Modeling Articulation and Acoustics with WavLM

Daniela Wiepert[1], *Lasya Yakkala*[1], *Rachel Yamamoto*[1]

[1]Department of Computer Science, University of Texas at Austin, USA

Abstract

Understanding how the sensorimotor cortex encodes information during speech perception remains a challenge for neuroscience due to the inherent connection between articulation and acoustics. Recent advances in explaining brain activity with self-supervised (SSL) speech representations have opened a new pathway for disentangling articulation and acoustics in brain activity by separating SSL feature spaces into orthogonal articulatory and acoustic components. In this work, we explore whether WavLM, an SSL speech model that has been used to extract a set of exact articulatory measurements with high accuracy, can be used to model articulation and acoustics separately. We extracted three primary feature sets (WavLM; estimated electromagnetic articulography (EMA) features derived from a linear mapping between WavLM and the EMA feature space; orthogonal residuals derived from a least-squares regressions of EMA features to WavLM features) and two dimension-matched feature sets (EMA-Wav, the predicted WavLM features from the least-squares regression; pca-residuals, a reduced dimension residual features set) and probed them for acoustic, articulatory, and semantic information.

We found that EMA-based features tended to capture more articulatory information and less acoustic information than the residual-based features, demonstrating potential for disentangling acoustic and articulatory components in the WavLM feature space. We also showed that the rank of a feature set had a significant impact on the type of information it captured. This highlighted the importance of rank-matching the acoustic and articulatory components of WavLM for accurate modeling.

Index Terms: SSL representations, linear probing, articulation, acoustics, fMRI

1. Introduction

Studies on the functional neuroanatomy of speech have established that speech production and perception tasks are linked with activity in the sensorimotor cortex (SMC), but there is debate on how this region represents and processes information during the two related, but distinct tasks [1, 2, 3, 4, 5]. Neural activity in the SMC is correlated with a range of motor control and sensory processing tasks. Activity in sub-regions of the SMC during speech production exhibit somatotopic organization of speech articulators, meaning that specific areas of the cortex correspond directly to specific articulators [2]. This region is also crucially involved in the sensory-motor integration during active articulation, wherein sensory information regarding the state of the vocal tract is used to guide the articulators to produce a desired sound [5]. It is generally agreed that the SMC encodes motor representations during speech production given the functional organization of neural activity and the motor output demanded by the task.

The role of the SMC is less clear for speech perception, which does not explicitly involve motor control or require sensory-motor integration. Two competing types of theories have emerged to describe SMC activity during speech perception, with one category claiming that the SMC encodes motor representations of articulation during speech perception and the other suggesting that the SMC is processing acoustic sensory representations. No theory has emerged as the clear winner as the inherent connection between acoustics and articulation has made it difficult to disentangle them in analyses.

The most well-known motor representation-based theory is the Motor Theory of Speech Perception, which posits that perceiving speech necessarily involves perceiving the articulatory gestures that produced it [6]. This theory has been supported by functional magnetic resonance imaging (fMRI) work showing that place of articulation can be decoded from SMC activity [3]. Correia et al. (2015) [3], claimed that the ability to decode articulatory features means that articulatory (motor) representations are encoded in the SMC. However, they did not consider the physical connection between acoustics and articulation when interpreting the results. It is possible that articulatory features could be recovered from acoustic sensory representations above chance. In direct contrast, Cheung et al. (2016) [4] found that activity in the SMC during speech perception was organized by acoustic features and not articulators, suggesting that the SMC only represents sensory information. This work also claimed that there was minimal overlap in SMC activity during speech production and perception [4]. In comparison to the previous study [3], this study used electrocorticogram (ECoG) recordings to measure activity, a method which involves placing an array of sensors directly on to the cortex. This physically limits the scope of the results to only the regions with sensors, whereas fMRI can map activity across the entire cortex.

