Fine-grained Artificial Neurons in Audio-transformers for Disentangling Neural Auditory Encoding

Mengyue Zhou[1], Xu Liu[1], David Liu[2], Zihao Wu[3], Zhengliang Liu[3], Lin Zhao[3]
Dajiang Zhu[4], Lei Guo[1], Junwei Han[1], Tianming Liu[3], Xintao Hu[1]*
[1] School of Automation, Northwestern Polytechnical University
[2] Athens Academy [3]School of Computing, University of Georgia
[4]Department of Computer Science and Engineering,University of Texas at Arlington
{zhou_my,liu_xu}@email.nwpu.edu.cn david.weizhong.liu@gmail.com
{zw63397,zl18864,lin.zhao,tliu}@uga.edu dajiang.zhu@uta.edu
{lguo,jhan,xhu}@nwpu.edu.cn

Abstract

The Wav2Vec and its variants have achieved unprecedented success in computational auditory and speech processing. Meanwhile, neural encoding studies that link representations of Wav2Vec to brain activities have provided novel insights into how auditory and speech processing unfold in the human brain. Most existing neural encoding studies treat each transformer encoding layer in Wav2Vec as a single artificial neuron (AN). That is, the layer-level embeddings are used to predict neural responses. The layer-level embedding aggregates multiple types of contextual attention captured by multi-head self-attention (MSA). Thus, the layer-level ANs lack fine-granularity for neural encoding. To address this limitation, we define the elementary units, i.e., each hidden dimension, as neuron-level ANs in Wav2Vec2.0, quantify their temporal responses, and couple those ANs with their biological-neuron (BN) counterparts in the human brain. Our experimental results demonstrated that: 1) The proposed neuron-level ANs carry meaningful neurolinguistic information; 2) Those ANs anchor to their BN signatures; 3) The AN-BN anchoring patterns are interpretable from a neurolinguistic perspective. More importantly, our results suggest an intermediate stage in both the computational representation in Wav2Vec2.0 and the cortical representation in the brain. Our study validates the fine-grained ANs in Wav2Vec2.0, which may serve as a novel and general strategy to link transformer-based deep learning models to neural responses for probing sensory processing in the brain.

1 Introduction

The Wav2Vec model and its variants (Schneider et al., 2019; Baevski et al., 2020) have achieved superb performance in learning acoustic information representations and on a variety of downstream tasks such as automatic speech recognition. Meanwhile, recent studies that link the computational representations in Wav2Vec to neural responses recorded by functional brain imaging techniques have provided novel insights into the model's interpretability and neural sensory perception of acoustic information(Li et al., 2022; Millet et al., 2022; Tuckute et al., 2022; Millet and Dunbar, 2022).

Such studies can be formulated as a general framework of brain encoding and decoding (Naselaris et al., 2011; Huth et al., 2016; Yamins and DiCarlo, 2016). In brief, a predictive model is trained to build a mapping between the computational feature representation (the feature space, referred to artificial neurons, ANs) of the input stimuli and the brain activities (the brain activity space, referred to biological neurons, BNs) evoked by the same set of stimuli. The fitness of the predictive model, also known as the "brain score", is used to infer the correspondence between specific features and the underlying brain regions.

In most existing studies that link audio-transformers to brain responses, the layer-level contextual embeddings in the transformer encoding layers are used as the feature space (Li et al., 2022; Millet et al., 2022; Tuckute et al., 2022). The layer-level representations aggregate multiple types of attentional relationships among the input sequences captured by multi-head self-attention (MSA) modules (Vaswani et al., 2017). The aggregation operation results in comprehensive representations. However, these representations lack specificity. Thus, treating each encoding layer as a single AN is relatively coarse and consequently degenerates the capability of audio-transformers in brain encoding and decoding studies.

Multi-level visualizations of transformer attentions (Vig, 2019b; Clark et al., 2019; Aken et al., 2020) may provide some inspirations to address

*The corresponding author

Findings of the Association for Computational Linguistics: ACL 2023, pages 7943–7956
July 9-14, 2023

this problem. For example, BertViz visualizes the attention at the model-level, head-level and neuron-level(Vig, 2019a). More specifically, the neuron-level visualization factorizes the attention score matrix in each head into a set of element-wise product matrices corresponding to the hidden dimensions. The neuron-level visualization enables computational interpretation of transformers with fine granularity. However, whether each hidden dimension can be defined as a fine-grained AN for neural encoding and decoding study is not clear. Do those ANs carry meaningful linguistic information? Do those ANs anchor to their BN signatures in the human brain? Are the coupled AN-BN pairs interpretable from a neurolinguistic perspective?

We sought to answer these questions in this study. To this end, we propose a general framework for coupling the fine-grained ANs in Wav2Vec2.0 (Baevski et al., 2020) and the BNs in the human brain. We adopt the pre-trained Wav2Vec2.0 to embed the spoken story stimuli in the Narratives functional magnetic resonance imaging (fMRI) dataset (Nastase et al., 2021). The temporal response of an AN is then quantified according to the element-wise product of the queries and keys. Functional brain networks (FBNs) are identified from the fMRI data and each FBN is regarded as a single BN. Afterwards, the coupling relationship between ANs and BNs are built by maximizing the synchronizations between their temporal responses.

Our experimental results show that those fine-grained ANs carry meaningful linguistic information and well synchronize to their BN signatures, and the anchored AN-BN pairs are interpretable. More importantly, our results suggest an intermediate stage in both the computational representation in Wav2Vec2.0 and the cortical representation in the brain. The proposed fine-grained ANs may also serve as a general strategy to link transformer-based deep learning models to neural responses for probing the sensory processing in the brain.

2 Related works

Features from computational models have long been used to model the feature space for exploring the auditory neural encoding. Conventional hand-crafted features that capture low-level acoustic properties (e.g., sound intensity, timbre, rhythm, pitch, and spectrograms) have been found to be closely correlated to brain responses (Potes et al., 2012; Daube et al., 2019; Alluri et al., 2012; Cong et al., 2013; Santoro et al., 2014; Toiviainen et al., 2014; Hu et al., 2017; Pasley et al., 2012; Leaver and Rauschecker, 2010; Berezutskaya et al., 2017; Ylipaavalniemi et al., 2009; Norman-Haignere et al., 2015). Some studies replicate similar findings for the combinations of those low-level features optimized for specific tasks such as auditory attention (Bordier et al., 2013) and melodic pitch expectations (Pearce et al., 2010).

