Exploring Self-Supervised Speech Representations for Cross-lingual Acoustic-to-Articulatory Inversion

Yun Hao, Reihaneh Amooie, Wietse de Vries, Thomas Tienkamp, Rik van Noord, Martijn Wieling

University of Groningen, The Netherlands

{yun.hao, r.amooie, wietse.de.vries, t.b.tienkamp, r.i.k.van.noord, m.b.wieling}@rug.nl

Abstract

Acoustic-to-articulatory inversion (AAI) is the process of inferring vocal tract movements from acoustic speech signals. Despite its diverse potential applications, AAI research in languages other than English is scarce due to the challenges of collecting articulatory data. In recent years, self-supervised learning (SSL) based representations have shown great potential for addressing low-resource tasks. We utilize wav2vec 2.0 representations and English articulatory data for training AAI systems and investigates their effectiveness for a different language: Dutch. Results show that using mms-1b features can reduce the cross-lingual performance drop to less than 30%. We found that increasing model size, selecting intermediate rather than final layers, and including more pre-training data improved AAI performance. By contrast, fine-tuning on an ASR task did not. Our results therefore highlight promising prospects for implementing SSL in AAI for languages with limited articulatory data.

Index Terms: acoustic-to-articulatory inversion, speech inversion, self-supervised learning

1. Introduction

Speech inversion or acoustic-to-articulatory inversion (AAI) is the process of inferring vocal tract movements from acoustic speech signals. Recently, there has been increasing interest in the development of AAI systems because of their potential application across various speech-related tasks, such as Automatic Speech Recognition (ASR) [1, 2], speech synthesis [3], pronunciation training [4] and speech therapy [5]. Electromagnetic articulography (EMA) is a widely used technique for gathering precise articulatory data from human subjects [6]. EMA is a point tracking method, where sensors placed on target articulators (including tongue, lips, and jaw) are used to track articulatory movements in real time in 3D. Despite its efficacy, the acquisition of EMA data remains challenging due to its cost and the need for specialized technical expertise.

To facilitate research in speech production and AAI, several EMA datasets are made publicly available, such as the Haskins Production Rate Contrast (HPRC) [7], the MOCHA-TIMIT [8], the MNGU0 [9], USC-TIMIT [10] and the EMA-MAE datasets [11]. As most of the publicly available datasets are in English, research on AAI tends to likewise prioritize English. For the vast majority of other languages spoken worldwide, however, sufficient data to train an AAI system is lacking.

Given that vocal tract anatomy and orofacial muscles are language independent, and articulatory processes heavily overlap across languages, AAI systems trained with a rich-resource language could potentially be adapted to relatively similar lower-resource languages. Sivaraman and colleagues [12] compared AAI models trained on English, Dutch, or Dutch-accented English, and observed that testing with non-matching languages resulted in diminished performance. Similar results were achieved by studies comparing English and four Indian languages [13], and English and Japanese [14]. Overall, cross-linguistic transferability of AAI systems has shown less than satisfactory results. This phenomenon may be attributed, in part, to the use of Mel-frequency Cepstral Coefficients (MFCC) as the acoustic features in these studies, which may lack the capacity to represent rich and robust speech information.

In recent years, self-supervised learning (SSL) based pre-trained models of speech, such as wav2vec 2.0 [15], have demonstrated remarkable performance across various downstream tasks, showing potential for addressing the challenges posed by limited training data. SSL features have also been introduced into English AAI tasks, achieving state-of-the-art results [16, 17, 18]. To investigate the cross-lingual and cross-speaker transferability of SSL features for AAI, Cho and colleagues [19] probed articulatory representation represented by the individual layers in speech SSL models trained on different languages. Although they found that the cross-lingual transferability of the representations was lower than within language performance, cross-lingual performance was relatively high and suggested the universality of the articulatory representation in speech SSL models. They also found that representations captured by intermediate layers outperformed those by the final layer. This has also been found for different tasks, such as the task of quantifying pronunciation differences on the basis SSL representations [20]. However, as the AAI systems developed by Cho and colleagues were trained in a speaker-dependent way, further research is needed to assess the full potential of SSL representations in enhancing speaker-independent AAI performance for languages lacking sufficient articulatory data.

