INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Similar hierarchical representation of speech and other complex sounds in the brain and deep residual networks: An MEG study

*Tzu-Han Zoe Cheng*[1]*,Kuan-Lin Chen*[1]*,Juliane Schubert*[2]*,Ya-Ping Chen*[2]*,Tim Brown*[1]*,John Iversen*[1]

[1]University of California, San Diego, USA
[2]University of Salzburg, Austria

tzcheng@ucsd.edu, kuc029@ucsd.edu, juliane.schubert@plus.ac.at, ya-ping.chen@plus.ac.at, ttbrown@health.ucsd.edu, jiversen@ucsd.edu

## Abstract

Listeners recognize a vast number of complex sounds, but vocal sounds, speech and song, are essential for communication. Recently, deep neural networks (DNNs) have achieved human-level accuracy in sound classification, but do they illuminate similar properties with biological brains? In this study, we compared DNNs to primary and secondary auditory cortex to understand the hierarchy of sound representations in the brain. Ten subjects listened to speech and other naturalistic sounds while their magnetoencephalography (MEG) signals were recorded. Widely-used DNNs were trained to classify the same sounds. Brain activity localized to secondary auditory areas decoded speech significantly more accurately than other non-human sounds. Secondary auditory selectivity best matched later, and more complex layers of DNNs. Our results are compatible with special coding for speech in the brain and suggest comparable hierarchical principles of DNNs and neural processing of sounds.

**Index Terms**: speech perception, decoding model, MEG, deep neural networks, sound classifications

## 1. Introduction

Humans hear and recognize a vast number of sounds in everyday life. Invasive ECoG, primate work, and fMRI studies have confirmed a general arrangement of primary core auditory areas coding low-level acoustic features (e.g. frequency, spectrotemporal modulation) and secondary belt auditory areas responding to sounds that are more complex and significant to humans [1, 2, 3]. However, how these complex sounds such as speech and song, which are particularly important in our life, are represented and categorized in high-level auditory areas is still a mystery. In our study, we propose a novel approach that used deep neural networks (DNNs) as working models of the human brain, seeking to identify similar principles in order to understand sound representations in the auditory cortex.

DNNs [4] are essential components of modern artificial intelligence (AI) systems and have achieved human-level or even super-human performance in specific tasks such as game playing [5], object detection [6], and sound classification [7]. Because DNNs share design principles inspired by the human brain, such as hierarchical layering and non-linear computation, they may provide insight into the human brain. Brain architectures have also inspired the adoption of skip connections or highways in DNNs, giving the ResNets[8, 9] or highway networks [10], which have been shown to empower a DNN to be scaled to have more than a thousand layers and still improve its representation power [11].

Despite many aspects of DNNs that are highly artificial, parallels between these hierarchical principles of DNNs and the human brain have been supported by empirical evidence that higher-level regions (V4 & IT) in the primate visual system are predicted better by the later layers of neural networks trained on the image recognition tasks [12, 13]. Yet, only a few studies have probed these parallels in the auditory system. The most relevant prior work is the seminal work by Kell *et al.* [14], who compared the functional magnetic resonance imaging (fMRI) of the brain and DNNs. They optimized hierarchical convolutional neural networks to perform word and music genre recognition tasks and then showed that differences in stimulus specificity in primary and non-primary auditory cortex were matched with different convolutional layers of their DNN, showing a similar hierarchical processing of sounds between DNNs and human brains and further claiming unique representations for speech and music in the brain are also seen in DNNs.

