EVIDENCE OF VOCAL TRACT ARTICULATION IN SELF-SUPERVISED LEARNING OF SPEECH

Cheol Jun Cho[1]   *Peter Wu*[1]   *Abdelrahman Mohamed*[2]   *Gopala K. Anumanchipalli*[1]

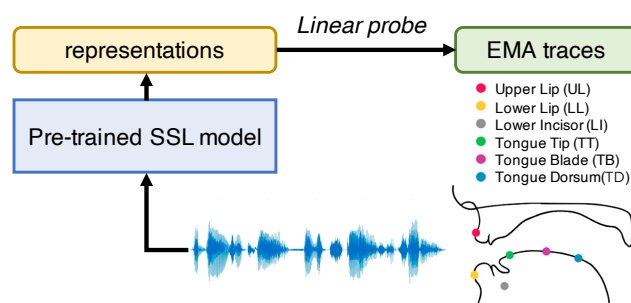[1]UC Berkeley, EECS, CA
[2]Meta AI

ABSTRACT

Recent self-supervised learning (SSL) models have proven to learn rich representations of speech, which can readily be utilized by diverse downstream tasks. To understand such utilities, various analyses have been done for speech SSL models to reveal which and how information is encoded in the learned representations. Although the scope of previous analyses is extensive in acoustic, phonetic, and semantic perspectives, the physical grounding by speech production has not yet received full attention. To bridge this gap, we conduct a comprehensive analysis to link speech representations to articulatory trajectories measured by electromagnetic articulography (EMA). Our analysis is based on a linear probing approach where we measure articulatory score as an average correlation of linear mapping to EMA. We analyze a set of SSL models selected from the leaderboard of the SUPERB benchmark [1] and perform further layer-wise analyses on two most successful models, Wav2Vec 2.0 [2] and HuBERT [3]. Surprisingly, representations from the recent speech SSL models are highly correlated with EMA traces (best: $r = 0.81$), and only 5 minutes are sufficient to train a linear model with high performance ($r = 0.77$). Our findings suggest that SSL models learn to align closely with continuous articulations, and provide a novel insight into speech SSL.

Index Terms— Speech, Self-supervised learning, Electromagnetic articulography (EMA), Speech representation, Probing analysis, Acoustic-to-articulatory inversion

1. INTRODUCTION

Self-supervised learning (SSL) has been suggested as a pre-training method to learn representations without requiring a huge amount of labeled data. Recently proposed SSL models for speech provide rich representations which can be readily utilized for a broad range of spoken language tasks. When fine-tuned to downstream tasks, the SSL-based models are able to surpass supervised-only models [4, 1]. Understanding how SSL models work is crucial to explain such success and to improve speech SSL. Previous studies have revealed acoustic, phonetic, and semantic representations encoded in representations of speech SSL, using diverse analytic tools in-

Fig. 1. General framework of our analysis approach: linear probing of representations from pre-trained SSL models on EMA

cluding (non-) linear probing, mutual information, or canonical correlation analysis [5, 3, 6, 7, 8]. Although those analyses have provided insights of information process in speech SSL models, the models still largely remain black box. Since the advent of data-driven deep learning strategies for spoken language engineering, they have grown increasingly disconnected from human mechanisms and insights from speech production.

To bridge this gap, we answer the question of how well the SSL representations are aligned with principles of speech production. We conduct a linear probing analysis of SSL representations against electromagnetic articulography (EMA) (Fig. 1). EMA measures real-time continuous displacements of 6 articulators (Fig. 1), which are timely synced with the produced speech [9, 10]. As EMA tracks the actual physical dynamics of vocal tract, the modality is suitable for investigating physical grounding of speech representation. Moreover, a large portion of speech features are naturally subsumed by articulatory trajectories and a full speech can be reconstructed from EMA [11]. Underlying neurobiological process of speech production can also be explained by vocal tract articulation [12]. Therefore, we target EMA traces as principled and grounded reference for probing representations of SSL models. We firstly introduce articulatory score as an average correlation of linear prediction and EMA. Then, we conduct a comprehensive analysis by controlling model space, data size, and layers for probing. Our work bridges

the gap between the current leading representation learning schemes and classical speech production research.