Even with the physical limitation of ECoG recordings, the results of Cheung et al. (2016) [4], are compelling, but both this study and the Correia et al. (2015) [3] study have a potential confound in their experimental design. Both studies rely on controlled speech stimuli consisting of short consonant-vowel syllable repetitions. While this stimulus format allowed them to directly compare articulatory features with simple machine learning models, it may also have deterministically linked acoustics and articulation in a way not seen in natural speech [7]. In order to make stronger claims regarding the nature of underlying representation in the SMC, we need natural stimuli-based analyses to maximally control for the correlation between acoustics and articulation.

Natural stimuli necessitates more sophisticated analysis methods due to the complexity of speech. Recent studies examining fMRI activity during speech perception in the audi-

tory cortex have leveraged representations from self-supervised (SSL) transformer speech models to predict fMRI response [8]. The representations were shown to better predictors of fMRI response than classical acoustic features, phonemic features, and supervised model features [8]. SSL speech representations also capture a variety of information about the speech signal and have been used to directly estimate articulatory movement measurements with a linear mapping from the SSL feature space [9]. As a result, SSL models may have potential for modeling acoustics and articulation separately which would allow direct comparison of the correlation between brain activity in the SMC and acoustics or articulation during both speech production and perception.

In this work, we aim to explore whether SSL representations can effectively model acoustics and articulation as separate components of a basis feature space. We focus on WavLM large, a self-supervised speech model that has been used to extract electromagnetic articulography (EMA) through a linear transformation [10, 9]. These estimated EMA features are purely articulatory, representing the movement of major articulators across time. Because of the linear relationship between estimated EMA features and WavLM representations, a feature set orthogonal to the EMA features can be represented by the residuals of a least-squares regression mapping EMA features back to WavLM features. The orthogonal residuals are theoretically the 'non-articulatory' or 'acoustic' component of the WavLM feature space. To examine the nature of the residuals as compared to EMA features, we probe each feature set for acoustic, articulatory, and semantic information.

2. Methods

We investigated WavLM's ability to model articulation and acoustics as separate components of the WavLM feature space using three primary feature sets and two dimension-matched feature sets, all derived from WavLM. These feature sets were extracted using the Speech Articulatory Coding (SPARC) model to generate articulatory movement measurements (estimated electromagnetic articulography features, EMA [9]), a least-squares regression to map articulatory movement features back into the WavLM feature space, and principal component analysis to reduce dimensions of high-rank features. The feature sets were then linearly probed using ridge regression and logistic regression. Higher prediction performance for a probe indicated that a feature set captured a higher proportion of that information type.

2.1. Data

We used speech recordings from a public fMRI dataset of scan data and paired stimulus recordings for a passive listening task [11]. The audio consisted of 27 natural narrative story recordings (10-15 minutes each, $\sim$6 hours total) clipped from *The Moth* podcast [11]. Each story was an autobiographical story without a script and was spoken by a single speaker. Time-aligned transcriptions of words and phones were available for all stories. The recordings were downsampled to 16 kHz and converted to single-channel audio. Time-aligned transcriptions were cleaned of invalid words and phones.

2.2. Feature sets

The primary feature sets were: (1) WavLM-Large layer 9 representations (WavLM, 1024-dim), (2) estimated electromagnetic articulography features (EMA, 13-dim [9]), and (3) a feature set of residuals from a least-squares regression mapping EMA features to WavLM feature (residuals, 1024-dim). We also generated rank-matched residuals (pca-residuals, 13-dim) and dimension-matched EMA features (EMA-Wav, 1024-dim).

2.2.1. WavLM Large

WavLM is a self-supervised (SSL) speech transformer model that jointly learns de-noising along with masked speech prediction to improve performance on downstream speech tasks outside of automatic speech recognition (ASR) [10]. Features from WavLM and other SSL speech models have been shown to be better predictors of the fMRI blood oxygen level-dependent (BOLD) signal, a proxy for neural activity, than acoustic baselines, phonemic features, and supervised model representations during speech perception [8, 12].