Regarding [14], the result is weakened by the fact that the separate representations for speech and music are a priori baked into the DNN, which had two independent branches for speech and music. Thus it is an open question if a general-purpose DNN would develop separate speech and music representations. Furthermore, auditory processing is fundamentally dynamic in time, and this is not captured by the fMRI. Magnetoencephalography (MEG), which has excellent temporal resolution and is more widely available for human study than invasive methods, could provide new and valuable insights into the nature of au-

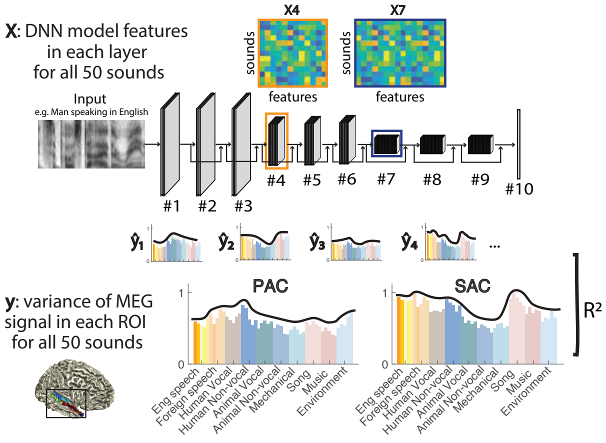
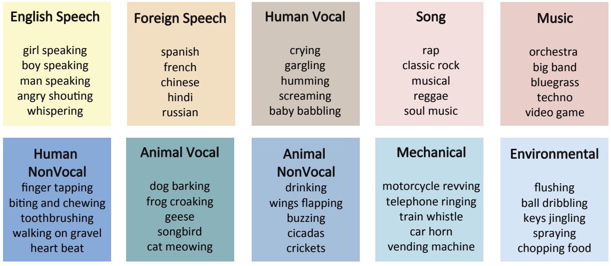


Figure 1: *Schematic analysis flow of fitting a linear model* $\mathbf{w}^*$ *between MEG signal and DNN model units. The DNN model units (*$\mathbf{X}$*) were collected from each layer for each sound. The predicting value was the variance of MEG for each sound (*$\mathbf{y}$*).* $R^2$ *was computed between the* $\mathbf{X}\mathbf{w}^*$ *and* $\mathbf{y}$*, used as a criterion to measure the model fit. See Section 2.5 for more details.*

10.21437/Interspeech.2023-1347



English Speech
girl speaking
boy speaking
man speaking
angry shouting
whispering

Foreign Speech
spanish
french
chinese
hindi
russian

Human Vocal
crying
gargling
humming
screaming
baby babbling

Song
rap
classic rock
musical
reggae
soul music

Music
orchestra
big band
bluegrass
techno
video game

Human NonVocal
finger tapping
biting and chewing
toothbrushing
walking on gravel
heart beat

Animal Vocal
dog barking
frog croaking
geese
songbird
cat meowing

Animal NonVocal
drinking
wings flapping
buzzing
cicadas
crickets

Mechanical
motorcycle revving
telephone ringing
train whistle
car horn
vending machine

Environmental
flushing
ball dribbling
keys jingling
spraying
chopping food

Figure 2: *Sound stimuli extracted from the Natural Sound Stimulus set. There are ten sound categories and each with 5 different sound examples.*

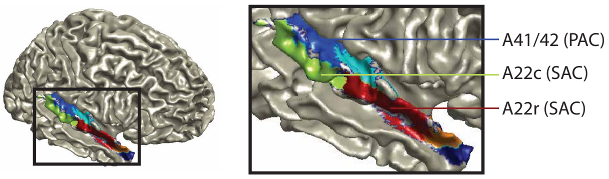


Figure 3: *Source localization of primary and secondary auditory cortex. The colored parts are the auditory cortex, including the primary auditory cortex (PAC) in blue and the caudal and rostral secondary auditory cortex (SAC) in green and red. The same regions of left hemisphere were identified, yielding six regions of interest used in the analysis.*

ditory coding. Moreover, convolutional neural networks used in past work [12, 14, 13] are outperformed by DNNs with skip connections such as the ResNet [8, 9], wide ResNet [15] and U-Net [16], which have achieved remarkable performance in many applications [17, 18, 19] and led to pivotal design principles for deep learning [11].