Our major findings are:

- SSL representations are highly correlated with articulations: a simple linear regression is able to yield high correlation above 0.8.
- Ranking by articulatory scores is coherent with the SU-PERB leaderboard [1], suggesting EMA prediction as a valid benchmark task.
- Articulatory encoding of SSL representations is extremely robust that only 5 minutes of EMA data are sufficient to obtain high performance.
- Layer-wise analysis reveals a shared pattern in Transformer based models except for last few layers, which is consistent with previous phoneme encoding analyses [2, 3, 6]. This trend is intact after fine-tuning for automatic speech recognition (ASR) task.
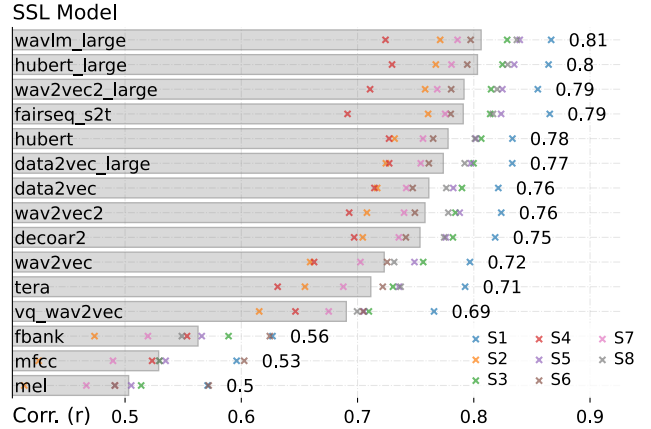
2. PRIOR WORKS

The speech SSL benchmarking provides a general sense of representational power of SSL models, based on the downstream task performance [1]. Several studies have probed the learned representations themselves [6, 8, 5, 3, 7]. [5] and [3] analyzes layer-wise phoneme encoding of their developed unsupervised models. [6] characterizes across layer patterns of acoustic, phonetic, and semantic encoding in speech SSL model. [8] conducted non-linear probing analyses using extensive and fine-grained labels of speech features. Articulations are also included in the scope of [8, 7], but those are hand labeled pronunciations which are discrete and lack context dependent information such as coarticulation. Up to our knowledge, none of the prior works directly deals with continuous articulations represented as EMA. Unlike the previous probing analyses focus on a limited set of models, this study covers a broad range of speech models from hand-engineered features to the current state-of-the-art SSL models.

3. METHOD

3.1. Articulatory Datasets

The articulatory dataset is composed of two EMA datasets, MNGU0 [10] and MOCHA-TIMIT [9]. Both datasets provide 12 EMA traces (X, Y midsagittal coordinates for each of 6 articulators) of spoken utterances. We use 1,189 utterances by a single speaker (S1) from MNGU0 and 470 utterances by each of 7 speakers (S2-8) from MOCHA-TIMIT. The total duration is up to 75 minutes for MNGU0 and 28 minutes in average for MOCHA-TIMIT. 100 utterances of MNGU0 speaker and 50 utterances of each MOCHA-TIMIT speaker were held out as test sets. The data are normalized by channel and down-sampled to 50 Hz to match the sampling rate of SSL models.

Fig. 2. Articulatory score for each model. Each individual speaker's scores are marked with color. The annotated numbers indicate the scores averaged across speakers.
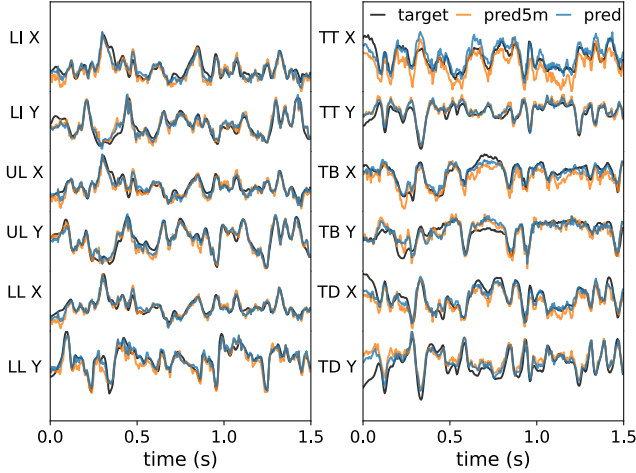
3.2. SSL models