The deep neural networks (DNNs) developed for auditory and speech processing bring new opportunities to model the feature space. The model architecture and the training objective are two basic ingredients of DNNs. Existing studies have investigated the similarity between brain responses and DNNs in different architectures including convolutional neural network (CNN) (Saddler et al., 2021; Francl and McDermott, 2022; Kell et al., 2018; Güçlü et al., 2016; Huang et al., 2018; Thompson et al., 2021), convolutional auto-encoder (CAE) (Wang et al., 2022), generative adversarial network (GAN) (Beguš et al., 2022), CNN followed by recurrent neural network (RNN) (Li et al., 2022; Tuckute et al., 2022; Vaidya et al., 2022; Millet and King, 2021), spiking neural networks (Khatami and Escabí, 2020), and transformers (Li et al., 2022; Millet et al., 2022; Tuckute et al., 2022; Vaidya et al., 2022). The training objectives include unsupervised, self-supervised and supervised by various tasks such as musical genre prediction, acoustic scene classification, and speech recognition.

These studies have provided fruitful insights into neural auditory encoding, model interpretation, and brain-like model development. For example, correlating the hierarchical representations derived from CNN-based models for automatic music tagging have revealed the representational gradients in the superior temporal gyrus (STG). The anterior STG (aSTG) and posterior STG (pSTG) have been shown to be more sensitive to low-level and high-level features encoded in shallow and deep layers, respectively (Güçlü et al., 2016). By optimizing a CNN-based model for dual-task of word and music genre classification, Kell et al. showed that the best-performing network may resemble the hierarchical organization of the human auditory cortex. That is, brain responses in the primary and non-primary auditory cortices are most well predicted by middle and late CNN layers, respectively (Kell et al., 2018). By modeling the feature space via CNN-RNN-based DeepSpeech2 (Amodei et al., 2016) op-

timized for acoustic scene classification and speech-to-text with different types of inputs (i.e., English, Dutch and Bengali), Millet et al. replicated such a hierarchy and suggested that the brain utilizes sound-generic representations in the first processing stage of its hierarchy, and then builds speech-specific representations in higher-level processing stages (Millet and King, 2021).

More recently, the transformer based on multi-head self-attention (MSA) has emerged as a powerful DNN architecture to learn comprehensive contextual representations (Vaswani et al., 2017). In this context, audio-transformers such as Wav2Vec 2.0 have also been used to model the feature space (Millet et al., 2022; Li et al., 2022; Tuckute et al., 2022; Vaidya et al., 2022). For example, Millet et al. compared Wav2Vec 2.0 to neural activities in a large cohort, and found that the representational hierarchy of Wav2Vec 2.0 aligns with the cortical hierarchy of speech processing. More specifically, Wav2Vec2.0 learns sound-generic, speech-specific and language-specific representations that are analogous to those of the temporal and prefrontal cortices (Millet et al., 2022). Li et al. compared the representational similarity of HuBERT (Hsu et al., 2021), Wav2Vec 2.0 (Baevski et al., 2020) and DeepSpeech2 (Amodei et al., 2016) with different training objectives to the human auditory pathway. They showed that the representational hierarchy in the DNNs correlates well to the ascending auditory pathway, and unsupervised models achieve optimal neural correlations (Li et al., 2022). Tuckute et al. examined brain-DNN similarities within the auditory cortex for a large set of models based on various architectures and trained on different tasks. They found that most DNNs predicted brain responses in the auditory cortex better than the filterbank models as a baseline, and the models trained on multiple tasks produced the best overall predictions. More importantly, they showed that most of the DNNs exhibited a correspondence between model stages and brain regions, for example, the neural responses in lateral, anterior and posterior non-primary auditory cortices were better predicted by deeper layers (Tuckute et al., 2022).

Despite those fruitful findings, the feature space defined in existing studies that assess the representational similarity between Wav2Vec2.0 and brain responses relies on layer-level embeddings. That is, these studies implicitly treat each layer as a single artificial neuron. Considering the heterogeneity of the attentional heads, this operation may lose the specificity of each head, which is designed to capture different types of contextual attention. As argued in the field of natural language processing (NLP), a fine decomposition of a model's components into elementary units is among the keys for mapping computational models to their neurobiological counterparts (Hale et al., 2022; Poeppel, 2012). This demand also applies to audio-transformers. Meanwhile, our previous study has shown the validity of fine-grained ANs defined as the hidden dimensions of the pre-trained BERT model (Liu et al., 2023). However, whether those fine-grained ANs hold similar premises in audio-transformers is unknown. Thus, the key objective of this work is to validate those fine-grained ANs in Wav2Vec 2.0 for neural encoding studies.

3 Methods

3.1 Synchronization between ANs and BNs

Similar to that in our previous study (Liu et al., 2023), the bridge that connects ANs in Wav2Vec2.0 and BNs in brain responses is defined as the synchronization between their temporal responses to the same set of external stimuli. Let F : X$\rightarrow Y_a$ represent ANs, and $f_i(X)$ represent the temporal response of AN f_i to stimuli X. Similarly, let G : X $\rightarrow Y_b$ represent BNs, and $g_j(X)$ denote the temporal response of BN g_j to X. The best synchronized BN for an AN f_i is identified according to Eq.1.

$$\mathrm{Sync}(f_i, G) = \arg\max_{g_j \in G} \delta(f_i, g_j) \quad (1)$$

where $\delta(\cdot)$ measures the synchronization between the two responses. Similarly, the best synchronized AN for a BN g_i is identified according to Eq.2.

$$\mathrm{Sync}(g_i, F) = \arg\max_{f_i \in G} \delta(g_i, f_j) \quad (2)$$

In this study, we adopt the Pearson correlation coefficient (PCC) as $\delta(\cdot)$ to measure the temporal synchronization. The ANs and BNs, as well as their temporal responses to the inputs are detailed in the following sections.

3.2 ANs and Their Temporal Responses

The transformer aggregates multiple attentional relationships captured by the MSA module. The attention score in a head is formulated as $\mathbf{A} = softmax(\mathbf{Q^T K}/\sqrt{d})$ (Fig. 1a), where $\mathbf{Q} = \{q_1, q_2, \cdots, q_n\}$ is the query set, $\mathbf{K} =$

$\{k_1, k_2, \cdots, k_n\}$ is the key set, d is the hidden dimension in a head, and n is the number of tokens in the input sequence. After removing the softmax operation for simplification, a single entry in the attention matrix is formulated as $a_{ij} = q_i \cdot k_j = \sum_1^d q_i \cdot \times k_j$ (Fig. 1b), where $\cdot\times$ denotes element-wise product. This means that the attention matrix can be factorized into d element-wise product (EP) matrices (Fig. 1c). Each EP matrix characterizes how the query-key interactions in a hidden dimension contribute to the attention matrix. Thus, an intuitive idea is to define each hidden dimension as a single AN, which largely increases the granularity of ANs. For example, we can define $N_L \times N_H \times d$ (e.g., 9216 in Wav2Vec2.0) ANs in audio-transformers, where N_H and N_H are the numbers of layers and heads, respectively.