In this paper, we asked to what degree dynamic responses of the human auditory cortex and DNNs may have similar hierarchical processing principles. We collected MEG signals while subjects listened to complex, naturalistic sounds in order to investigate how the human brain and advanced general-purpose DNNs with skip connections process complex sounds. We found brain selectivity for speech and vocal sounds, including song, compared to other non-human sounds in the secondary auditory cortex and, critically — without presuming this difference in the network architecture, found that general-purpose DNNs recaptured this distinction, and primarily in the later layers of DNNs. To the best of our knowledge, this is the first work using MEG to demonstrate similar hierarchical processing of sounds, as well as speech selectivity, in the human auditory cortex and DNNs trained to perform sound classification.

# 2. Materials and Methods

## 2.1. Stimuli

Fifty naturalistic sounds were selected from the Natural Sound Stimulus set used in [2], each sound is 2-seconds long and has been categorized into one of ten categories, resulting in 5 sounds in each category (Figure 2). The ten sound categories are English speech, foreign speech, music (instrumental), song (music with vocals), human non-speech vocal sound, human non-vocal sound, animal vocalization, animal non-vocal sound, mechanical sound, or environmental sound. These sounds were played in random order with each sound repeated overall 10 times resulting in 500 trials, recorded by MEG. All sounds were presented binaurally at comfortable volume while participants were looking at a fixation cross in the center of the screen, that changed color (switching from black to red for 100ms) at random intervals in order to probe participants' attention (they had to respond with a button press as quickly as possible).

## 2.2. MEG data acquisition

Ten individuals (2 females, 8 males, 0 nonbinary, mean age = 26.1) participated in this study. They were native German speakers who were fluent in English, reported normal hearing and no history of psychological or neurological disease, and were eligible for MEG recordings (i.e. without ferromagnetic metals in or close to their bodies). This study was in accordance with the Declaration of Helsinki and approved by the Ethics Committees of the Department of Psychology, University of Salzburg. Participants signed informed consent forms to participate. The participants were compensated monetarily or with course credit.

MEG was recorded at 1000 Hz using a 306-channel Triux MEG system (Elekta-Neuromag Ltd., Helsinki, Finland) with 102 magnetometers and 204 planar gradiometers in a magnetically shielded room (AK3B, Vakuumschmelze, Hanau, Germany). The MEG signal was online band-pass filtered between 0.1 Hz and 330 Hz. A signal space separation algorithm implemented in the Maxfilter software (version 2.215) provided by the MEG manufacturer was used to remove external noise from the MEG signal (mainly 16.6 Hz from Austrian local train power and 50 Hz and harmonics from the power line). Data were collected across 10 50-trial blocks and were aligned to an individual common head position (based on the measured head position at the beginning of each block).

## 2.3. MEG data analysis

Data preprocessing was performed using Matlab R2020b (The MathWorks, Natick, Massachusetts, USA) Fieldtrip toolbox [20] and in-house scripts. To identify eye blinks and heartbeat artifacts, 50 independent components were identified from filtered (1 - 100 Hz) continuous magnetometer data of the first block (10 blocks in total). On average 2.5 ± 0.53 (SD) artifact components were removed per subject. All data were filtered between 0.1 Hz and 30 Hz using a finite impulse response (FIR) filter with Kaiser window. Then, the data were resampled to 100 Hz to save computational power and were epoched from -0.5s to 3s for further analysis.

To localize the activity, individual head shapes were digitized for each participant including fiducials and at least 300 points on the scalp using a Polhemus Fastrak system (Polhemus, Vermont, USA). A template structural magnetic resonance image (MRI) from Montreal Neurological Institute (MNI) and a corresponding 3-d volumetric source space grid (1 cm resolution) was warped to match the individual head shape and fiducials. Source reconstruction was then conducted by a linearly constrained minimum variance (LCMV) beamformer [21]. Source-space grid points localized to six regions of interest (ROIs) were used in the further analysis: bilateral primary auditory cortex (PAC: Brodmann A41/42) and rostral and caudal secondary auditory cortex (SAC: Brodmann A22r and A22c), based on the brainnetome atlas [22] (Figure 3).