We evaluate a pool of self-supervised speech representations by retrieving official models using S3PRL [1] and Hugging-face.[1] To give a fair comparison, we attempted to constrain the scope of analysis to the models pre-trained on LibriSpeech 960 hours corpus [13] or LibriLight 60K hours corpus [14], and we exclude extensions or variants. Although WavLM [15] is an extension of HuBERT [3], the large version of the model is included as a reference of the best score we can get from SSL models. As a supervised counterpart, the encoder of Fairseq S2T [16] is added, which was trained for ASR using LibriSpeech. Finally, some handcrafted acoustic features are included as the baselines. The final list includes: (*contrastive model*) Wav2Vec [17], VQ-Wav2Vec [18] and Wav2Vec 2.0 [2], (*predictive model*) HuBERT, Data2Vec [19] and WavLM, (*generative model*) TERA [20] and DeCoAR 2.0 [21], (*supervised model*) Fairseq S2T, and (*hand crafted*) filter bank, mel-spectrogram and MFCC (Fig. 2). If applicable, both base and large versions are included.

The large models were trained on LibriLight 60K hours except for WavLM which was trained on 57% more data from other corpus in addition to LibriLight. All the other models used LibriSpeech for training. By feeding speech audio of the EMA dataset, we extract representations from each layer of the models. For some models, features are down-/up-sampled to 50 Hz, so all representations are paired with synchronized EMA targets with the consistent sampling rate.

3.3. Articulatory score

Articulatory score is defined as an average correlation of linear probing. For each representational space from Sec. 3.2, a linear model is trained to predict each of 12 EMA traces,

---

Fig. 3. Example from test set. Separate models are trained on full training set (*pred*) and 5 minute subset (*pred5m*).

Fig. 4. Articulatory scores by different amount of training data for three speakers: S1 (*left*), S5 (*top right*) and S8 (*bottom right*)

frame by frame.[2] The correlation is measured by Pearson's correlation coefficient (r) on held-out test set. Finally, articulatory score is obtained by averaging over 12 EMA channels and 8 speakers, if no articulator or speaker is specified.

We train separate models for each speaker since the individual variability in EMA is not only affected by speaker identity but also, or more, affected by the sensor configuration.
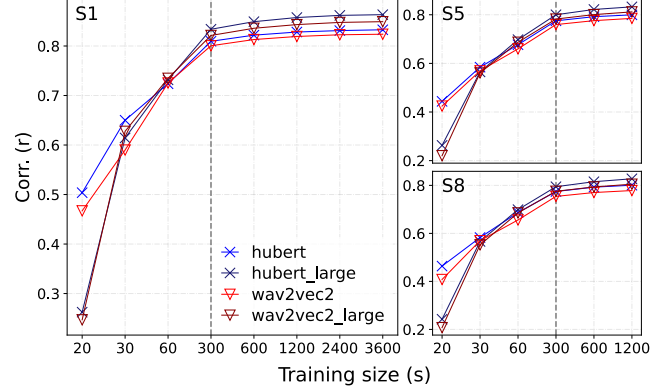
4. RESULTS

4.1. Comparison across SSL models

To compare representational power of SSL models for articulations, we assess articulatory scores of SSL models. Highest scores are chosen among the multiple layers, and reported on Fig. 2. The overall pattern is consistent with SUPERB [1], where Transformer based models trained with contrastive or predictive objectives (e.g., Wav2Vec 2.0, HuBERT, WavLM) are generally better than the other. The ranks of the models mostly match those from the SUPERB leaderboard with high rank correlation (Kendall tau=0.82, Spearman=0.93). One unexpected deviation is Data2Vec Large, which shows lower score than the score of HuBERT Base.

Overall scores are surprisingly high that even with base model such as HuBERT Base, we can get 0.78 correlation and for the best model, WavLM Large, the average correlation reaches 0.81, which are far beyond the handcrafted baselines (Fig. 2). Increasing model capacity can help build better articulatory representations, as the large models produce higher score than base versions. The supervised model, Fairseq S2T, has higher score than SSL models with comparable capacity,

suggesting that ASR task is relevant for learning articulatory representations.

