INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speech Taskonomy: Which Speech Tasks are the most Predictive of fMRI Brain Activity?

Subba Reddy Oota[1*], *Veeral Agarwal*[2*], *Mounika Marreddy*[2], *Manish Gupta*[2,3], *Raju Bapi*[2]

[1]Inria Bordeaux, France, [2]IIIT Hyderabad, India, [3]Microsoft, India

subba-reddy.oota@inria.fr,{veeral.agarwal,mounika.marreddy}@research.iiit.ac.in
gmanish@microsoft.com,raju.bapi@iiit.ac.in

Abstract

Self-supervised speech based models have been found to be successful in predicting brain recordings of subjects experiencing naturalistic story listening. Inspired by the recent progress on deep learning models for various speech-processing tasks, existing literature has leveraged pretrained speech Transformer models for brain encoding. However, there is no work on exploring the efficacy of *task-specific finetuned* Transformer representations for this task. Hence, in this paper, we explore transfer learning from representations finetuned for eight different tasks from Speech processing Universal PERformance Benchmark (SUPERB) for predicting brain responses. Encoding models based on task features are used to predict activity in different regions across the whole brain, and also in language and auditory brain regions. Our experiments on finetuning the Wav2Vec2.0 model for these eight tasks show that the model finetuned on automatic speech recognition (ASR) yields the best encoding performance for the whole brain, language and auditory regions.
Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

In computational cognitive science, brain encoding is the problem of predicting brain activations from stimuli. For the past two decades, researchers have focused on mapping stimulus representations to brain activations through encoding models for text and vision. For text, researchers have explored both syntactic as well as semantic representations, the most recent ones using Transformer-based deep learning (DL) methods [1, 2, 3, 4]. For vision, the encoding models have leveraged our algorithmic understanding of visual hierarchy (V1, V2, V4, IT) in the visual cortex and hence follow the convolutional neural network-based design [5, 6]. Recent advancements in deep learning models for speech [7, 8, 9, 10] have motivated neuroscience researchers to leverage them for auditory brain encoding [11, 12, 13, 14] starting early 2022.

Millet et al. [11] used a pretrained deep learning model Wav2Vec2.0 [8] to learn latent representations of the speech waveform similar to those of the human brain. They find that the functional hierarchy of its transformer layers aligns with the cortical hierarchy of speech in the brain, and reveals the whole-brain organisation of speech processing. Similarly, Vaidya et al. [12] experimented with four pretrained speech representation methods (APC [7], Wav2Vec [15], Wav2Vec2.0 [8], and HuBERT [9]) and found that Wav2Vec2.0 aligns best with the human auditory system. Further, Oota et al. [14] and Tuckute et al. [13] extended this analysis to more such deep learning based speech models. However, no existing work has investigated the implications of finetuning speech pretrained models for various speech-processing tasks for speech encoding in the brain.

Although pretrained speech models can understand broad aspects of speech in general, their representations are not specifically tuned towards capturing distinctive characteristics of speech tasks. When participants are listening to speech narratives, it is plausible that they are engaged in various speech tasks such as phoneme and speech recognition, emotion recognition, comprehending the intent, keyword identification, attending to speaker characteristics, etc. Thus we expect that the brain activation recorded during narrative listening would be related to various speech tasks. Hence, toward designing a better brain speech-processing pipeline when subjects listened to naturalistic stories, in this paper, we finetune a pretrained speech model (Wav2Vec2.0) on eight different speech-processing tasks. Our goal in this paper is to find which of these eight finetuned models best captures distinctive characteristics of story listening and hence leads to the best encoding accuracy.

Indeed explorations around which finetuned models lead to better encoding accuracy compared to pretrained ones, have already been done rigorously for text and vision. For text, several researchers [3, 16, 17, 18, 19] finetuned a pretrained language model (BERT [20]) for multiple tasks from General Language Understanding Evaluation (GLUE) benchmark [21], and found that using a finetuned BERT leads to improved brain predictions. For visual stimuli, Wang et al. [22] finetuned a pretrained vision model (ResNet50 [23]) for multiple 2D and 3D computer vision tasks. They found that models finetuned for 3D vision tasks lead to better encoding accuracy compared to pretrained models alone. Inspired by the success of finetuning in the language and vision fields, we build neural speech taskonomy models for brain encoding and aim to find speech-processing tasks that have the most explanatory capability of brain activation during naturalistic story listening experiments.