We then quantify the temporal response of an AN. It is notable that the ANs respond to the input tokens (25ms per token with 5ms overlap) but the fMRI observes the brain in the temporal resolution of repetition time (TR, 1.5s in the Narratives fMRI dataset). Thus, it is a prerequisite to temporally align the ANs' responses to fMRI volumes to measure the synchronization between them. To this end, the input audio stories are tokenized via the convolutional layers in Wav2vec2.0, and partitioned into subsets according to the TR. Let $\{t_1, t_2, \cdots, t_m\}$ denote the m tokens (m=75 in this study) in the j-th subset (corresponding to the j-th time point in fMRI), $\mathbf{Q}_j^{l,h} = \{q_1^{l,h}, q_2^{l,h}, \cdots, q_m^{l,h}\}$ and $\mathbf{K}_j^{l,h} = \{k_1^{l,h}, k_2^{l,h}, \cdots, k_m^{l,h}\}$ denote the queries and keys in the h-th head and l-th layer in Wav2Vec2.0, respectively. The i-th dimension of the corresponding element-wise product $\mathbf{EP}_j^{l,h,i} \in \mathbf{R}^{m\times m}$ (Fig. 1c) measures how a single AN selectively responds to all the m queries and m keys. Thus, we define the response of a single AN at time point j as the mean of the entries in $\mathbf{EP}_j^{l,h,i}$ (Fig. 1d). The temporal response of an AN to the entire input sequence is derived by iterating through all the token subsets (time points). Afterwards, it is convoluted with a canonical hemodynamic response function (HRF) implemented in SPM[1] to count for compensation for hemodynamic latency in fMRI.

3.3 BNs and Their Temporal Responses

The human brain is intrinsically organized as a complex networked system, and brain functions essentially rely on functional interactions among

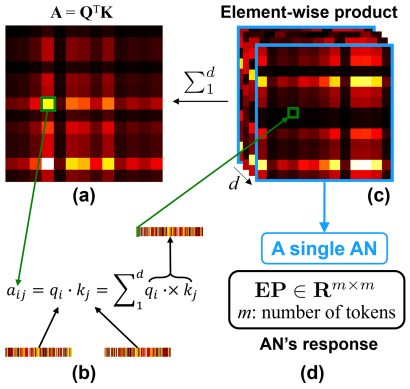

Figure 1: The definitions of AN and its response. The attention matrix (a) can be factorized as the summation of the element-wise products of queries and keys over hidden dimensions (b). A fine-grained AN is defined as each hidden dimension (c). The response of an AN can be derived from its element-wise product matrix (d).

functional brain networks (FBNs) (Park and Friston, 2013). Compared to the isolated voxels (an elementary structural unit in fMRI) that are used to quantify the brain activity space in most existing neural encoding studies (Tuckute et al., 2022; Millet et al., 2022; Vaidya et al., 2022), FBNs capture inter-regional functional interactions. Thus, we define each FBN as a single BN in neural recordings.

Various methods have been developed to identify FBNs in fMRI. Here, we adopt an open access model, the volumetric sparse deep belief networks (VS-DBN) [2] to identify FBNs (Dong et al., 2019). In brief, the VS-DBN learns a set of latent variables embedded in fMRI. Each latent variable consists of voxels exhibiting similar fluctuation patterns over time and represents the spatial map of an FBN.

The VS-DBN consists of an input layer and three layers of restricted Boltzmann machines (RBMs). It takes an fMRI volume as a feature and each time frame as a sample. The first RBM is with N visible units, where N is the number of voxels in a volume. The number of hidden units (m) in the third RBM determines the number of FBNs. The weights in RBMs are trained layer-wisely. The linear combination that performs successive multiplication of weights from the third to the first RBM is used to generate the global latent variables $\mathbf{W}$. Each column in $\mathbf{W}$ represents an FBN's spatial map. The responses of a single hidden unit in the third RBM to the entire input fMRI sequence are the corresponding time series of an FBN and are regarded as the temporal response of an FBN.

[1] https://www.fil.ion.ucl.ac.uk/spm/

[2] https://github.com/QinglinDong/vsDBN

4 Experiments

4.1 Dataset and Preprocessing

We use the open source "Narratives" fMRI dataset (Nastase et al., 2021) in the experiments. The "Narratives" fMRI data were acquired while human subjects listened to 27 diverse spoken stories. We select two sessions with moderate duration, the "Pie man" (Pieman) and "The Man Who Forgot Ray Bradbury" (Forgot). The Pieman is a story about a journalist writing reports of a man with supernatural abilities (duration 422s, word count 957). FMRI data were acquired for 82 subjects (282 volumes, spatial resolution $3 \times 3 \times 4\text{mm}^3$, TR=1.5s). The Forgot is about a man confronting a gradual loss of memory (duration 837s, word count 2135). FMRI data were acquired for 46 subjects (558 volumes, spatial resolution $2.5 \times 2.5 \times 2.5\text{mm}^3$, TR=1.5s). The "Narratives" fMRI data were released with various preprocessed versions and we use the AFNI-smooth version. The spoken story was released with time-stamped word-level transcripts and the onset and duration of each phoneme in a word. We use this information to temporally align phonemes and fMRI volumes. In addition, we tag phonemes in the audio-story with typical categories (vowel, mixed, fricative, affricate, nasal and stop) defined previously that cover the phonetic inventory of 38 unique phonemes (Hamooni and Mueen, 2014).

4.2 Implementation Details

We use the pre-trained Wav2Vec2.0-base maintained by HuggingFace [3] in the experiments. We partition the stories into short segments by balancing the token capacity of Wav2Vec2.0 and sentence integrity. It is notable that the story Forgot is much longer than Pieman. Thus, we crop the Forgot from the beginning to have the same number of TRs as that in Pieman to facilitate a cross-validation. As such, both spoken stories are partitioned into 25 segments (duration: 16.62±6.74s in Pieman and 15.30±3.50s in Forgot).