The goal of this study is therefore to investigate the cross-lingual generalizability of SSL representations for AAI. Specifically, we examine the performance of speaker-independent AAI systems with English HPRC corpus and Dutch EMA data. By comparing the performance of AAI systems trained with different wav2vec 2.0 representations, we aim to answer the following research questions:

- **RQ1**: How well do AAI models trained using SSL representations generalize to an unseen language?
- **RQ2**: How does AAI performance relate to features extracted from different layers, models differing in model size, and pre-training/fine-tuning datasets?

2. Method

2.1. EMA Datasets

English data For the English EMA dataset, we use the HPRC dataset [7], which contains parallel acoustic and EMA data from eight American English native speakers (four male, four female). The reading material contains 720 sentences, each repeated in both normal and fast speaking rates. The data consist of eight EMA sensor trajectories at a sampling rate of 100 Hz and synchronized acoustics at 44,100 Hz, collected using a Northern Digital WAVE system. We selected six sensors as the articulatory target for our experiment, which are tongue tip (1cm back from apex, TT), tongue blade (TB), tongue root (TR), upper lip (UL), low lip (LL) and lower incisor (LI).

Dutch data The Dutch EMA data was collected originally for another project to compare the articulation of oral cancer speakers to control speakers [21]. This study utilizes the data of three (one male, two female) Dutch control speakers, which was collected in the Netherlands using a Northern Digital VOX [22] at a sampling rate of 400 Hz, and synchronized acoustic data at a sampling rate of 22,050 Hz. All speakers provided written informed consent and the protocol was approved by the Medical Ethical Review Board of the University of Groningen (NL79242.042.21). The dataset contains words in carrier phrases and sentences that contain the phonemes of Dutch with their respective distribution. We placed five sensors following the sensor adhesion procedure specified in [6]. Specifically, we placed two tongue sensors: the TT sensor was placed 1 cm behind the anatomical tip whereas the tongue back sensor was placed at the /k/ constriction. One sensor was placed on the lower incisor to track jaw movements and two sensors were placed on the vermillion border of the lower and upper lips.

For both English and Dutch EMA data, we consider the sensor's movement at the midsagittal plane in the anterioposterior and vertical direction (x and y-axis, respectively).

2.2. Speech SSL models

Wav2vec 2.0 is one of the best-performing acoustic pre-trained models, trained with raw waveforms as input using CNN and transformer-based networks and a contrastive loss. Previous study indicated that wav2vec 2.0-based speech representations were able to effectively recover articulatory data [23].

In this study, we select seven wav2vec 2.0 (w2v2) pre-trained models from Facebook's public repository on Huggingface[1], with different sizes, pre-training data, and fine-tuning data for representation extraction. The w2v2-base and w2v2-large models share the same pre-training data, which is 960h English speech from LibriSpeech [24], while the MMS (massively multilingual speech) models [25] are all pre-trained using a combination of several datasets containing 491kh of speech data covering 1,406 languages in total. The mms-1b model was further fine-tuned using MMS-lab, which contains 44.7kh of labeled speech in 1,107 languages. Adapter modules were introduced for ASR fine-tuning, with different sets of adapter weights trained specifically for each language. For more details, please refer to [25]. In this study, we use the fine-tuned model with either an English or a Dutch adapter, which are dubbed mms-1b-eng and mms-1b-nld, respectively. An overview of the SSL models is provided in Table 1.

[1] https://huggingface.co/facebook

Table 1: *Comparison of SSL models. The mms-1b-eng and mms-1b-nld models are different in their specification of language-specific adapter modules.*

Model	Dim	Layer	Pre-train	Fine-tune
w2v2-base	768	12	Libri-speech	-
w2v2-base-ft	768	12		Librispeech
w2v2-large	1024	24		-
w2v2-large-ft	1024	24		Librispeech
mms-300m	1024	24	491kh, 1406 languages	-
mms-1b	1280	48		-
mms-1b-eng	1280	48		MMS-lab
mms-1b-nld	1280	48		MMS-lab

2.3. Speaker-independent AAI systems

The pipeline of our speaker-independent AAI systems is shown in Figure 1. The acoustic signals are first preprocessed to extract SSL or MFCC (as a baseline) features. The extracted features are then fed into a BLSTM network to predict EMA trajectories. The detailed process of data preprocessing, model training, and evaluation is described below.

