.1093/acrefore/9780190264086.013.46

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 491. 10.1038/s41467-022-28091-4, PubMed: 35078981

LibriVox: Free public domain audiobooks. (2014). *Reference Reviews*, *28*(1), 7–8. 10.1108/RR-08-2013-0197

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346. 10.1038/s41583-020-0277-3, PubMed: 32303713

Maison, S., Micheyl, C., & Collet, L. (2001). Influence of focused auditory attention on cochlear activity in humans. *Psychophysiology*, *38*(1), 35–40. 10.1111/1469-8986.3810035, PubMed: 11321619

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. 10.1038/nature11020, PubMed: 22522927

Nakamoto, K. T., Jones, S. J., & Palmer, A. R. (2008). Descending projections from auditory cortex modulate sensitivity in the midbrain to cues for spatial position. *Journal of Neurophysiology*, *99*(5), 2347–2356. 10.1152/jn.01326.2007, PubMed: 18385487

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., . . . Mesgarani, N. (2019). Hierarchical encoding of attended auditory objects in multi-talker speech perception. *Neuron*, *104*(6), 1195–1209.e3. 10.1016/j.neuron.2019.09.007

Pérez-González, D., & Malmierca, M. S. (2014). Adaptation in the auditory system: An overview. *Frontiers in Integrative Neuroscience*, *8*. 10.3389/fnint.2014.00019

Price, C. N., & Bidelman, G. M. (2021). Attention reinforces human corticofugal system to aid speech perception in noise. *NeuroImage*, *235*, 118014. 10.1016/j.neuroimage.2021.118014, PubMed: 33794356

Prokott, K. E., Tamura, H., & Fleming, R. W. (2021). Gloss perception: Searching for a deep neural network that behaves like humans. *Journal of Vision*, *21*(12), 14. 10.1167/jov.21.12.14, PubMed: 34817568

Rinne, T., Balk, M. H., Koistinen, S., Autti, T., Alho, K., & Sams, M. (2008). Auditory selective attention modulates activation of human inferior colliculus. *Journal of Neurophysiology*, *100*(6), 3323–3327. 10.1152/jn.90607.2008, PubMed: 18922948

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. arXiv:1505.04597.

Saiz-Alía, M., & Reichenbach, T. (2020). Computational modeling of the auditory brainstem response to continuous speech. *Journal of Neural Engineering*, *17*(3), 036035. 10.1088/1741-2552/ab970d

Slee, S. J., & David, S. V. (2015). Rapid task-related plasticity of spectrotemporal receptive fields in the auditory midbrain. *Journal of Neuroscience*, *35*(38), 13090–13102. 10.1523/JNEUROSCI.1671-15.2015, PubMed: 26400939

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *Interspeech 2012*, 194–197. 10.21437/Interspeech.2012-65

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (1:1556–1566). Stroudsburg, PA: Association for Computational Linguistics. 10.3115/v1/P15-1150

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4879–4883). Piscataway, NJ: IEEE. 10.1109/ICASSP.2018.8462665

Whittington, J. C. R., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, *23*(3), 235–250. 10.1016/j.tics.2018.12.005, PubMed: 30704969

Wittekindt, A., Kaiser, J., & Abel, C. (2014). Attentional modulation of the inner ear: A combined otoacoustic emission and EEG study. *Journal of Neuroscience*, *34*(30), 9995–10002. 10.1523/JNEUROSCI.4861-13.2014, PubMed: 25057201

Woldorff, M. G., Gallen, C. C., Hampson, S. A., Hillyard, S. A., Pantev, C., D. Sobel, & Bloom, F. E. (1993). Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proceedings of the National Academy of Sciences*, *90*(18), 8722–8726. 10.1073/pnas.90.18.8722

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. 10.1038/nn.4244, PubMed: 26906502

Zhang, Q., Hu, X., Hong, B., & Zhang, B. (2019). A hierarchical sparse coding model predicts acoustic feature encoding in both auditory midbrain and cortex. *PLoS Computational Biology*, *15*(2), e1006766. 10.1371/journal.pcbi.1006766, PubMed: 30742609

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., . . . Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron*, *77*(5), 980–991. 10.1016/j.neuron.2012.12.037, PubMed: 23473326

Received February 17, 2022; accepted June 25, 2022.