We extracted features from layer 9 of a pretrained WavLM Large model (316.62M parameters) implemented with HuggingFace [13]. The pretrained model was English-only and output 1024-dimensional features [10]. All other feature sets were derived from WavLM, making the WavLM feature space the basis feature space. Probe performance for WavLM should therefore serve as an upper-bound for probe performance. We selected features for layer 9 for consistency with other WavLM-derived features.

2.2.2. Electromagnetic articulography (EMA) features

Electromagnetic articulography (EMA) features describe the kinematics of articulation through XY measurements of six major articulators (upper lip, lower lip, lower incisor, tongue tip, tongue blade, and tongue dorsum) during speech [9]. Ground truth measurements are captured with sensors placed on the midsagittal plane and recorded with a sampling rate of 200 Hz. Voicing features such as pitch and loudness are often appended to the measurements to provide voicing information.

We did not have access to ground truth EMA features due to the pre-recorded nature of the audio data. Instead, we used the recently released Speech Articulatory Coding (SPARC) model [9], designed for projecting speech into an EMA and speaker feature space and performing speech synthesis and editing, to estimate EMA features using a linear mapping from layer 9 of WavLM Large. SPARC also returns pitch and loudness values to model voicing, resulting in a 14-dimensional feature. We excluded the loudness measurement as it is purely acoustic. The final EMA feature set had 13 dimensions.

2.2.3. Least-squares regression

We used a least-squares regression mapping EMA features to WavLM features to generate a set of residual features (1024-dim) orthogonal to EMA features within the WavLM feature space. Orthogonality is ensured due to the linear relationship between estimated EMA features and WavLM features.

We additionally used the predicted WavLM features as a higher-dimensional EMA feature set (EMA-Wav, 1024-dim). These features were not rank matched with the residuals as there are still only 13 independent dimensions when projecting into the WavLM feature space. We chose to examine this feature set because of a key difference in EMA and WavLM features properties: WavLM features can be pooled across time and still retain critical information for the entire time duration within the mean pooled feature. This is not true for EMA features as they are XY measurements. The downstream task of predicting fMRI responses relies on downsampling with mean pooling to

align features with scans, meaning that we need to demodulate the EMA features so that they have the same pooling property as WavLM features. The least-square regression may have some ability to demodulate the EMA features.

The regression was trained using EMA and WavLM features from all 27 stimuli stories without any normalization. We did not include an intercept. The R-squared (R^2) value on the training data was ~ 0.5136, meaning that the EMA features explained about 51.4% of the variance in wavLM features.

Because EMA features are low rank (13) but the residuals of the least squares regression are high rank (1024), we generated a rank-matched residual feature set using principal component analysis (PCA) to reduce the dimensionality of the residuals to 13. The PCA method was fit with the least-squares residual features. We transformed the data with a standard scaler (mean removal, unit variance).

2.3. Probes

The linear probes were grouped into four categories: acoustics, articulatory, mixed acoustic and articulatory, and semantic. We extracted two acoustic probe features: FBANK spectral features and ComParE 2016 features. The FBANK features were extracted using torchaudio's kaldi compliance module with 40 mel bins (40-dim) [14]. The ComParE 2016 low-level descriptor features (65-dim) were extracted using openSMILE [15]. These features included 65 acoustic measures such as loudness, zero-crossing rate, MFCCs (14), spectral energy, harmonicity, F0, voice probability, jitter, shimmer, and more [15]. The articulatory probe was a set of categorical articulatory features extracted from the time-aligned phone transcriptions. These categorical articulatory features represent phones as a set of 22 binary features including places of articulation, manners of articulation, and voicing. We used word identity and phone identity as mixed probes since both articulation and acoustics can provide information about a word or phone. We prepared word and phone identity probes by generating a vocabulary of unique words (>1000) or phones (59) and converting the transcriptions into target classes. Word embeddings (985-dim) were used as a semantic probe. The word embeddings were extracted using an English1000 semantic model, trained on 1000 of the most common English words, that projected words into a 985-dim embedding space [16]. In this embedding space, semantically similar words are generally closer together.