We train the VS-DBN model to extract FBNs for each fMRI session independently. The fMRI volumes of multiple subjects (randomly selected 75 subjects in Pieman, and all the 46 subjects in Forgot) are aggregated as samples (20775/25668 in Pieman/Forgot). The parameters are set as follows: 512/256/128 hidden units in the 1st/2nd/3rd RBM layer, Gaussian initialization with zero-mean and a standard deviation of 0.01, learning rate 0.001/0.0005/0.0005, batch-size 20, L1 weight-decay rate 0.001/0.00005/0.00005, 100 training epochs, batch normalization. In each session, the resulted FBNs in all the subjects share the same set of spatial maps but have subject-specific temporal responses. The subject-specific temporal responses are averaged over subjects to characterize the temporal responses of FBNs in population.

[3] https://huggingface.co/docs/transformers/main/en/model_doc/wav2vec2

5 Results

5.1 Synchronization between ANs and BNs

We first assess the intra-session synchronization between ANs and BNs. The distributions of the AN's maximum PCC to BNs for Pieman (Pieman-Pieman, 0.3305±0.0042) and Forgot (Forgot-Forgot, 0.3142±0.0049) are shown in Fig. 2(a). Permutation tests with 5000 randomizations show that the PCCs are significant ($p < 0.01$, FDR corrected) for 9203/9192(99.86%/99.74%) ANs in Pieman/Forgot. In both sessions, the average PCC in each layer (Fig. 2b) is relatively stable in the first ten layers but increases sharply in the last two layers, indicating that the ANs in the last two layers better synchronize to BNs. We then evaluate the inter-session synchronization between ANs and BNs, which may server as a stronger baseline control. We identify the best correlated AN in one session and the BN in the other. The inter-session PCCs are significantly ($p < 10^{-10}$) lower compared to the intra-session one in both sessions (Pieman-Forgot, 0.1912±0.0016; Forgot-Pieman, 0.2097±0.0032; Fig. 2a). The AN that is anchored by an BN is identified according to Eq. 2. The PCCs (0.4262±0.0027 in Pieman and 0.4199±0.0029 in Forgot) are statistically significant ($p < 10^{-10}$) for all the 128 BNs in both sessions. These observations show that the temporal responses of ANs and BNs are well synchronized.

5.2 The Global BN anchored by ANs

We identify the global BN as the one that is the most frequently anchored by ANs after applying a PCC threshold of 0.25 (Fig. 3a). The spatial distributions of the global BNs in Pieman (BN#47) and Forgot (BN#42) are similar. They mainly encompass the Heschl's gyrus (HG) and nearby superior temporal gyrus (STG), posterior superior temporal sulcus (pSTS), posterior inferior temporal gyrus (pITG), temporal pole (TP), temporo-parietal junction (TPJ), Broca's and Wernicke's areas in the in-

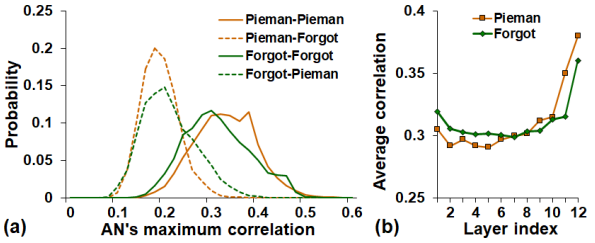

Figure 2: (a) The distributions of intra-session (Pieman-Pieman and Forgot-Forgot) and inter-session (Pieman-Forgot and Forgot-Pieman) temporal synchronization between ANs and BNs. (b) The average temporal synchronization in each layer.

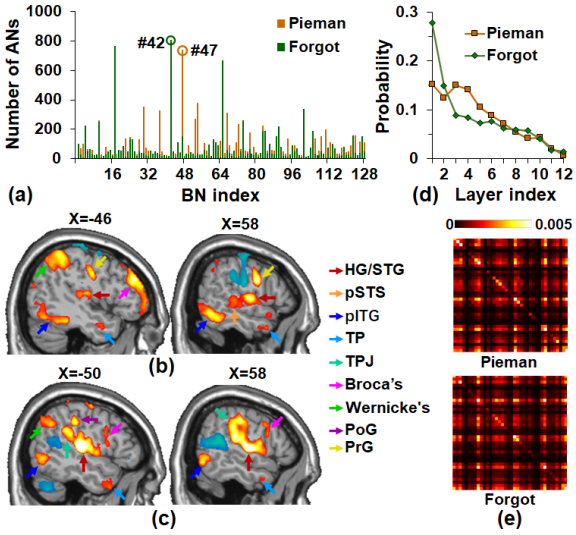

Figure 3: (a) The number of ANs that are anchored by a BN. The BN that is the most frequently anchored is considered as the global BN (#47 in Pieman and #42 in Forgot). (b) The BN#47 in Pieman. (c) The BN#42 in Forgot. (d) The distribution of the global BN in different layers. (e) The average phoneme distribution matrix. Left/right: Pieman/Forgot. HG: Heschl's gyrus; STG: superior temporal gyrus; pSTS: posterior superior temporal sulcus; pITG: posterior inferior temporal gyrus; TP: temporal pole; TPJ: temporo-parietal junction; PoG: post-central gyrus; PrG: pre-central gyrus.

ferior frontal gyrus (IFG), pre-central gyrus (PrG), and post-central gyrus (PoG) (Fig. 3b-c). These brain regions well match the cortical anatomy of the dual-stream language model (Hickok and Poeppel, 2007). The earliest stage of neural speech processing involves spectrotemporal analysis in HG and STG, followed by phonological-level representation in STS. Subsequently, the system diverges into two streams, a dorsal steam (TPJ, IFG and sensorymotor cortex of PrG and PoG) and a ventral stream (pITG and TP) map sensory or phonological representations onto articulatory motor representations and lexical conceptual representations, respectively (Hickok and Poeppel, 2007).

Intriguingly, the ANs that synchronize with the global BN are widely distributed across layers 1-10, but predominantly located on lower layers (i.e., 1-4, Fig. 3d). We then assess the phonemic patterns of query-key pairs that those ANs selectively respond. In each of the 25 audio segments we select 1500 query-key pairs that have top values in the EP matrix corresponding to each of the ANs, and construct a 38×38 phoneme distribution matrix (PDM), in which rows are queries and columns are keys. Each entry in the PDM is the proportion of the query-key pairs falling into the entry. The average PDMs over all the ANs are quite homogeneous in both sessions (Fig. 3e), showing that the global BN responds to general attentional relationships among phonemes. Meanwhile, some ANs are selective to phonemic relationships (e.g., vowel-vowel, see details in section 5.4, Fig. 8). Taken together, these observations reinforce the prevalent functional interactions (Cloutman, 2013; Bhaya-Grossman and Chang, 2022) among the brain regions covered by the global BN and suggest that the lower layers in Wav2Vec2.0 are responsive to learn phonemic relationships.