Wav2Vec2.0 [8] is a popular state-of-the-art model for several speech-processing tasks. The robust pretraining helped the model to achieve the state-of-the-art word error rate on the benchmark *LibriSpeech* speech-to-text dataset requiring finetuning with just one hour worth of labeled data [8]. Hence, we experiment with Wav2Vec2.0 as our pretrained model. Further, Speech processing Universal PERformance Benchmark (SUPERB) [24] is a collection with a wide range of speech processing tasks that captures different abilities of human speech processing related to how humans produce, perceive, and understand speech. Therefore, in the hope of discovering a task that best captures distinctive characteristics of story listening, we finetune Wav2Vec2.0 on the following eight SUPERB tasks, and evaluate their brain encoding performance: Phoneme Recognition (PR), Automatic Speech Recognition (ASR), Key-

The first two authors made equal contribution.

10.21437/Interspeech.2023-121

word Spotting (KS), Intent Classification (IC), Speaker Diarization (SD), Speaker Verification (SV), Speaker Identification (SID), and Emotion Recognition (ER). We chose these eight tasks of the thirteen available in SUPERB as these would benefit from finetuning of pretrained models and are also more relevant for the speech encoding models we plan to investigate.

Overall, we make the following contributions in this paper.

- Given Transformer models finetuned for eight speech tasks, we propose the problem of finding which of these are the most predictive of fMRI brain activity for story listening.
- We show that task-specific (ASR, ER, SID and IC) speech representations lead to a significant improvement in brain alignment compared to the pretrained Wav2Vec2.0 model for specific brain regions. Finetuning on ER, SID and IC leads to the best alignment for the early auditory cortex; finetuning on ASR provides the best encoding for the auditory associative cortex and language regions.
- Layer-wise analysis of the effect of each speech task on the alignment with whole brain activity shows that (a) the ASR task is better aligned in middle layers, and (b) performance degrades drastically for later layers, specifically for SD and PR tasks.

2. Task Descriptions

We experiment with the eight SUPERB tasks briefly described as follows. **Phoneme Recognition (PR)** transcribes an utterance into the smallest content units. **Automatic Speech Recognition (ASR)** transcribes utterances into words. **Keyword Spotting (KS)** detects preregistered keywords by classifying utterances into a predefined set of words. **Intent Classification (IC)** classifies utterances into predefined classes to determine the intent of speakers. **Speaker Diarization (SD)** classifies each utterance for its speaker identity as a multi-class classification, where speakers are in the same predefined set for both training and testing. **Speaker Verification (SV)** verifies whether the speakers of a pair of utterances match as a binary classification, and speakers in the testing set may not appear in the training set. **Speaker Identification (SID)** predicts who is speaking when for each timestamp, and multiple speakers can speak simultaneously. **Emotion Recognition (ER)** predicts an emotion class for each utterance. These tasks check for performance across several cognitive speech perception skills like recognition (PR and ASR), detection (KS), semantics (IC, SF, and ST), speaker-related (SV, SD, and SID), and paralinguistics (ER). Several speech tasks, such as Query by Example (QbE), Slot Filling (SF), Speech Enhancement (SE), and Speech Separation (SS), do not require fine-tuning. Hence, we have not chosen these models for our study.

Our selection of these tasks was based on the following design principles: (1) We wanted to select a set of tasks covering diverse cognitive speech perception skills. (2) We wanted to select tasks that are a part of popular speech benchmark like SUPERB [21]. (3) We selected tasks for which Wav2vec2.0-base finetuned models were available.

3. Dataset and Experiments

The "Narratives" collection aggregates a variety of fMRI datasets collected while human subjects listened to real spoken stories [25]. We analyze data from 82 subjects listening to the story titled 'PieMan' with 282 TRs (repetition time – fMRI recorded every 1.5 sec.). The dataset is in English and contains 957 words across 67 sentences. The dataset was already preprocessed and projected on the surface space ("fsaverage6").