Using a support vector machine (SVM), we first tested the sound selectivity of the six ROIs, hypothesizing that the primary and secondary auditory cortex represent different types

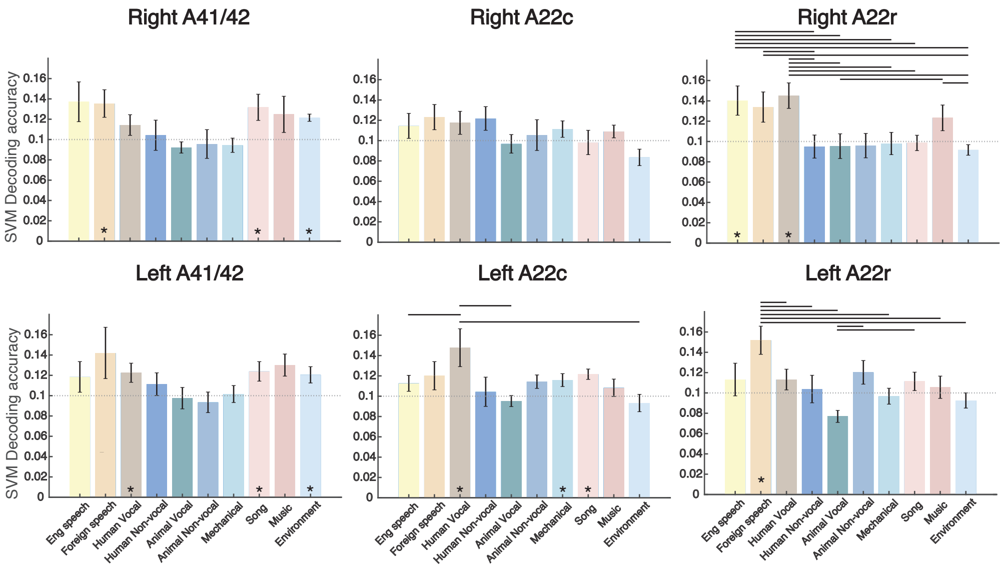


Figure 4: *SVM decoding accuracy of the 6 ROIs across ten sound categories. The asterisks close to the bottom indicate significantly higher accuracy than 10% (chance level). The black horizontal bars indicate significant pairwise comparisons between sounds.*

of sounds, with more significant sounds such as human vocalization, speech, song and music evoking responses in more specialized areas in the secondary auditory cortex [2, 3, 23, 24, 25]. Time series of MEG responses pooled by averaging within each ROI were used to decode the ten sound categories in a ten-class classification task by SVM [26], implemented in Python based on the Scikit-learn library [27]. The MEG data and corresponding labels were split into the training set (70%) and the testing set (30%). An RBF kernel was chosen with a penalty parameter C = 0.7 and a "one-vs-rest" decision function. The sound selectivity was quantified by the main effect of sound categories on the SVM decoding accuracy on the testing set, using a one-way analysis of variance (ANOVA) and then pairwise comparisons.

### 2.4. DNNs trained by the sound classification task

We trained three widely used DNNs, standard pre-activation ResNet-20 [8, 9], wide ResNet-40-4 [15] and U-Net [16], to classify sounds from the same Natural Sound Stimulus set used in the MEG experiments, using all sounds from the dataset to increase the sample size. In total, there are 11 sound classes with 98 training examples and 67 validation examples. Each example is a 2-second wav file at 44.1 kHz converted to spectrogram magnitude for model input. Model output is an 11-element vector representing the probability of each class. All models are trained for 240 epochs by minimizing the cross-entropy loss with the stochastic gradient descent using the initial learning rate 0.1, weight decay 0.0005, momentum 0.9, and batch size 8. The learning rate is decreased by a factor of 5 after 60, 120, 160, and 200 epochs. For the ResNet-20 and wide ResNet-40-4, we extracted the representations generated by the last layer in each block. For the U-Net, we extracted the representations given by each level in the encoder.