5.3 The Local BNs in Each Layer

We then identify the frequently anchored BN in each layer (local BNs). We define the anchoring frequency of a BN in a layer as the ratio between the number of ANs it anchors in that layer and the total number of anchored AN-BN pairs in that layer. The BNs' anchoring frequency shows distinctive patterns across layers (Fig. 4). In lower layers 1-5 and upper layers 11-12, the local BNs are very sparse and limited to one or two predominant BNs. For example, the anchoring frequency of the BN#47 is much higher compared to those of the rest in layers 1-5. Meanwhile, the local BNs in layers 6-10 are widely spread. That is, the ANs in those layers tend to anchor to different BNs. We highlight the local BNs in each layer by circles and show their spatial maps in Fig. 5 for Pieman. The BN#47 is referred Fig. 3(b). The anchoring frequency and the local BNs in Forgot are shown in Fig. A.1-A.2, respectively.

The global BN (BN#47) is identified as the local BN in layers 1-5, however, its anchoring frequency decreases as the layer goes deeper (Fig. 6a). The BN#54 encompasses the working memory network (WM, retains short-term temporal memory) and

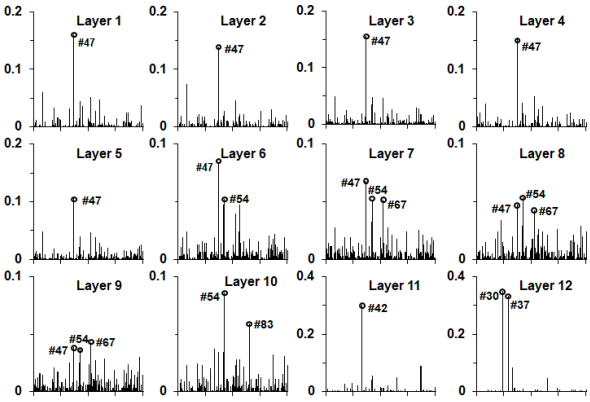

Figure 4: The BN's anchoring frequency in each layer. In each subplot, the x-axis is BN index and the y-axis is anchoring frequency. Circles highlight local BNs.

the language network (Broca's and Wernicke's areas), reflecting the functional interactions between them. It is identified as the local BNs in the intermediate layers 6-10 and its anchoring frequency increases and reaches the peak in layer 10 (Fig. 6b). The BN#67 involves the activations in PrG and the deactivations in PrG and HG/STG, reflecting the functional competition between them. It is identified as the local BNs in layers 7-9. Its anchoring frequency indicates that it is widely anchored by ANs in layers 1-9 but fades out sharply in the last three layers (Fig. 6c). The BN#83, which is one of the local BNs in layer 10, exhibits activations in precuneus cortex (PcC, which is considered as part of the brain's semantic system (Binder et al., 2009)) and frontal pole (FP), as well as deactivations in PrG and PoG. The BN#42 reflects functional interactions among the primary auditory cortex (HG/STG), the language network and visual cortex (intracalcarine cortex, IcC, and PcC) and it predominates the local BNs in layer 11. There are two local BNs in layer 12, BN#30 and BN#37. The BN#30 shows complex co-activations in the ventral and dorsal streams in speech processing, and semantic related regions including the angular gyrus (AG), posterior supramarginal gyrus (SmGp), and lateral occipital cortex (LOCs). The BN#37 mainly covers the ventral and dorsal streams in speech processing.

Using cumulative attention diagonality (CAD) applied to head-level attention score, Shim et al. have shown distinctive global and local attention patterns in lower (1-8) and upper (9-16) layers in Wav2Vec2.0, respectively. The former integrates long-range attention to form phonetic localization, while the latter focuses on short-range diagonal attention for language identification (Shim et al., 2021). We apply the same metric to the EP matrix of ANs rather than the head-level attention score. Fig. 7 shows the average CAD over top 1% and top 2% ANs in each layer for a randomly selected segment in the two sessions. We identify a transient stage (layers 6-10) between global (layers 1-4) and local (layers 11-12) ones. Combined with the fade-out of BN#47 (layers 1-5) and the fade-in of BN#54 (layers 6-10) along the layer depth (Fig. 6), we suggest that there is an intermediate level between the global and local ones in Wav2Vec2.0. That is, the layers 6-10 may gradually integrate global phonetic localization encoded in the early stages of cortical speech hierarchy (BN#47) through the functional interactions between WM and the language network (BN#54) to form local language localization. In addition, the good predictive performance in WM has rarely been reported in exiting neural encoding studies of Wav2Vec2.0 (Li et al., 2022; Millet et al., 2022; Tuckute et al., 2022; Vaidya et al., 2022), which may be partly due to the relatively coarse layer-level ANs used in these studies. Thus, the fine-grained ANs defined in this study enable us to preliminarily reveal this intermediate-level representation in Wav2Vec2.0 and map it to its neurobiological counterparts.

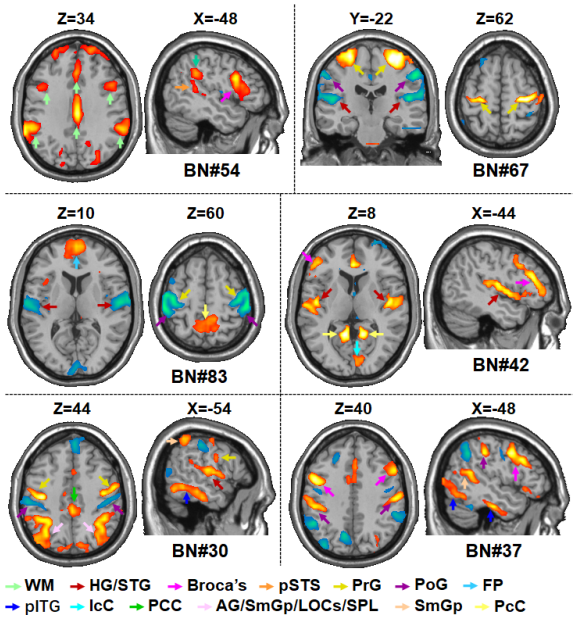

Figure 5: The local BNs in Pieman. WM: working memory; FP: frontal pole; IcC: intra-calcarine cortex; PCC: posterior cingulate cortex; AG: angular gyrus; SmGp: posterior supramarginal gyrus; LOCs:lateral occipital cortex; SPL: superior parietal lobule; PcC: precuneus cortex.