There is also a notable tendency across speakers that some speakers (e.g., S1, S5 and S8) show consistently high performance but some show low performance (e.g., S4, S2). This is due to noise ceiling of EMA data, where some speakers' data underwent unstable measurement.

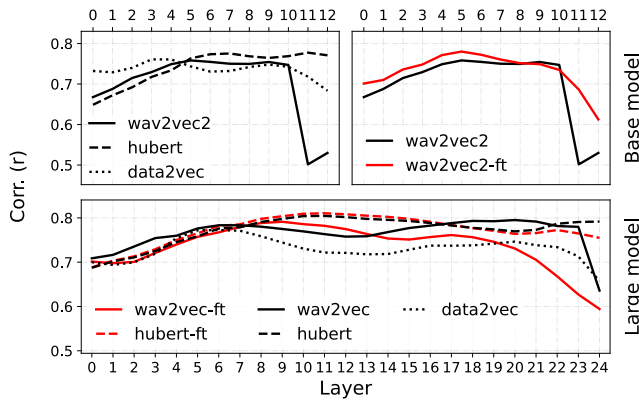## 4.2. Effects of training data size for EMA probing

To further test how robustly representations are aligned with articulations, we control the size of training data for training linear models. The experiments were done using Wav2Vec 2.0 and HuBERT on three best speakers, S1, S5 and S8. The training size is one of [20s, 30s, 1m, 5m, 10m, 20m] and additional [40m, 60m] for S1. As shown in Fig. 4, the scores increase dramatically until 5 minutes of training data and then start to saturate in all cases. This suggests that only 5 minutes of data are sufficient to build a high-performing linear mapping model. To verify this, we conducted the experiment with all 8 speakers using HuBERT Large layer-11 which shows the highest score among 24 layers (Fig. 5 *bottom*). The performance of different size of training data is reported on Table 1. Strikingly, for all speakers, the extremely low-resource setting (5 min) shows sufficiently high-performance. This is also true when we qualitatively check the prediction (Fig. 3). This finding suggests that representations from the state-of-the-art SSL models efficiently encode a well generalizable representation of articulation.

4.3. Encoding patterns across layer

To investigate how information evolves through the layers, we evaluate articulatory score for each layer (Fig. 5) from Wav2Vec 2.0, HuBERT and Data2Vec. These models were selected since they share model architecture with a convolu-

---

[2]By solving ordinary least squares regression without regularization.

Speaker	Training size		
	5 min	10 min	All
S1	0.83	0.85	0.87
S2	0.74	0.76	0.77
S3	0.79	0.81	0.82
S4	0.67	0.70	0.73
S5	0.80	0.82	0.83
S6	0.76	0.78	0.79
S7	0.74	0.76	0.78
S8	0.80	0.82	0.83
Average	0.77	0.79	0.80

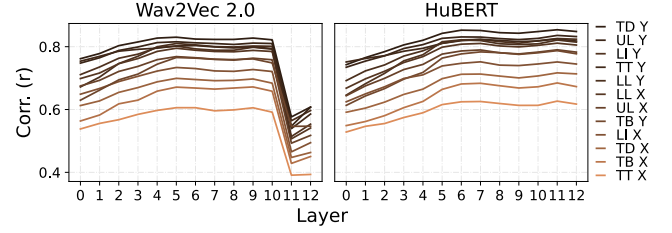**Table 1**. Articulatory scores using HuBERT Large layer-11 trained on 5m, 10m and all training data.

Fig. 5. Layer-wise articulatory scores of base models (*top left*), vs. fine-tuned model of Wav2Vec 2.0 (*top right*) and any applicable large models (*bottom*). Layer-0 is the output of CNN encoder. The red line, *model*-ft, indicates the fine-tuned version of *model*.

tional waveform encoder followed by Transformer encoder, but these are trained with different learning objectives.