### 2.5. Compare the hierarchical processing between DNNs and auditory cortical responses

We quantify the similarity between the representations learned from a DNN and the MEG signals measured from a brain by the correlation coefficient under a linear relationship for a given matrix of model units $\mathbf{X}$ and a cortical response $\mathbf{y}$. Because the data is limited, we find the optimal solution in the ridge regression sense by $\mathbf{w}^* = \arg\min_{\mathbf{w}} \quad \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$ to prevent overfitting where there are 50 sounds and $\lambda > 0$ is a regularization parameter chosen by a grid search between $10^{-30}$ and $10^{30}$ with a 10-fold cross-validation. The model units of each layer with 3 dimensions, including time, frequency, and kernel, were averaged across kernels and transformed into a long vector for each of the 50 sounds. The variance of the cortical responses to each sound was extracted for each ROI. Figure 1 illustrates our evaluation. We predict that the primary "core" area's source activities are better correlated with earlier layers than later layers, while the secondary "belt" areas' activities are better correlated with later layers than earlier layers.

## 3. Results

### 3.1. Cortical selectivity

To test cortical selectivity for sound classes, we first verified if the SVM decoding accuracy was significantly higher than the chance level (i.e. 10%) for any sound classes, marked by asterisks in Figure 4. The secondary auditory areas showed significantly higher decoding accuracy for English speech (right A22r $t(9)$ = 2.654, $p$ = 0.026), song (left A22c $t(9)$ = 4.011, $p$ = 0.003), foreign speech (left A22r $t(9)$ = 3.548, $p$ = 0.006), and human vocal sounds (right A22r $t(9)$ = 3.407, $p$ = 0.008). In contrast, there was marginal to no significance for the primary auditory cortex except for the foreign speech (right A41/42 $t(9)$ = 2.48, $p$ = 0.035) and environmental sounds (right A41/42 $t(9)$ = 5.744, $p < 0.000$; left A41/42 $t(9)$ = 2.424, $p$ = 0.038).

A one-way ANOVA was used to test selectivity. Only the secondary auditory areas showed a significant main effect of sound categories on SVM decoding accuracy, including left A22c ($F(9, 90)$ = 1.95, $p$ = .054), right A22r ($F(9, 90)$ = 3, $p$ = .004) and left A22r ($F(9, 90)$ = 2.84, $p$ = .005). These ROIs were sensitive to specific sound categories such that the left A22c had the highest decoding accuracy for the human vo-

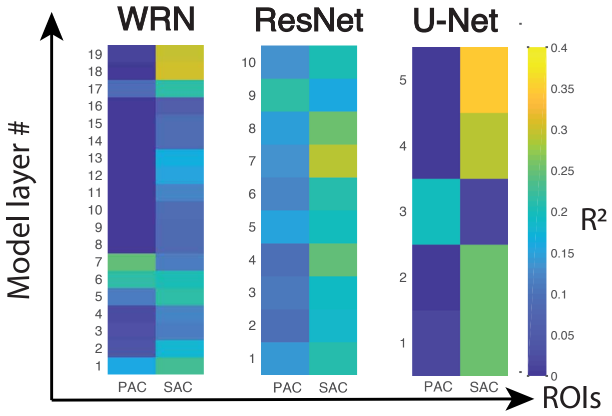


Figure 5: *Brain sound selectivity correlated with model units from each DNN layer. The columns represent the left A41/42 from the PAC and the right A22r from the SAC, which was highly selective to speech, song, and human vocal sounds. The y-axis represents layers, depending on different DNN architectures. The colors indicate the correlation coefficient squared.*

cal sounds. The right A22r had high decoding accuracy for human vocal sounds, English speech, and foreign speech. The left A22r had the highest decoding accuracy for foreign speech. The three significant one-way ANOVA analysis results were further broken down to pairwise $t$-tests between sound categories (See Figure 4 horizontal bars). Results generally showed higher decoding accuracy for speech, human vocal sounds, and song than other non-human sounds in the secondary auditory areas, one exception being right A22c, which was not significant ($F(9, 90) = 1.17$, $p = .327$). There was no significant selectivity observed in the primary auditory areas, right A41/42 ($F(9, 90) = 1.66$, $p = .112$) and left A41/42 ($F(9, 90) = 1.28$, $p = .261$). Note that a few accuracy values are lower than the chance level, which is likely due to the small sample size.