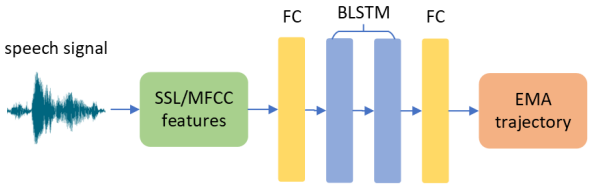


Figure 1: *Pipleline of our AAI systems.*

2.3.1. Data Preprocessing

The EMA data is first passed through a Butterworth low-pass filter of 10 Hz, and then z-scored within each utterance. The English and Dutch acoustic data are both downsampled to 16 kHz, and silent segments are removed from the data before feature extraction. In addition, we select MFCC as the baseline acoustic feature for comparison with the SSL representations. The 40-dim MFCCs are extracted using a 25 ms windows with a 10 ms shift. The EMA data is downsampled before training the AAI model to be aligned with the sampling rate of each acoustic feature, which is 100 Hz for the MFCC features and 50 Hz for the SSL representations.

2.3.2. Model Training and Evaluation

AAI is a time-related task, where the prediction of each articulatory movement is correlated with previous and following movements. Bidirectional Long Short-term Memory (BLSTM) has been shown to be a useful neural network for the AAI task, as it can learn proper temporal correlations of the corresponding contexts for predicting smooth articulatory trajectories [26, 27]. Specifically, our network architecture is based on the best architecture in [26], which consists of a fully-connected layer followed by two BLSTM layers with 150 hidden states. The final layer is another fully connected layer which outputs the EMA trajectories. The model thus comprises about 1.1 million parameters. The model is initialized using PyTorch's default initialization method, optimized using the root mean square error

Table 2: *Best layers per model based on six-fold cross-validation. The total nr. of layers is added between parentheses.*

Model	Best layer	Model	Best layer
w2v2-base	10 (12)	mms-300m	15 (24)
w2v2-base-ft	10 (12)	mms-1b	36 (48)
w2v2-large	17 (24)	mms-eng	28 (48)
w2v2-large-ft	11 (24)	mms-nld	29 (48)

(RMSE) loss and Adam optimizer, with a batch size of 64 and a learning rate of 0.0001. We used an NVIDIA Tesla A100 80GB PCIe GPU, with an average runtime of approximately 0.5 hours for training each model.[2]

To train speaker-independent AAI systems, we exclude two speakers, F04 and M04, from the HPRC dataset for model testing. For model comparison, early stopping, and hyperparameter selection, we employ a six-fold cross-validation methodology. Among the six remaining speakers from the HPRC dataset, five are allocated to the training set, and one is reserved for validation in each fold. Subsequently, the evaluation result on each test set is averaged across six folds to provide a comprehensive assessment. The AAI performance is evaluated using the Pearson Correlation Coefficient (PCC) between the ground truth articulatory data and the predicted trajectories, averaged over each utterance and each dimension. For the English data, the PCC is averaged over the 12-dimensional EMA data. For the Dutch EMA data, since we collected data of only two sensors on the tongue, we disregard the prediction results of the English model in the TR dimension and compute the average PCC over the remaining 10 dimensions of data.[3]

3. Results and Discussion

3.1. Layer-wise analysis of SSL representations

Previous research on utilizing SSL representations to enhance AAI performance typically selects the last layer of the model as input features [16, 17, 18]. However, recent studies on layer-wise analysis of SSL models in speech-related tasks have found that this is often not the optimal choice [20, 23]. Based on six-fold cross-validation, we chose the layers with the highest PCCs for each SSL model (see Table 2).

We observe similar trends of the PCC pattern for each model: PCCs increase rapidly from the intial layers, stabilize and peak around the middle-to-later layers, then decrease before the final three-to-four layers (see Figure 2). Our experiments suggest that the average best layer is not the final layer (cf. [20, 23]), but can be found around two thirds of the total number of layers.

3.2. Comparison across input features and test languages

The PCC results of AAI models trained on the English dataset using different input features for predicting English and Dutch test data are shown in Table 3. Instead of using a model trained on all six speakers, we use the six cross-validation models for each input feature (each trained on five speakers) to evaluate the performance on the held-out test set for both English and Dutch. Consequently, average and 95% confidence intervals of PCCs can be computed, which provides information about the robustness of the results. Furthermore, the relative performance drop, indicating the decrease in PCC from the matched English test set to the non-matching Dutch test set, is displayed in the last column of Table 3. Note that the optimal layers per model where selected using six-fold cross-validation when being evaluated on the single held-out speaker (not the test set).

[2]Code for feature extraction and model training can be found at `https://github.com/haoyunlf/aai`

[3]Through preliminary experiments, we found that the TB sensor in the HPRC dataset is closer to the tongue back sensor in our Dutch dataset than the TR sensor.