For all models, there are two local peaks in scores through the layers: the first one around the middle and the second one a few layers before the last layer (Fig. 5 *top left*). This pattern is also found in large models and remained intact even after fine-tuning models for ASR on LibriSpeech 960 hours (Fig. 5 *bottom and top right plots*). However, scores differ dramatically by models near the last layers where the effects of learning objective are strong. This is also supported by the observation that the later layers show larger changes than the earlier layers when fine-tuned. There are slight gains in performance after fine-tuning for all three models: Wav2Vec 2.0 Base 0.76 → 0.78, Large 0.79 → 0.80 and HuBERT Large 0.80 → 0.81. Combined with the finding from Sec. 4.1, training for ASR task is contributing to learning articulatory representations.

Compared to previous probing analyses, the layer-wise

Fig. 6. Layer-wise scores for each articulator from Wav2Vec 2.0 (*left*) and HuBERT (*right*) Each articulator is colored by average correlation.

patterns in Wav2Vec 2.0 and HuBERT are similar to those of the phoneme encoding analyses by [2, 3, 6], which is natural since phonemes can be characterized by articulations.

Layer-wise scores for each articulator are plotted on Fig. 6. The same across layer trend is consistently found in individual articulator for both models. There exists a variability among EMA channels: Y-coordinates of tongue dorsum, upper lip, lower incisor, and tongue tip are well-decoded while X-coordinates of tongue tip, tongue blade, tongue dorsum, and lower incisor show relatively low scores. The articulators with low scores are indeed known to be less salient than other articulators in vocal tract dynamics.

4.4. Speaker independent articulatory representations

Although it is challenging to study speaker independent feature of EMA due to variable sensor configuration, we repeated experiments in two speaker independent schemes: 1) sharing models across speakers with the same train-test split, and 2) leave-one-speaker out cross-validation. In both settings, we were able to replicate the patterns reported in Sec. 4.1 and Sec. 4.3, with some degradation in scores. Since speaker identity gets marginalized in deep layers of SSL models [22], these consistent findings suggest that there exist speaker independent articulatory representations in SSL representations.

5. CONCLUSION

We probe recent SSL models with EMA traces and find that the learned representations have highly correlated and well-generalizable articulatory representations. Our findings suggest that the recent SSL models are well-grounded by the physical process of speech production. This motivates an interesting future direction of developing a more efficient model by directly grounding the model with articulations.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 2020-Decem, pp. 1–19, 2020.

[3] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[4] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, et al., "Self-supervised speech representation learning: A review," *arXiv preprint arXiv:2205.10643*, 2022.

[5] Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli, "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27826–27839, 2021.

[6] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 914–921.

[7] Danni Ma, Neville Ryant, and Mark Liberman, "Probing acoustic representations for phonetic properties," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 311–315.

[8] Jui Shah, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[9] Alan Wrench, "Mocha: multichannel articulatory database," http://www.cstr.ed.ac.uk/research/project/artic/mocha.html, 1999.

[10] Korin Richmond, Phil Hoole, and Simon King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[11] Peter Wu, Shinji Watanabe, Louis Goldstein, Alan W Black, and Gopala K Anumanchipalli, "Deep speech synthesis from articulatory representations," *arXiv preprint arXiv:2209.06337*, 2022.

[12] Josh Chartier, Gopala K Anumanchipalli, Keith Johnson, and Edward F Chang, "Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex," *Neuron*, vol. 98, no. 5, pp. 1042–1054, 2018.

[13] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[14] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, et al., "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[15] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[16] Changhan Wang, Yun Tang, Xutai Ma, et al., "fairseq s2t: Fast speech-to-text modeling with fairseq," in *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*, 2020.

[17] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Proc. Interspeech 2019*, pp. 3465–3469, 2019.

[18] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2019.

[19] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, et al., "Data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint arXiv:2202.03555*, 2022.

[20] Andy T Liu, Shang-Wen Li, and Hung-yi Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[21] Shaoshi Ling and Yuzong Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *arXiv preprint arXiv:2012.06659*, 2020.

[22] Sanyuan Chen, Yu Wu, Chengyi Wang, et al., "Why does self-supervised learning for speech recognition benefit speaker recognition?," *arXiv preprint arXiv:2204.12765*, 2022.