We use the multi-modal parcellation of the human cerebral cortex (Glasser Atlas: consists of 180 ROIs (regions of interest) in each hemisphere) to display the brain maps [26], since the Narratives dataset contains annotations tied to this atlas. The data covers both auditory and language brain ROIs with the following subdivisions: (i) early auditory cortex (EAC: A1, A2, LBelt, MBelt, PBelt); (ii) auditory association cortex (AAC: STSda, STSva, STSdp, STGa, TE1a, TE2a); and (iii) inferior frontal gyrus (IFG: 44, 45, IFJa, IFSp) [27, 28, 29]. The dataset has been made available freely without restrictions by [25].

The input audio story is first segmented into clips corresponding to 1 TR. Each audio clip is input to the speech models one by one to obtain stimulus representations per clip. The representations are obtained by probing the pretrained speech model and taking the output from different encoder layers. For all the models, we used the checkpoints provided by the huggingface library. Since we extracted the features at each TR, downsampling is not required for further sampling the frequency of each feature dimension.

BOLD (Blood oxygenation level-dependent) fMRIs measure brain activity by detecting changes associated with blood flow. When an area of the brain is in use, blood flow to that region also increases. It takes a while for the vascular system to respond to the brain's need for glucose. Thus, blood flow lags the neuronal events triggering by a few seconds. This hemodynamic response is typically modeled using a finite response filter (FIR) per voxel. We model this for each subject separately with eight temporal delays corresponding to around 12 secs. This means that we predict the brain activations at the t^{th} time point based on the concatenation of speech representation for audio clips corresponding to $(t-8)^{th}$ to t^{th} time point.

4. Methodology

Voxelwise Encoding Model The main goal of each fMRI encoder model is to predict brain responses associated with each brain voxel given a stimuli. To explore how speech signals are encoded in the brain when listening to stories, we use layerwise pretrained Wav2Vec2.0 features in a voxelwise encoding model to predict brain responses. We train fMRI encoding models using Banded ridge regression [30] on stimuli representations from various feature spaces. Before doing regression, we first z-scored each feature channel separately for training and testing. This was done to match the features to the fMRI responses, which were also z-scored for training and testing. The solution to the banded regression approach is given by $f(\hat{\beta}) = \underset{\beta}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\beta\|_F^2 + \lambda\|\beta\|_F^2$, where $\mathbf{Y}$ denotes the voxels matrix across TRs, β denotes the learned regression coefficients, and $\mathbf{X}$ denotes stimulus representations.

Cross-Validation We follow 4-fold (K=4) cross-validation. All the data samples from K-1 folds were used for training, and the model was tested on samples of the left-out fold.

Evaluation Metric: We evaluate our models using popular brain encoding evaluation metric, Pearson Correlation (PC), as described in the following. Given a subject and a brain region, let N be the number of samples. Let $\{Y_i\}_{i=1}^N$ and $\{\hat{Y}_i\}_{i=1}^N$ denote the actual and predicted voxel value vectors for the i^{th} sample. Thus, $Y \in \mathbb{R}^{N\times V}$ and $\hat{Y} \in \mathbb{R}^{N\times V}$ where V is the number of voxels in that region. Let $\{\hat{E}_i\}_{i=1}^N$ denote the stimuli representation for the i^{th} sample. Thus, $E \in \mathbb{R}^{N\times D}$ where D is the dimensionality of the encoded representation. PC is