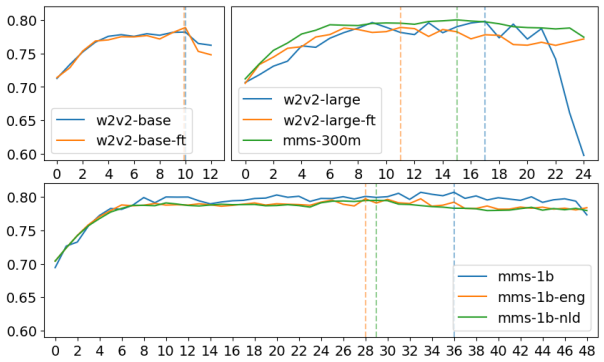


Figure 2: *Layer-wise PCCs based on six-fold cross-validation of base models (top left), large models (top right) and huge models (bottom). Dashed vertical lines denote optimal layers for each feature. Layer-0 is the output of CNN encoder.*

Table 3: *PCC results for each input feature on English and Dutch test sets. Average and 95% confidence intervals of PCCs for each input feature were computed over the prediction by six AAI models trained with six-fold cross-validation. Relative drop denotes the relative drop of average PCC from matched language (English) to non-matched language (Dutch).*

Feature	PCC English	PCC Dutch	Relative drop
MFCC	0.683 ±0.007	0.330 ±0.018	51.7%
w2v2-base	0.781 ±0.006	0.484 ±0.010	38.0%
w2v2-base-ft	0.779 ±0.005	0.410 ±0.024	47.3%
w2v2-large	0.793 ±0.004	0.526 ±0.007	33.7%
w2v2-large-ft	0.786 ±0.005	0.488 ±0.007	37.9%
mms-300m	0.795 ±0.004	0.553 ±0.012	30.4%
mms-1b	**0.796** ±0.004	**0.564** ±0.012	**29.2%**
mms-1b-eng	0.794 ±0.003	0.506 ±0.011	36.2%
mms-1b-nld	0.793 ±0.002	0.502 ±0.020	36.8%

MFCC vs SSL The observed pattern is similar for both test sets. All SSL features significantly outperform MFCC features. This indicates that SSL models have learned more effective information about speech articulation. For both test sets, mms-1b performs best, followed by mms-300m. The relative performance drop for the mismatched Dutch test set also decreased to less than 30%.

Capacity When considering the impact of model size on AAI performance, we observe that the w2v2-large outperforms w2v2-base, and mms-1b outperforms mms-300m. This suggests that increasing the model capacity is an effective way to better learn articulatory representations, consistent with the findings of [23].

Pre-training To assess the influence of model pre-training data on AAI performance, we compare w2v2-large and mms-

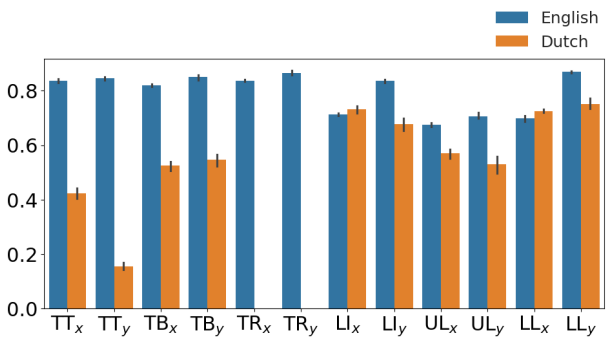


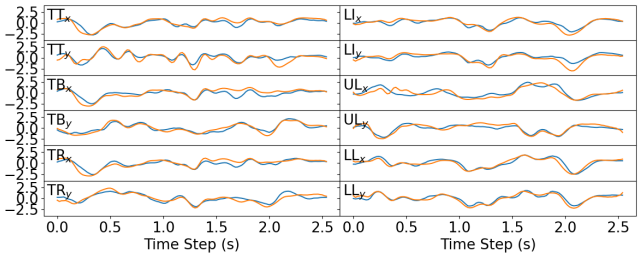
Figure 3: *PCC values for each articulator, predicted using mms-1b features for English and Dutch test sets. For TR, there are only predictions for English, as we only collected data for two sensors on the tongue for Dutch EMA data.*

300m features which share the same number of parameters but differ in their training data. The results indicate that mms-300m performs considerably better, and shows that more acoustic data with more languages in the pre-training stage also facilitate the model in learning articulatory representations.