### 3.2. Brain signal variance correlated with DNN model units

In all three DNN models, the validation accuracy of the sound classification was higher than the chance level (i.e. 10%), with WRN-40-4 (9.0M parameters) having the best validation accuracy 47.76% with equal performance in the ResNet-20 (0.27M parameters) and U-Net (31M parameters) at 40.30%. We focused on comparing sound selectivity profiles in model units of the best-performing model, WRN-40-4, to selectivity described in the primary and secondary auditory cortices. Among the 6 ROIs, the variance of neural signals from the right A22r, the region with the greatest selectivity, was highly correlated (i.e. higher R-squared) with later WRN-40-4 model layers compared to earlier layers. In contrast, the variance of the left A41/42 was slightly higher correlated in the earlier WRN-40-4 layers. Results for all three candidate models are shown in Figure 5.

## 4. Conclusion and Discussion

We asked how human brains and DNN models process naturalistic sounds. We specifically investigated whether the human auditory cortex and DNNs use similar hierarchical principals and if they similarly privilege more relevant human vocal sounds such as speech from other everyday sounds in higher-order levels in the cortex and in the DNNs. Indeed, we found that higher decoding accuracy was found for speech, human vocal sounds, and song compared to other non-human sounds and this effect was most pronounced in the secondary auditory cortex compared to the primary auditory cortex. Furthermore, these highly selective secondary areas bore the most resemblance to the later, and more complex DNN layers (as assessed by correlation coefficient). This finding suggests comparable hierarchical principles of DNN and neural processing of sounds. Our finding, in human auditory processing, using MEG brain recording and advanced DNNs is consistent with previous research in the visual system primarily using primate electrophysiology [12, 13, 28].

This is the first study to our awareness using quiet, non-invasive, and time-sensitive MEG to probe the parallel hierarchical processing of naturalistic sounds in human auditory cortex and advanced DNNs with skip connections trained to perform sound classification tasks. Our findings of sound selectivity, especially speech selectivity, in the secondary auditory cortex, are consistent with previous fMRI studies [2, 23, 24] and ECoG works [3, 25]. Moreover, the finding of similar hierarchical processing in human brains and current DNN models is consistent with [14], but further corroborates it with a general-purpose neural network architecture that did not have any a priori assumptions about sound categories built into its architecture.

We used broadband MEG signals without much data preprocessing, but it is well known that there are many specialized temporal and frequency-specific signals found within the broadband MEG signal. Future research may try seeking and engineering the most relevant time-frequency features from the MEG signal for a better performance. For example, applying temporal response function (TRF, forward modeling) on natural continuous sounds may be particularly relevant [29]. Another useful tool in addition to the source localization used in this study may be principal/ independent component analysis, which can localize cortical sources based on the specific sound features [25, 30]. Regarding our SVM results, the MEG-based sound classification accuracy was above chance and comparable to a similar study conducting electroencephalography(EEG)-based classification of natural sounds [31]. Future studies may recruit a larger sample size to increase SNR and statistical power, and more importantly for a better SVM model fitting. Similarly, while we found significant and interpretable results from DNNs, the sample sizes were small in comparison to many DNN training studies and thus future studies with larger sample sizes would be a useful replication. Finally, other DNNs may have even better performance by incorporating temporal information such as recurrent networks and Transformers [32], or even more high-level language models such as Bidirectional Encoder Representations from Transformers (BERT) [33].