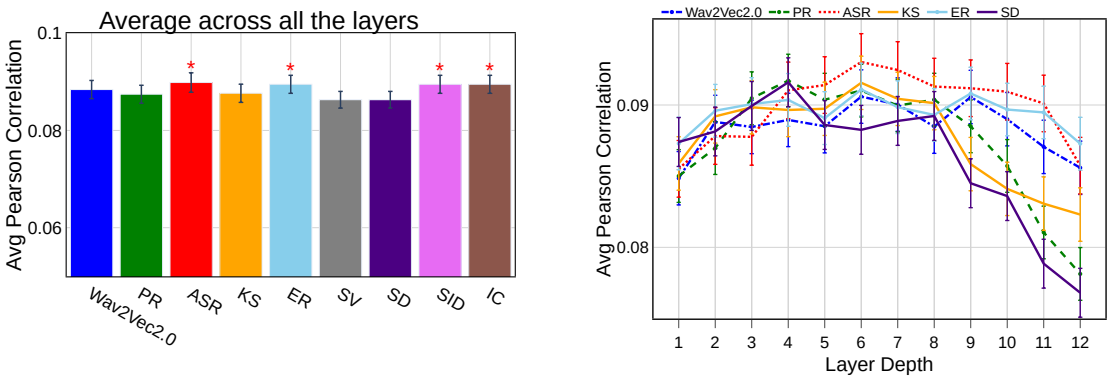

Figure 1: *Brain alignment of pretrained Wav2Vec2.0 (blue) and different finetuned tasks. The left plot compares the average Pearson correlation across all layers of pretrained Wav2Vec2.0 and all voxels, and the same quantity for each downstream task. Error bars indicate the standard error of the mean across participants and '*' indicates that the particular task PC is significantly higher than pretrained Wav2Vec2.0. The right plot compares the layer-wise performance of pretrained Wav2Vec2.0 and the finetuned speech tasks.*

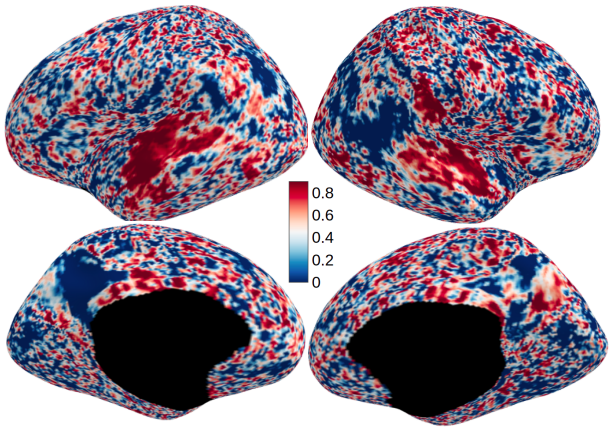

Figure 2: *Voxel-wise correlation values for the brain alignment of pretrained Wav2Vec2.0 and ASR task across all the layers for the ASR finetuned model.*

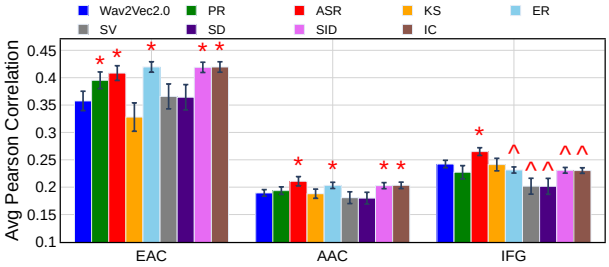

Figure 3: *Brain alignment of pretrained Wav2Vec2.0 (blue) and different finetuned tasks. Plot compares the average Pearson correlation across all layers of pretrained Wav2Vec2.0 and brain ROIs. Error bars indicate the standard error of the mean across participants. A '*' at a particular bar indicates that the task is significantly better than pretrained Wav2Vec2.0, whereas ∧ denotes that the task performance is significantly lower.*

then computed as PC=$\frac{1}{N}\sum_{i=1}^{N} corr[Y_i, \hat{Y}_i]$ where corr is the correlation function.

Statistical Significance To estimate the statistical significance of the performance differences, we performed two-tailed paired-sample t-tests on the mean correlation values for the subjects. Further, the Benjamni-Hochberg False Discovery Rate (FDR) correction [31] is used for all tests (appropriate because fMRI data is considered to have positive dependence [32]). The correction is performed by grouping all the subject-level p-values (i.e., across each speech task and pretrained Wav2Vec2.0) and choosing one threshold for all results.

Implementation Details for Reproducibility All experiments were conducted on a machine with 1 NVIDIA GEFORCE-GTX GPU with 16GB GPU RAM. We used banded ridge-regression with the following parameters: MSE loss function, and L2-decay (λ) varied from 10^{-1} to 10^{-3}; best λ was chosen by tuning on validation data.

5. Results

In order to assess the performance of the fMRI encoder models learned using the representations from a variety of speech tasks, we computed the Pearson correlation coefficient between the predicted and true responses across whole brain voxels, various auditory and language ROIs, and sub-ROIs.