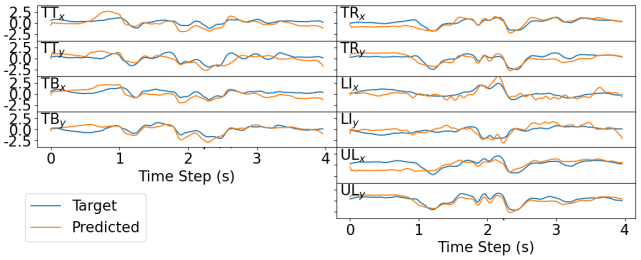
Fine-tuning When comparing the impact of fine-tuning data, we found that all fine-tuned models (w2v2-base-ft, w2v2-large-ft, mms-1b-eng , and mms-1b-nl) perform worse than their non-fine-tuned counterparts, regardless of language. This finding therefore contradicts the findings of Cho and collegues [23]. In our (cross-speaker) evaluation procedure, fine-tuning on ASR does not help the model to better learn articulatory representations, even when the fine-tuning language matches the AAI target language. This could be due to the training objective of ASR, which prioritizes differences between phonemes while overlooking variations within phoneme categories. Instead, the objective of AAI requires models to predict finer phonetic variations for each phoneme. This finding is particularly meaningful for low-resource languages, as other methods besides fine-tuning on an ASR task should be used to better utilize (limited) available data for improving AAI performance.

3.3. Comparison across articulatory sensors

Figure 3 illustrates the performance for each articulatory sensor predicted by the model using the best (mms-1b) features. This graph visualizes that for English the predictions for tongue sensors (TT, TB, TR) are better than those for other sensors. This trend is consistent with the findings obtained by [12, 16, 17], but differs from those reported in [23]. This difference could be caused by the use of different corpora, a limited number of speakers, and a many-to-one correspondence between articulatory configurations and resulting acoustic output. When the test data is in Dutch, the pattern is opposite, with predicted tongue sensor positions being inferior to those for lip and lower incisor sensors. This may be due to the tongue, being the primary active articulator, requiring more delicate control to produce nuanced speech differences than is required for the movement of lips and jaw. As a consequence, a greater mismatch between acoustics and tongue movement compared to jaw and lip movement may be expected in a cross-language setting. An example of the target and predicted trajectories for each language (see Figure 4) also shows that predictions for the Dutch tongue sensors shows a greater deviation from the targets compared to English.



(a) *English utterance 'The birch canoe slid on the smooth planks', produced by speaker F04.*



(b) *Dutch utterance 'hij heeft tamme baat gezegd' (he has said tame benefit), produced by speaker S01.*

Figure 4: *Examples of target and predicted articulatory trajectories by using mms-1b input features.*

4. Conclusion

This study is the first to investigate the effectiveness of SSL speech representations for cross-lingual AAI. Based on experiments with English training data and testing on (unseen) Dutch articulatory data, we found that using the best-performing SSL features reduces the performance drop to less than 30% (**RQ1**). Furthermore, our cross-lingual performance (PCC of 0.564) represents a 9.6% improvement over the 0.51 reported in the only other study on Dutch AAI [12]. Additionally, we found that increasing model size, selecting appropriate intermediate layers, and including more pre-training data may help improve AAI performance (**RQ2**). Fine-tuning the SSL model on an ASR task reduced the performance of AAI, however.

Limitations A limitation of this study is that since the two languages were collected by different labs using different EMA devices (the NDI Vox is more precise than the NDI Wave; [22]) and slightly different sensor placement, we are not able to disentangle the reduction in performance due to a different native language versus using a different articulatory corpus. Future research therefore ideally would compare cross-language AAI performance in a single dataset containing speakers of multiple languages.

Future work As we have achieved promising results in cross-lingual AAI by selecting appropriate SSL representations with a simple BLSTM network, future studies should explore further enhancement of cross-lingual AAI by integrating SSL with more complex networks such as transformer [16] or other techniques such as multi-task learning [28]. Our ultimate goal is to be able to predict accurate articulatory trajectories from acoustics for any unseen language or speaker. Such a system can then be used to generate additional articulatory features on the basis of acoustics which may help in enhancing performance in various speech-related tasks, such as low-resource ASR [29, 30].

5. Acknowledgments

This work was partly supported by the China Scholarship Council (CSC). We are also grateful to the ILSE project of the Center for Information Technology, University of Groningen, for providing access to the computing resources utilized in this research.