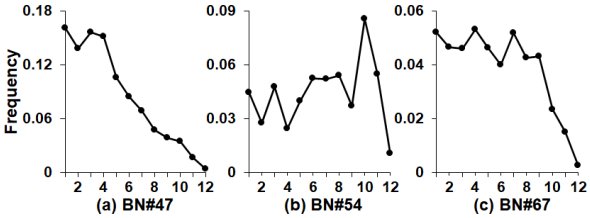

Figure 6: The anchoring frequency of three local BNs in different layers. (a) BN#47. (b) BN#54. (c) BN#67.

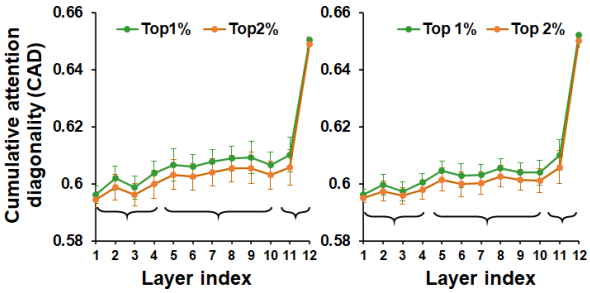

Figure 7: The average cumulative attention diagonality (CAD) of each layer. Left/right: Pieman/Forgot.

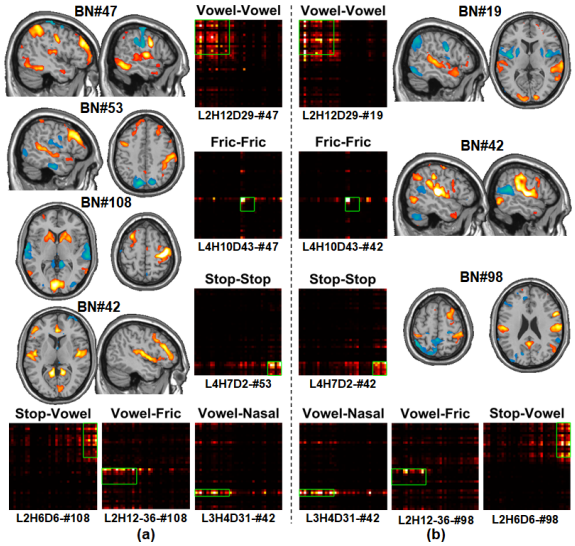

Figure 8: Phoneme-selective AN-BN pairs. (a) Pieman. (b) Forgot. The indices of AN and BN in a pair are below the phoneme distribution matrix. L2H12D29-#47 denotes that the AN located on layer 1, head 12, and dimension 29 synchronizes with BN#47.

5.4 Phoneme-selective AN-BN Paris

We identify some ANs that are selective to different categories of phonemes, as shown in Fig. 8 for some examples. In each example, we show the phoneme distribution matrix (PDM) of the AN, and the BN anchored by the AN. It is notable that the ANs in the two sessions are identical and the corresponding BNs are similar, showing good reproducibility across sessions. Brain regions including HG/STG, STS and sensorimotor areas are frequently observed in those BNs, which is partly in line with previous studies (Kim et al., 2021).

Despite some interesting findings in computational interpretation of audio-transformers (Shim et al., 2021; Yang et al., 2020), the neural basis of phoneme-selectivity in the brain is still under debate (Mesgarani et al., 2014; Gwilliams et al., 2022; Bhaya-Grossman and Chang, 2022; Sohoglu, 2019). What we intend to convey here is that the fine-grained ANs defined in this study, applied in a neural encoding framework, may provide an alternative strategy to probe this problem.

6 Discussion and Conclusion

We proposed to define fine-grained artificial neurons (ANs) in the audio-transformer Wav2Vec2.0 and map them to their neurobiological counterparts. Our experimental results showed that the fine-grained ANs carried meaningful linguistic information and well synchronized to their BN signatures. Moreover, the anchored AN-BN pairs are partly interpretable in a neurolinguistic view.

Although a comprehensive mapping of the cortical speech hierarchy is out of the scope of this study, we observed some interesting results, facilitated by the fine-grained ANs. First, the alignment between the computational hierarchy in Wav2Vec2.0 and the cortical speech hierarchy is largely in line with existing studies (Li et al., 2022; Millet et al., 2022; Tuckute et al., 2022; Vaidya et al., 2022). Second, and more importantly, we preliminarily discovered an intermediate stage in both the computational representation in Wav2Vec2.0 and the cortical representation in the brain. It gradually integrates global phonetic localization encoded in the early stages of neural speech hierarchy through the functional interactions between the working memory and language networks to form local language localization. In comparison, a good predictive performance from computational representation in audio-transformers to brain activities has rarely been reported previously. Third, we observed phoneme-selective neural-level ANs in Wav2Vec2.0, and the associated BNs are partly in line with existing studies (Kim et al., 2021). Thus, the fine-grained ANs defined here may potentially provide an alternative approach to explore whether there are phoneme-selective neural activities in the brain.

The fine-grained ANs defined in this study may also serve as the brain-based test-bed to evaluate

and interpret audio-transformers, and provide neurolinguistic support for better understanding of the role of self-attention for efficient speech computation. For example, after interpreting distinctive attentional profiles in different attention heads, Shim et al. applied a layer-wise attention map reuse strategy to improve model performance (Shim et al., 2021). A similar but with more fine-grained strategy could further improve model performance.

In conclusion, we defined and validated neuron-level ANs in Wav2Vec2.0. This definition may serve as a general strategy to link transformer-based deep learning models to neural responses for probing the sensory processing in the brain.

7 Limitation

The current study is has some limitations. First, we used a single audio-transformer model, the pre-trained Wav2Vec2.0-base, as a test bed to validate the fine-grained ANs and couple them to their BN signatures. On the one hand, various audio-transformers have been proposed in the literature. On the other hand, the parameters of a pre-trained model are fine-tuned by downstream tasks and previous studies have shown that fine-tuning may lead DNNs to increase their brain similarity (Millet and King, 2021; Tuckute et al., 2022). Thus, it would be interesting to explore whether there are consistent AN-BN coupling patterns across different models, either pre-trained or fine-tuned. In addition, it is necessary to investigate these patterns across different languages (e.g., English VS Mandarin).