Whole Brain Results In Fig. 1 (left), we present the average brain alignment across all layers of pretrained Wav2Vec2.0 and for each speech task. In comparison to Wav2Vec2.0, ASR shows improvement in correlation, whereas tasks including ER, SID and IC report similar correlation performance. Further, tasks such as SD and SV yield lower correlation, which suggests that speaker diarization and verification are not important in listening to stories. This result suggests that there are certain speech tasks that are important for improved brain alignment over pretrained Wav2Vec2.0. The effect of the two-tailed test was significant for the tasks with pretrained Wav2Vec2.0, p-values are as follows: ASR (0.0287), ER (0.0341), SID (0.0341), and IC (0.0341). For the remaining tasks, the p-values with pretrained Wav2Vec2.0 as follows: PR (0.893), KS (0.862), SV (0.944), and SD (0.944). Overall, ASR, ER, SID and IC are statistically significant across all speech tasks with pretrained Wav2Vec2.0.

In Fig. 1 (right), we also report the layer-wise performance for Wav2Vec2.0 and all speech tasks. Similar to previous work [11], we observe that pretrained Wav2Vec2.0 has the best brain alignment in the early and in later layers. On the other

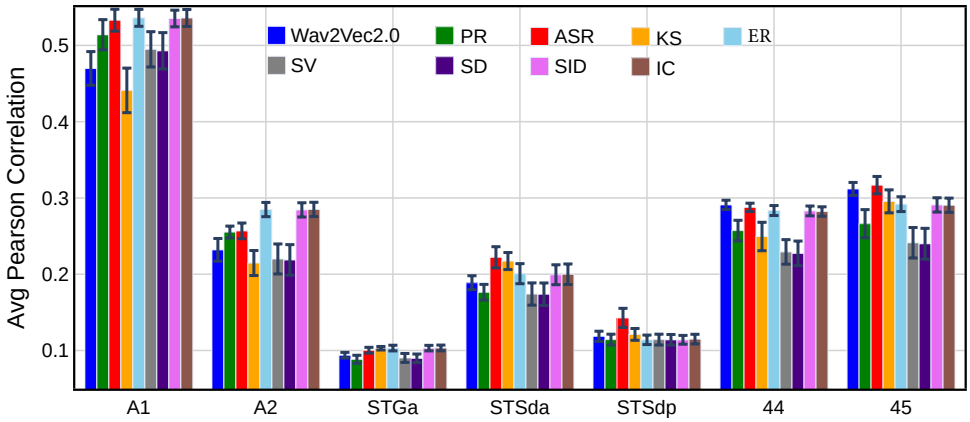

Figure 4: *Brain alignment of pretrained Wav2Vec2.0 (blue) and different finetuned tasks. Plot compares the average Pearson correlation across all layers of pretrained Wav2Vec2.0 and sub-ROIS. The error bars indicate the standard error of the mean across participants.*

hand, the ASR task has the best brain alignment in the middle layers. We further observe that the alignment after finetuning on a speech task is significantly worse, mainly for later layers. This pattern holds across speech tasks, including PR and SD. These results provide direct evidence that the later layers of finetuned task representations have more task-specific information. Since phoneme parsing is an early processing step in speech understanding, as expected, the early layers of PR (phoneme recognition) finetuned model are well aligned with brain activity. Overall, these results demonstrate that when listening to a story, information processing operations related to recognizing words (ASR), intent present in the sentence (IC), the emotion of the story (ER), and speaker identification (SID) processing may be engaged.

Fig. 2 presents the voxel-wise correlation values for the brain alignment of pretrained Wav2Vec2.0 and each speech task across all the layers for the ASR task. Low correlations in some regions indicate that finetuning changes predictions for those regions. We observe that the correlation is high in temporal lobes but not in language regions and parietal regions. Thus, ASR leads to better language understanding compared to pretrained Wav2Vec2.0. Perhaps that is why, like language models, the ASR model also has the best performance for middle layers [2].