6. References

[1] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory information for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1913–1924, 2010.

[2] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4624–4627.

[3] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into hmm-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[4] A. Suemitsu, J. Dang, T. Ito, and M. Tiede, "A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning," *The Journal of the Acoustical Society of America*, vol. 138, no. 4, pp. EL382–EL387, 2015.

[5] O. Engwall, O. Bälter, A.-M. Öster, and H. Kjellström, "Designing the user interface of the computer-based speech training system artur based on early user tests," *Behaviour & Information Technology*, vol. 25, no. 4, pp. 353–365, 2006.

[6] T. Rebernik, J. Jacobi, R. Jonkers, A. Noiray, and M. Wieling, "A review of data collection practices using electromagnetic articulography," *Laboratory Phonology*, vol. 12, no. 1, p. 6, 2021.

[7] M. Tiede, C. Y. Espy-Wilson, D. Goldenberg, V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017.

[8] A. Wrench, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th seminar on speech production: models and data, 2000*, 2000.

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

[10] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[11] A. Ji, J. J. Berry, and M. T. Johnson, "The electromagnetic articulography mandarin accented english (ema-mae) corpus of acoustic and 3d articulatory kinematic data," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7719–7723.

[12] G. Sivaraman, C. Espy-Wilson, and M. Wieling, "Analysis of Acoustic-to-Articulatory Speech Inversion Across Different Accents and Languages," in *Proc. Interspeech 2017*, 2017, pp. 974–978.

[13] A. Illa, A. Nair, and P. K. Ghosh, "The impact of cross language on acoustic-to-articulatory inversion and its influence on articulatory speech synthesis," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8267–8271.

[14] T. Yan, K. Maekawa, Y. Nota, and M. Hirata, "Combining language corpora in a Japanese electromagnetic articulography database for acoustic-to-articulatory inversion," in *Proc. INTERSPEECH 2023*, 2023, pp. 1464–1467.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[16] S. Udupa, S. C, and P. K. Ghosh, "Improved acoustic-to-articulatory inversion using representations from pretrained self-supervised learning models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[17] S. K. Maharana, K. K. Adidam, S. Nandi, and A. Srivastava, "Acoustic-to-articulatory inversion for dysarthric speech: Are pretrained self-supervised representations favorable?" in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, accepted.

[18] P. Wu, L.-W. Chen, C. J. Cho, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Speaker-independent acoustic-to-articulatory speech inversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[19] C. J. Cho, A. Mohamed, A. W. Black, and G. K. Anumanchipalli, "Self-supervised models of speech infer universal articulatory kinematics," *arXiv preprint arXiv:2310.10788*, 2023.

[20] M. Bartelds, W. de Vries, F. Sanal, C. Richter, M. Liberman, and M. Wieling, "Neural representations for modeling variation in speech," *Journal of Phonetics*, vol. 92, p. 101137, 2022.

[21] T. B. Tienkamp, T. Rebernik, R. Buurke, K. Polsterer, R. J. J. H. van Son, M. B. Wieling, M. J. H. Witjes, S. A. H. J. de Visscher, and D. Abur, "The effect of speaking style on the articulatory-acoustic vowel space in individuals with tongue cancer before and after surgical treatment," in *Proceedings of the 13th International Seminar on Speech Production*, 2024, pp. 65–68.

[22] T. Rebernik, J. Jacobi, M. Tiede, and M. Wieling, "Accuracy assessment of two electromagnetic articulographs: Northern digital inc. wave and northern digital inc. vox," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 7, pp. 2637–2667, 2021.

[23] C. J. Cho, P. Wu, A. Mohamed, and G. K. Anumanchipalli, "Evidence of vocal tract articulation in self-supervised learning of speech," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[25] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[26] P. Zhu, L. Xie, and Y. Chen, "Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[27] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4450–4454.

[28] Y. M. Siriwardena, G. Sivaraman, and C. Espy-Wilson, "Acoustic-to-articulatory Speech Inversion with Multi-task Learning," in *Proc. Interspeech 2022*, 2022, pp. 5020–5024.

[29] M. Morshed and M. Hasegawa-Johnson, "Cross-lingual articulatory feature information transfer for speech recognition using recurrent progressive neural networks," in *Proc. Interspeech 2022*, 2022, pp. 2298–2302.

[30] Z. Yue, E. Loweimi, Z. Cvetkovic, H. Christensen, and J. Barker, "Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7372–7376.