Second, existing studies have shown that audio-transformers are able to learn sound-generic, speech-specific and language-specific representations and those hierarchical representations are akin to the cortex (Li et al., 2022; Millet et al., 2022; Vaidya et al., 2022). Thus, it would be interesting to explore whether the fine-grained ANs carry such multi-level representations, and link them to brain responses.

Third, the reproducibilty between the two sessions was high regarding to most of the results (e.g., the global BNs and the phoneme-selective AN-BN pairs), but it was relatively low in some results (e.g., the local ANs in some layers). We speculate that this is the consequence of relatively smaller fMRI training samples but much larger amount of VS-DBN model parameters in the session of Forgot, in which the number of subjects is smaller but the fMRI spatial resolution are higher. Higher spatial resolution results in much larger number of valid voxels (120,506) compared to that in Pieman (50,065) and consequently more visible units in the VS-DBN model.

Last but not least, the analyses presented in this study are intrinsically limited by the coarseness of spatial (voxels in millimeters) and temporal resolution (volumes in seconds) of fMRI data. Mapping from sound to an interpretable representation involves integrating neural activities on different spatial-scales down to sub-millimeters and on different timescales down to milliseconds. Thus, it would be of great interest in the future to apply the fine-grained ANs to auditory magnetoencephalogram (MEG) dataset to disentangle the symbiosis of model computation and brain responses in both space and time (Bhaya-Grossman and Chang, 2022; Gwilliams et al., 2022).

8 Acknowledgements

This work was partly supported by National Key R&D Program of China (2020AAA0105701), National Natural Science Foundation of China (62076205, 61936007 and 61836006).

References

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2020. Visbert: Hidden-state visualizations for transformers. In *Companion Proceedings of the Web Conference 2020*, pages 207–211.

Vinoo Alluri, Petri Toiviainen, Iiro P Jääskeläinen, Enrico Glerean, Mikko Sams, and Elvira Brattico. 2012. Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *Neuroimage*, 59(4):3677–3689.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Gašper Beguš, Alan Zhou, and T Christina Zhao. 2022. Encoding of speech in convolutional layers and the brain stem based on language experience. *bioRxiv*.

Julia Berezutskaya, Zachary V Freudenburg, Umut Güçlü, Marcel AJ van Gerven, and Nick F Ramsey. 2017. Neural tuning to low-level features of

speech throughout the perisylvian cortex. *Journal of Neuroscience*, 37(33):7906–7920.

Ilina Bhaya-Grossman and Edward F. Chang. 2022. Speech computations of the human superior temporal gyrus. *Annual Review of Psychology*, 73(1):79–102.

Jeffrey R. Binder, Rutvik H. Desai, William W. Graves, and Lisa L. Conant. 2009. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, 19(12):2767–2796.

Cecile Bordier, Francesco Puja, and Emiliano Macaluso. 2013. Sensory processing during viewing of cinematographic material: Computational modeling and functional neuroimaging. *Neuroimage*, 67:213–226.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Lauren L. Cloutman. 2013. Interaction between dorsal and ventral processing streams: Where, when and how? *Brain and Language*, 127(2):251–263.

Fengyu Cong, Vinoo Alluri, Asoke K Nandi, Petri Toiviainen, Rui Fa, Basel Abu-Jamous, Liyun Gong, Bart GW Craenen, Hanna Poikonen, Minna Huotilainen, et al. 2013. Linking brain responses to naturalistic music through analysis of ongoing eeg and stimulus features. *IEEE Transactions on Multimedia*, 15(5):1060–1069.

Christoph Daube, Robin A.A. Ince, and Joachim Gross. 2019. Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Current Biology*, 29(12):1924–1937.e9.

Qinglin Dong, Fangfei Ge, Qiang Ning, Yu Zhao, Jinglei Lv, Heng Huang, Jing Yuan, Xi Jiang, Dinggang Shen, and Tianming Liu. 2019. Modeling hierarchical brain networks via volumetric sparse deep belief network. *IEEE transactions on biomedical engineering*, 67(6):1739–1748.

Andrew Francl and Josh H McDermott. 2022. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, 6(1):111–133.

Umut Güçlü, Jordy Thielen, Michael Hanke, and Marcel Van Gerven. 2016. Brains on beats. *Advances in Neural Information Processing Systems*, 29.

Laura Gwilliams, Jean-Remi King, Alec Marantz, and David Poeppel. 2022. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature communications*, 13(1):1–14.

John T Hale, Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, and Jonathan R Brennan. 2022. Neurocomputational models of language processing. *Annual Review of Linguistics*, 8:427–446.

Hossein Hamooni and Abdullah Mueen. 2014. Dual-domain hierarchical classification of phonetic time series. In *2014 IEEE international conference on data mining*, pages 160–169. IEEE.

Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. IEEE.

Xintao Hu, Lei Guo, Junwei Han, and Tianming Liu. 2017. Decoding power-spectral profiles from fmri brain activities during naturalistic auditory experience. *Brain imaging and behavior*, 11(1):253–263.

Nicholas Huang, Malcolm Slaney, and Mounya Elhilali. 2018. Connecting deep neural networks to physical, perceptual, and electrophysiological auditory signals. *Frontiers in neuroscience*, 12:532.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.

Fatemeh Khatami and Monty A Escabí. 2020. Spiking network optimized for word recognition in noise predicts auditory system hierarchy. *PLOS Computational Biology*, 16(6):e1007558.

Seung-Goo Kim, Federico De Martino, and Tobias Overath. 2021. Linguistic modulation of the neural encoding of phonemes. *bioRxiv*.

Amber M Leaver and Josef P Rauschecker. 2010. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *Journal of Neuroscience*, 30(22):7604–7612.

Yuanning Li, Gopala K Anumanchipalli, Abdelrahman Mohamed, Junfeng Lu, Jinsong Wu, and Edward F Chang. 2022. Dissecting neural computations of the human auditory pathway using deep neural networks for speech. *bioRxiv*.

Xu Liu, Mengyue Zhou, Gaosheng Shi, Yu Du, Lin Zhao, Zihao Wu, David Liu, Tianming Liu, and Xintao Hu. 2023. Coupling artificial neurons in bert and biological neurons in the human brain. In *AAAI 2023*.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. 2014. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. 2022. Toward a realistic model of speech processing in the brain with self-supervised learning. *arXiv preprint arXiv:2206.01685*.

Juliette Millet and Ewan Dunbar. 2022. Do self-supervised speech models develop human-like perception biases? *arXiv preprint arXiv:2205.15819*.