2.4. Feature extraction pipeline

Any features or probes extracted directly from the audio signal were extracted using a sliding window with a window size of 8.1 seconds and a shift of 0.1 seconds. These parameters were selected based on the expected downstream use of these features for fitting encoding models. Previous work showed that providing context to WavLM generated features with better encoding model performance [12]. In practice, providing context means that we conceptualize windows as a context portion and chunk portion, with the window shift always equivalent to the chunk size. The context portion of a window is the initial part of the window that is provided to the model during feature generation but ignored in the output (i.e., we do not save the features). We set context size to 8s and chunk size to 0.1s.

Following WavLM parameters, we subdivided the windows into frames of 25ms with a stride of 20ms and extracted a single feature for each frame across all windows. We then downsampled the features by saving only the final feature from each window. For word and phone-based probes, we required an additional downsampling step to align the feature sets with the probes. All features that fell within the start and end times of a word or phone were mean pooled.

We extracted features for full windows and full frames to maintain consistency in the number of total features extracted for each feature set or probe. We used a NVIDIA GeForce RTX 2080 Ti GPU for PyTorch-based feature extraction.

2.5. Linear probing

We examined the nature of the main feature sets by predicting a variety of probe features with linear models. We randomly divided the 27 stimulus stories into 5 train/test splits (80/20) and fit models for each split. The feature sets were scaled using a standard scaler (mean removal, unit variance) fit on the training data. Performance was interpreted as expressing the amount of information captured by a feature set.

We predicted FBANK, openSMILE ComParE 2016, and word embedding probes with ridge regression [8]. The ridge regression models were fit with repeated k-fold cross validation (5 splits, 3 repeats) and ten alphas ranging (0.00001-100, log spaced). We did not include an intercept. Performance was measured as the average Pearson correlation (r) across the probe feature dimension and the R^2 value. We predicted word identity, phone identity, and categorical articulatory feature probes using logistic regression [8]. Each categorical articulatory feature was treated as a separate binary classification task. We measured performance with balanced accuracy due to class imbalance across the classification probes.

3. Results

WavLM features consistently captured the most information regardless of probe ($p < 0.001$, Figure 1), highlighting WavLM's ability to capture a variety of speech-related information. This result aligned with our expectations for the basis feature set.

Across all acoustic probes, the residuals and pca-residuals had significantly higher correlation compared to the corresponding EMA-based feature ($p < 0.001$, Figure 1), demonstrating that the residuals contain more acoustic information than EMA-based features. The residuals had significantly higher correlation with ComParE 2016 features than pca-residuals ($p < 0.001$). This indicates that some acoustic information is lost with the PCA dimensionality reduction. There was no significant difference in performance between the EMA-based features.

A contradictory trend appeared for the other probe categories. For articulatory, mixed articulatory and acoustic, and semantic probes, the low-rank pca-residuals (13-dim) had significantly lower performance than the EMA features ($p < 0.001$) while the high-rank residuals (1024-dim) exhibited significantly higher performance ($p < 0.001$), second only to the WavLM features. This means that while EMA-based features captured more semantic and articulatory information than pca-residuals, they did not capture more than the high-rank residuals. Again, there was no significant difference in performance for EMA-based features.

For ridge regression-based probes, WavLM features explained significantly more variance ($p < 0.05$, Table 1) than other features, followed by the residuals. Both residual-based features had higher R^2 values than EMA-based features for FBANK features, but pca-residuals explained approximately the same amount of variance as EMA-based features for the ComParE 2016 probe and less variance than EMA-based fea-