ROI Level Results We further examine the effect on the alignment specifically in a set of ROIs that are thought to underlie speech and language comprehension, as shown in Fig. 3. We make the following observations: (1) All speech tasks are better aligned with EAC compared to AAC and IFG regions. (2) ASR task has a higher brain Pearson correlation than other tasks in AAC and IFG ROIs (3) Since ASR task handles both phoneme and word recognition, we observe that ASR task is responsible for the better brain alignment in language ROIs, AAC and IFG. On the other hand, the KS task is based on keyword extraction; hence, IFG has a better correlation, whereas EAC has a lower correlation. Similarly, PR is better aligned with EAC which is responsible for the processing of phonemes and sounds. (4) ER, SID, IC and pretrained Wav2Vec2.0 showed similar correlation for whole brain (Fig. 1). But Fig. 3 shows that these sub-regions have a higher correlation compared to Wav2Vec2.0 in the EAC region. (5) Finally, across all ROIs, the pretrained Wav2Vec2.0 model has a worse correlation compared to at least five other speech task models.

Sub-ROI Level Results We further demonstrate the prediction performance of the encoder model for sub-ROIs across speech tasks, as shown in Fig. 4. It can be observed that the sub-ROI of early auditory cortex (EAC: A1) has a higher Pearson correlation than other sub-ROIs. On the other hand, sub-ROIs A2 from AAC, 44 and 45 from IFG display similar correlation performance. However, the sub-ROIs in the AAC (STGa and STSdp) yield a lower correlation. The language ROIs 44 and 45, together with STSda and STSdp in the AAC, are part of the well-known language network associated with narrative comprehension [33], and it is heartening to see that task features from ASR task show the best correlation in these regions.

6. Discussion and Conclusion

We evaluated the processing of finetuned representations of eight speech tasks and their alignment with brain responses. We showed that the representations of multiple speech task finetuned models (ASR, PR, ER, SID and IC) compared to pretrained Wav2Vec2.0 lead to a significant increase in brain alignment across auditory regions. To understand the contribution of each speech task to the brain alignment better, we performed multiple analyses: layer-wise correlations, and analyses at ROI and sub-ROI levels. We find that ASR is better aligned than Wav2Vec2.0 in all analyses – whole brain, ROI and sub-ROI level. Our results (Figs. 3 and 4) also seem to support that ASR best captures activity in both auditory and language regions.

Recently, Shah et al. [34] used 43 probing tasks to uncover the parts of the two popular self-supervised speech models (Mockingjay and Wav2Vec2.0) that encode specific speech properties like (i) audio features: total duration, local jitter, (ii) fluency features: filled pause rate, mean silence, (iii) pronunciation features: stressed syllable percent, mean stress distance syllable, and (iv) semantic level text features: total nouns and total adjectives. These techniques have revealed a hierarchy of information processing in multi-layered speech models that progresses from simple to complex with increased depth. To better understand the reasons for the better alignment between task-specific speech models and the brain, future work can focus on investigating the correspondence between the detailed processing of underlying speech properties by the human brain vs. self-supervised speech models, similar to alignment between linguistic properties of language models and human brains [35].

This work was done on native English speakers' data related to English stories only. More work needs to be done to verify which of these insights hold for datasets in other languages, and for non-native English speakers.

7. References

[1] J. Gauthier and R. Levy, "Linking artificial and human neural representations of language," *arXiv:1910.01244*, 2019.

[2] M. Toneva and L. Wehbe, "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)," *arXiv:1905.11833*, 2019.

[3] D. Schwartz, M. Toneva, and L. Wehbe, "Inducing brain-relevant bias in natural language processing models," *NeurIPS*, vol. 32, pp. 14 123–14 133, 2019.

[4] S. R. Oota, F. Alexandre, and X. Hinaut, "Long-term plausibility of language models and neural dynamics during narrative listening," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 44, no. 44, 2022.

[5] R. Beliy, G. Gaziv, A. Hoogi, F. Strappini, T. Golan, and M. Irani, "From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri," *arXiv:1907.02431*, 2019.

[6] C. Du, C. Du, L. Huang, and H. He, "Conditional generative neural decoding with structured cnn feature prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2629–2636.

[7] Y.-A. Chung, H. Tang, and J. Glass, "Vector-quantized autoregressive predictive coding," *Interspeech*, pp. 3760–3764, 2020.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, vol. 29, pp. 3451–3460, 2021.