In sum, this study establishes a reliable hierarchical mapping between neural networks and human brain dynamics measured using a widely-available, high temporal resolution brain imaging modality (MEG) and general-purpose DNNs. These results can inform future studies targeting larger populations and real-life applications. Our findings speak for the sound processing of the brain, and further suggest similar hierarchical principals in DNNs and the human auditory cortex when processing naturalistic sounds. This study has the potential to launch a new way of studying human brains with computational models and could inspire the architectural design for state-of-the-art neural networks used for processing complex and naturalistic sounds.

# 5. References

[1] J. H. Kaas and T. A. Hackett, "Subdivisions of auditory cortex and processing streams in primates," *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 11 793–11 799, 2000.

[2] S. Norman-Haignere, N. G. Kanwisher, and J. H. McDermott, "Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition," *Neuron*, vol. 88, no. 6, pp. 1281–1296, 2015.

[3] L. S. Hamilton, Y. Oganian, and E. F. Chang, "Topography of speech-related acoustic and phonological feature encoding throughout the human core and parabelt auditory cortex," *BioRxiv*, pp. 2020–06, 2020.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[7] S. K. Zieliński, H. Lee, P. Antoniuk, and O. Dadan, "A comparison of human against machine-classification of spatial audio scenes in binaural recordings of music," *Applied sciences*, vol. 10, no. 17, p. 5956, 2020.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[9] ——, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[10] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, vol. 28, 2015.

[11] K.-L. Chen, C.-H. Lee, H. Garudadri, and B. D. Rao, "ResNEsts and DenseNEsts: Block-based DNN models with improved representation guarantees," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 3413–3424.

[12] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the national academy of sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[13] P. Bao, L. She, M. McGill, and D. Y. Tsao, "A map of object space in primate inferotemporal cortex," *Nature*, vol. 583, no. 7814, pp. 103–108, 2020.

[14] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," *Neuron*, vol. 98, no. 3, pp. 630–644, 2018.

[15] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[18] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 1, pp. 189–198, 2018.

[19] K.-L. Chen, C.-H. Lee, B. D. Rao, and H. Garudadri, "A DNN based normalized time-frequency weighted criterion for robust wideband DoA estimation," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023.

[20] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data," *Computational intelligence and neuroscience*, vol. 2011, pp. 1–9, 2011.

[21] B. D. Van Veen, W. Van Drongelen, M. Yuchtman, and A. Suzuki, "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering," *IEEE Transactions on biomedical engineering*, vol. 44, no. 9, pp. 867–880, 1997.

[22] L. Fan, H. Li, J. Zhuo, Y. Zhang, J. Wang, L. Chen, Z. Yang, C. Chu, S. Xie, A. R. Laird *et al.*, "The human brainnetome atlas: a new brain atlas based on connectional architecture," *Cerebral cortex*, vol. 26, no. 8, pp. 3508–3526, 2016.

[23] W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, and F. E. Theunissen, "The hierarchical cortical organization of human speech processing," *Journal of Neuroscience*, vol. 37, no. 27, pp. 6539–6557, 2017.

[24] J. M. Fisher, F. K. Dick, D. F. Levy, and S. M. Wilson, "Neural representation of vowel formants in tonotopic auditory cortex," *NeuroImage*, vol. 178, pp. 574–582, 2018.

[25] S. V. Norman-Haignere, J. Feather, D. Boebinger, P. Brunner, A. Ritaccio, J. H. McDermott, G. Schalk, and N. Kanwisher, "A neural population selective for song in human auditory cortex," *Current Biology*, vol. 32, no. 7, pp. 1470–1484, 2022.

[26] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.

[29] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mTRF) toolbox: A matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in human neuroscience*, vol. 10, p. 604, 2016.

[30] N. J. Zuk, J. W. Murphy, R. B. Reilly, and E. C. Lalor, "Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies," *PLoS computational biology*, vol. 17, no. 9, p. e1009358, 2021.

[31] N. J. Zuk, E. S. Teoh, and E. C. Lalor, "Eeg-based classification of natural sounds reveals specialized responses to speech and music," *NeuroImage*, vol. 210, p. 116558, 2020.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.