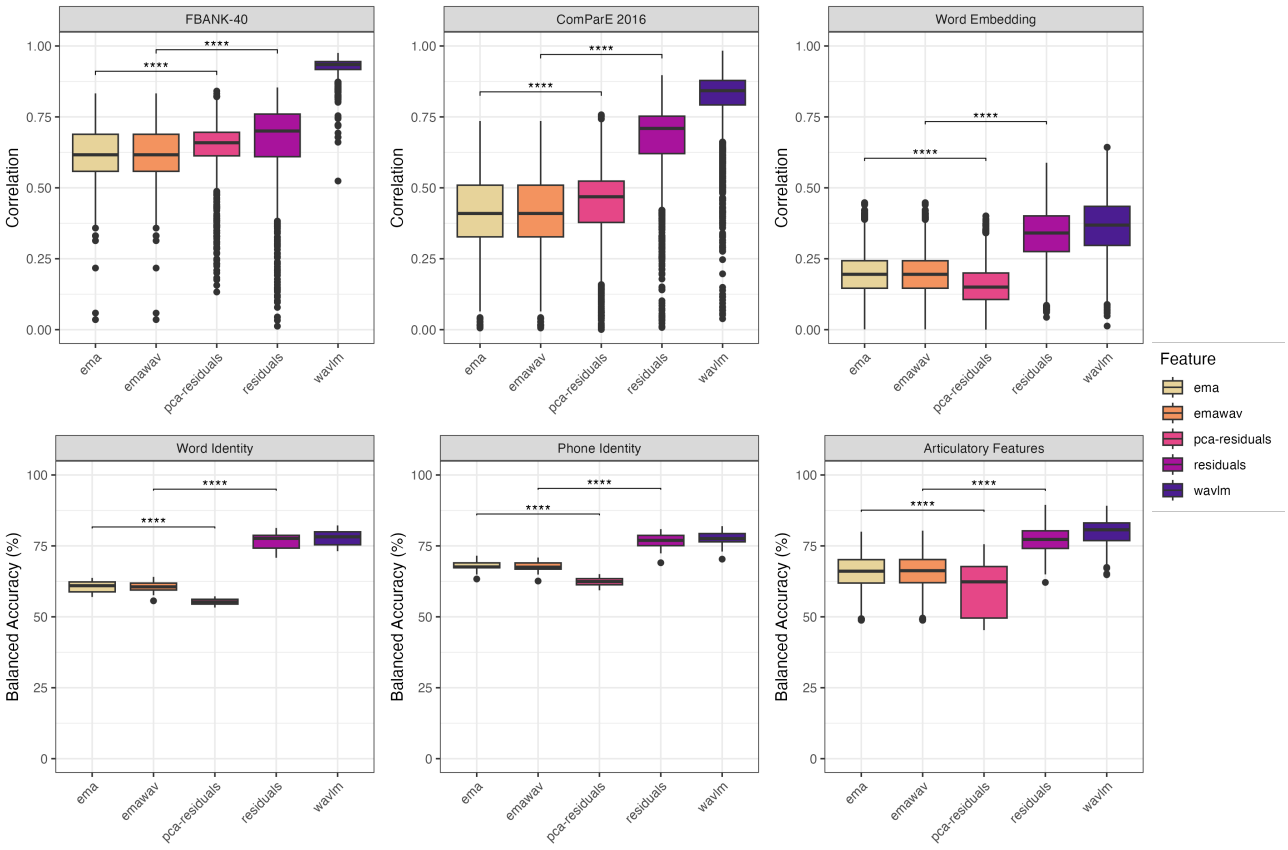

Figure 1: *Probe performance evaluated across feature sets. Ridge regression models evaluated with average correlation across the probe feature dimension. Logistic regression evaluated as balanced accuracy. Significance included for dimension-matched EMA and residual features (p<0.001).*

tures for the word embedding probe. The differences between EMA-based and residual-based features were not significant. As seen with the correlations, the difference between residuals and pca-residuals was slightly significant for ComParE 2016 features, with residuals explaining more variance. The most notable trend was that the R^2 values for word embeddings were significantly lower across all feature sets (Average R^2=0.319, $p < 0.001$, Table 1). This indicates that all feature sets did not capture much semantic information.

For classification-based probes, WavLM had significantly higher performance ($p < 0.001$) than all other feature sets, while pca-residuals had significantly lower performance ($p < 0.001$) (Table 2). The original residuals were significantly higher than both pca-residuals and EMA-based features ($p < 0.001$). Phone identity had overall higher performance than word identity ($p < 0.001$) across all feature sets and had comparable performance to categorical articulatory features, though with predictably less variance ($\sigma = 0.061$ vs. $\sigma = 0.089$). Performance did not vary significantly across classification tasks for the high-rank residuals and WavLM.