[10] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv:2202.03555*, 2022.

[11] J. Millet, C. Caucheteux, P. Orhan, Y. Boubenec, A. Gramfort, E. Dunbar, C. Pallier, and J.-R. King, "Toward a realistic model of speech processing in the brain with self-supervised learning," *arXiv:2206.01685*, 2022.

[12] A. R. Vaidya, S. Jain, and A. G. Huth, "Self-supervised models of audio effectively explain human cortical responses to speech," *arXiv preprint arXiv:2205.14252*, 2022.

[13] G. Tuckute, J. Feather, D. Boebinger, and J. H. McDermott, "Many but not all deep neural network audio models capture brain responses and exhibit hierarchical region correspondence," *bioRxiv*, pp. 2022–09, 2022.

[14] S. R. Oota, K. Pahwa, M. Marreddy, M. Gupta, and B. S. Raju, "Neural architecture of speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Proc. Interspeech 2019*, pp. 3465–3469, 2019.

[16] S. Kumar, T. R. Sumers, T. Yamakoshi, A. Goldstein, U. Hasson, K. A. Norman, T. L. Griffiths, R. D. Hawkins, and S. A. Nastase, "Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model," *BioRxiv*, pp. 2022–06, 2022.

[17] K. L. Aw and M. Toneva, "Training language models for deeper understanding improves brain alignment," *arXiv preprint arXiv:2212.10898*, 2022.

[18] G. Merlin and M. Toneva, "Language models and brain alignment: beyond word-level semantics and prediction," *arXiv preprint arXiv:2212.00596*, 2022.

[19] S. R. Oota, J. Arora, V. Agarwal, M. Marreddy, M. Gupta, and B. R. Surampudi, "Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity?" *arXiv preprint arXiv:2205.01404*, 2022.

[20] 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.

[21] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[22] A. Wang, M. Tarr, and L. Wehbe, "Neural taskonomy: Inferring the similarity of task-derived representations from brain activity," *NeurIPS*, vol. 32, pp. 15 501–15 511, 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech*, 2021, pp. 1194–1198.

[25] S. A. Nastase, Y.-F. Liu, H. Hillman, A. Zadbood, L. Hasenfratz, N. Keshavarzian, J. Chen, C. J. Honey, Y. Yeshurun, M. Regev *et al.*, "The "narratives" fmri dataset for evaluating models of naturalistic language comprehension," *Scientific data*, vol. 8, no. 1, pp. 1–22, 2021.

[26] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson *et al.*, "A multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, 2016.

[27] C. M. Baker, J. D. Burks, R. G. Briggs, A. K. Conner, C. A. Glenn, K. N. Taylor, G. Sali, T. M. McCoy, J. D. Battiste, D. L. O'Donoghue *et al.*, "A connectomic atlas of the human cerebrum—chapter 7: the lateral parietal lobe," *Operative Neurosurgery*, vol. 15, no. suppl_1, pp. S295–S349, 2018.

[28] C. K. Milton, V. Dhanaraj, I. M. Young, H. M. Taylor, P. J. Nicholas, R. G. Briggs, M. Y. Bai, R. D. Fonseka, J. Hormovas, Y.-H. Lin *et al.*, "Parcellation-based anatomic model of the semantic network," *Brain and behavior*, vol. 11, no. 4, p. e02065, 2021.

[29] R. H. Desai, U. Tadimeti, and N. Riccardi, "Proper and common names in the semantic system," *Brain Structure and Function*, vol. 228, no. 1, pp. 239–254, 2023.

[30] A. N. Tikhonov, V. J. Arsenin, and V. Arsenin, *Solutions of ill-posed problems*. V H Winston, 1977.

[31] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[32] C. R. Genovese, "A bayesian time-course model for functional magnetic resonance imaging data," *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 691–703, 2000.

[33] S. A. Nastase, Y.-F. Liu, H. Hillman, K. A. Norman, and U. Hasson, "Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space," *NeuroImage*, vol. 217, p. 116865, 2020.

[34] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[35] S. R. Oota, M. Gupta, and M. Toneva, "Joint processing of linguistic properties in brains and language models," *arXiv preprint arXiv:2212.08094*, 2022.