Juliette Millet and Jean-Remi King. 2021. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. *arXiv preprint arXiv:2103.01032*.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. 2011. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410.

Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, et al. 2021. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific data*, 8(1):1–22.

Sam Norman-Haignere, Nancy G. Kanwisher, and McDermott Josh H. 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 88(6):1281–1296.

Hae-Jeong Park and Karl Friston. 2013. Structural and functional brain networks: from connections to cognition. *Science*, 342(6158):1238411.

Brian N Pasley, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, and Edward F Chang. 2012. Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251.

Marcus T Pearce, María Herrojo Ruiz, Selina Kapasi, Geraint A Wiggins, and Joydeep Bhattacharya. 2010. Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, 50(1):302–313.

David Poeppel. 2012. The maps problem and the mapping problem: two challenges for a cognitive neuroscience of speech and language. *Cognitive neuropsychology*, 29(1-2):34–55.

Cristhian Potes, Aysegul Gunduz, Peter Brunner, and Gerwin Schalk. 2012. Dynamics of electrocorticographic (ecog) activity in human temporal and frontal cortical areas during music listening. *NeuroImage*, 61(4):841–848.

Mark R Saddler, Ray Gonzalez, and Josh H McDermott. 2021. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature communications*, 12(1):1–25.

Roberta Santoro, Michelle Moerel, Federico De Martino, Rainer Goebel, Kamil Ugurbil, Essa Yacoub, and Elia Formisano. 2014. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS computational biology*, 10(1):e1003412.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Kyuhong Shim, Jungwook Choi, and Wonyong Sung. 2021. Understanding the role of self attention for efficient speech recognition. In *International Conference on Learning Representations*.

Ediz Sohoglu. 2019. Auditory neuroscience: sounding out the brain basis of speech perception. *Current Biology*, 29(12):R582–R584.

Jessica AF Thompson, Yoshua Bengio, Elia Formisano, and Marc Schönwiesner. 2021. Training neural networks to recognize speech increased their correspondence to the human auditory pathway but did not yield a shared hierarchy of acoustic features. *bioRxiv*.

Petri Toiviainen, Vinoo Alluri, Elvira Brattico, Mikkel Wallentin, and Peter Vuust. 2014. Capturing the musical brain with lasso: Dynamic decoding of musical features from fmri data. *Neuroimage*, 88:170–180.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. 2022. Many but not all deep neural network audio models capture brain responses and exhibit hierarchical region correspondence. *bioRxiv*.

Aditya R Vaidya, Shailee Jain, and Alexander G Huth. 2022. Self-supervised models of audio effectively explain human cortical responses to speech. *arXiv preprint arXiv:2205.14252*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig. 2019a. Bertviz: A tool for visualizing multihead self-attention in the bert model. In *ICLR Workshop: Debugging Machine Learning Models*.

Jesse Vig. 2019b. Visualizing attention in transformer-based language representation models. *arXiv preprint arXiv:1904.02679*.

Liting Wang, Huan Liu, Xin Zhang, Shijie Zhao, Lei Guo, Junwei Han, and Xintao Hu. 2022. Exploring hierarchical auditory representation via a neural encoding model. *Frontiers in neuroscience*, 16.

Daniel LK Yamins and James J DiCarlo. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.

Shu-wen Yang, Andy T Liu, and Hung-yi Lee. 2020. Understanding self-attention of self-supervised audio transformers. *arXiv preprint arXiv:2006.03265*.

Jarkko Ylipaavalniemi, Eerika Savia, Sanna Malinen, Riitta Hari, Ricardo Vigário, and Samuel Kaski. 2009. Dependencies between stimuli and spatially independent fmri sources: Towards brain correlates of natural stimuli. *NeuroImage*, 48(1):176–185.

A Appendix

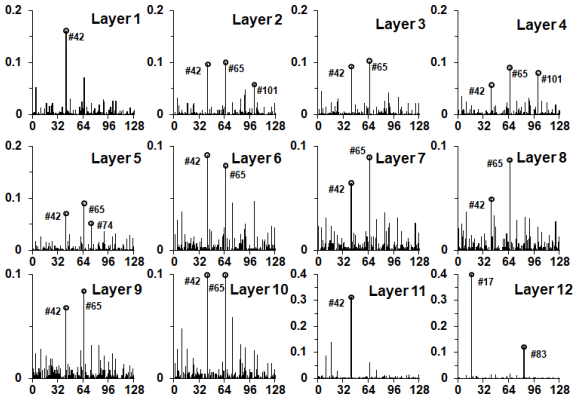

Figure A.1: The anchoring frequency of BNs in each layer in Forgot. In each subplot, the *x*-axis is BN index and the *y*-axis is anchoring frequency.Circles highlight the indices of local BNs.

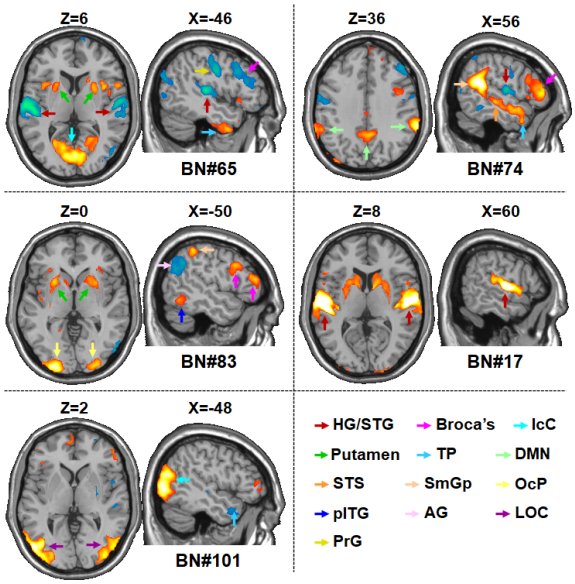

Figure A.2: The local BNs in Forgot. HG: Heschl's gyrus; STG: superior temporal gyrus; STS: superior temporal sulcus; pITG: posterior inferior temporal gyrus; TP: temporal pole; PrG: pre-central gyrus; IcC: intracalcarine cortex; DMN: default mode network; SmGp: posterior supramarginal gyrus; OcP: occipital pole; LOC: lateral occipital cortex; AG: angular gyrus.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract and section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3, 4, 5

☑ B1. Did you cite the creators of artifacts you used?
3, 4, 5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3, 4, 5

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3, 4, 5

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4, 5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4, 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.