4. Discussion

We found evidence that it is possible to disentangle articulation and acoustics in the WavLM feature space using EMA features and orthogonal residual features. EMA-based features (13-dim EMA, 1024-dim EMA-Wav) captured significantly less acous-

Table 1: *Regression: R^2 values averaged over splits. For global averages across features and probes, * indicates a significance of p<0.001. Features: W=WavLM, E=EMA, EW=EMA-Wav, PR=pca-residuals, R=residuals, Probes: F40=FBANK, C16=ComParE 2016, Word=word embedding.*

	W	E	EW	PR	R	Avg
F40	0.88	0.64	0.64	0.68	0.68	0.70
C16	0.74	0.53	0.53	0.53	0.62	0.59
Word	0.36	0.30	0.30	0.29	0.35	**0.32***
Avg	**0.72***	0.51	0.51	0.53	0.56	

Table 2: *Classification: Balanced accuracy averaged over splits. For global averages across features and probes, * indicates a significance of p<0.001. Features: W=WavLM, E=EMA, EW=EMA-Wav, PR=pca-residuals, R=residuals, Probes: WI=word identity, PI=phone identity, CA=categorical articulatory features.*

	W	E	EW	PR	R	Avg
WI	0.70	0.58	0.56	0.52	0.70	**0.62***
PI	0.78	0.67	0.67	0.62	0.76	0.70
CA	0.80	0.66	0.66	0.62	0.77	0.70
Avg	**0.80***	0.65	0.65	*0.62**	0.77	

tic information than residual based features (1024-dim residuals, 13-dim pca-residuals) but more articulatory information than pca-residuals. Additionally, the feature sets did not capture semantic information, indicating they are more purely articulatory or acoustic in nature.

For articulatory probes, the residuals broke the trend, exhibiting an unexpected ability to capture articulatory information. They outperformed EMA-based features for all articulatory probes. This is likely explained by two factors: (1) the increased information capacity due to the rank of the residual features, and (2) the negative influence of mean pooling on EMA-based features. Theoretically, EMA-based features should have higher balanced accuracy than the residuals when predicting articulatory and mixed articulatory-acoustic probes as these features can be used to synthesize speech with high fidelity [9]. However, both EMA and EMA-Wav features had a lower rank than the residuals. Given that the low-rank pca-residuals followed the expected trend, the higher rank of the residuals appears to be the main reason they outperformed. Additionally, the downsampling required to align the feature sets with the probes required mean pooling, which is incompatible with the 13-dimensional EMA features. We had hypothesized that the least squares regression would demodulate the EMA features, but this was not seen in our results. The EMA-Wav features had the same low performance as the original EMA features.

This has downstream implications for using EMA features to fit fMRI encoding models, as further downsampling is required to align features with the repetition time (TR), or the time between consecutive images, of an fMRI scan. More work needs to be done to demodulate and increase the rank of EMA features prior to fitting encoding models. Some potential avenues include more sophisticated neural-nets for mapping EMA features to WavLM features or extracting articulatory trajectory features representing longer time-scale patterns of movement for the articulators rather than XY measurements [17].

The initial downsampling procedure during feature extraction (prior to aligned with probe features) may also have influenced the results. We only kept one out of five feature frames which significantly reduced the amount of training data available for fitting the linear models. This would have a larger effect on word and phone classification tasks where mean pooling over the word/phone duration is required. We may have excluded frames with important information about word/phone identity. Future work may consider evaluating linear probes without extra downsampling.

We were also limited by choosing only layer 9 of the WavLM model. This limitation came from how estimated EMA features are extracted with the SPARC framework, but there may be merit to exploring how much variance EMA features explain for different layers and how probing performance varies when extracting residuals from different layers of WavLM.

5. Conclusion

We explored whether the WavLM feature space can be used to separately model articulation and acoustics through linear probing. We found that using estimated electromagnetic articulography (EMA) features and simple regression and principal component analysis to extract a rank-matched orthogonal feature set shows potential for disentangling the articulatory and acoustic components of WavLM. However, the dimensionality of EMA features, along with other properties the feature set, may limit the downstream applicability of the derived features for studying brain activity.

6. Code availability

Feature extraction and linear probing code is available at https://github.com/dwiepert/audio_features. Audio preprocessing code is available at https://github.com/dwiepert/audio_preprocessing. Both repositories require an additional private support code package with utilities for accessing data buckets.

7. References

[1] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature reviews. Neuroscience*, vol. 8, pp. 393–402, 06 2007.

[2] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation," *Nature*, vol. 495, no. 7441, p. 327–332, Feb. 2013. [Online]. Available: http://dx.doi.org/10.1038/nature11911

[3] J. M. Correia, B. M. Jansma, and M. Bonte, "Decoding articulatory features from fmri responses in dorsal speech regions," *The Journal of Neuroscience*, vol. 35, no. 45, p. 15015–15025, Nov. 2015. [Online]. Available: http://dx.doi.org/10.1523/JNEUROSCI.0977-15.2015

[4] C. Cheung, L. S. Hamilton, K. Johnson, and E. F. Chang, "The auditory representation of speech sounds in human motor cortex," *eLife*, vol. 5, Mar. 2016. [Online]. Available: http://dx.doi.org/10.7554/eLife.12577

[5] S. E. Blumstein and S. R. Baum, "Chapter 55 - neurobiology of speech production: Perspective from neuropsychology and neurolinguistics," in *Neurobiology of Language*, G. Hickok and S. L. Small, Eds. San Diego: Academic Press, 2016, pp. 689–699. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124077942000559

[6] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[7] L. Hamilton and A. Huth, "The revolution will not be controlled: natural stimuli in speech neuroscience," *Language, Cognition and Neuroscience*, vol. 35, pp. 1–10, 07 2018.

[8] A. R. Vaidya, S. Jain, and A. Huth, "Self-supervised models of audio effectively explain human cortical responses to speech," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 21 927–21 944. [Online]. Available: https://proceedings.mlr.press/v162/vaidya22a.html

[9] C. J. Cho, P. Wu, T. S. Prabhune, D. Agarwal, and G. K. Anumanchipalli, "Coding speech through vocal tract kinematics," 2024. [Online]. Available: https://arxiv.org/abs/2406.12998

[10] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[11] A. LeBel, L. Wagner, S. Jain, A. Adhikari-Desai, B. Gupta, A. Morgenthal, J. Tang, L. Xu, and A. G. Huth, "A natural language fmri dataset for voxelwise encoding models," *Scientific Data*, vol. 10, no. 1, Aug. 2023. [Online]. Available: http://dx.doi.org/10.1038/s41597-023-02437-z

[12] A. R. Vaidya, "Speech encoding models: Wavlm & context," Online, March 2022. [Online]. Available: https://github.com/HuthLab/lab-meeting-slides/blob/master/2022-Spring/20220309_wavlm-context.pdf

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[14] J. Hwang, M. Hira, C. Chen, X. Zhang, Z. Ni, G. Sun, P. Ma, R. Huang, V. Pratap, Y. Zhang, A. Kumar, C.-Y. Yu, C. Zhu, C. Liu, J. Kahn, M. Ravanelli, P. Sun, S. Watanabe, Y. Shi, Y. Tao, R. Scheibler, S. Cornell, S. Kim, and S. Petridis, "Torchaudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for pytorch," 2023.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462.

[16] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, p. 453–458, Apr. 2016. [Online]. Available: http://dx.doi.org/10.1038/nature17637

[17] J. Chartier, G. K. Anumanchipalli, K. Johnson, and E. F. Chang, "Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex," *Neuron*, vol. 98, no. 5, pp. 1042–1054.e4, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0896627318303398