# End-to-end acoustic-articulatory dysarthric speech recognition leveraging large-scale pretrained acoustic features

Yue, Z.; Zhang, Y.

**DOI**
10.1109/ICASSP49660.2025.10888412

**Publication date**
2025

**Document Version**
Final published version

**Published in**
International Conference on Acoustics, Speech and Signal Processing (ICASSP)

ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888412

# End-to-end acoustic-articulatory dysarthric speech recognition leveraging large-scale pretrained acoustic features

Zhengjun Yue
*Multimedia Computing Group*
*Delft University of Technology*
the Netherlands
z.yue@tudelft.nl

Yuanyuan Zhang
*Multimedia Computing Group*
*Delft University of Technology*
the Netherlands
y.zhang-44@tudelft.nl

***Abstract*—Automatic dysarthric speech recognition (ADSR) remains challenging due to the irregularities in speech caused by motor control impairments and the limited availability of dysarthric speech data. This paper explores the integration of articulatory features, captured using Electromagnetic Articulography (EMA), with both conventional acoustic features and those extracted from large-scale pretrained models including Whisper and XLSR-53 as well as the fine-tuned Whisper model. We propose end-to-end (E2E) Conformer-based acoustic-articulatory models for ADSR and compare their performance against the corresponding hybrid TDNNF models. The experimental results show that using the fine-tuned Whisper features (Whisper-FT) fused with articulatory features achieves the lowest (10.5%) word error rate (WER) on dysarthric speech, with particularly significant improvements for severely dysarthric speech, reaching a WER of 20.8%.**

***Index Terms*—dysarthric speech recognition, articulatory-acoustic multi-modal, large-scale pretrained acoustic features**

## I. Introduction

Dysarthria is a motor speech disorder caused by a neuromotor interface disruption [1] that affects the clarity and intelligibility of spoken language. Individuals with dysarthria often struggle to control their articulatory movements, resulting in irregular articulatory patterns and less intelligible speech. Automatic dysarthric speech recognition (ADSR) is a highly challenging task due to the large speech variability and data scarcity. Moreover, due to the large mismatch between typical and dysarthric speech, mainstream ASR systems, such as Whisper [2], designed for typical speech do not perform well on dysarthric speech [3], [4]. There is a growing need for more robust models capable of recognising dysarthric speech, especially as individuals with speech impairments increasingly rely on ASR technologies for their daily communication and accessibility.

To address these challenges, integrating features from other modalities that are correlated with the speech signal, such as visual and articulatory data, has shown potential for improving ADSR performance [5]–[7]. Articulatory measurements, which capture the movements of speech articulators such as lips and tongue reflect a direct representation of the speech production process. Compared to acoustic features, articulatory data has been found to be less sensitive to speaker variability [8] and more effective in modelling coarticulation effects [9]. Therefore, incorporating articulatory information alongside acoustic data has the potential to offer complementary insights for improving ADSR performance.

Because of the lack of multi-modality dysarthric datasets, most previous research on dysarthric speech recognition relies on acoustic features such as mel-frequency cepstral coefficients (MFCCs) [10] or filterbank (FBank) features [11], [12]. However, these features may not capture the irregular patterns in dysarthric speech. Articulatory features, derived from the movements of the speech articulators, have been shown to provide valuable complementary information [13], [14], especially when acoustic features alone are insufficient. End-to-end (E2E) models have been more and more widely used for dysarthric speech recognition [15]–[17]. However, due to the limited amount of paired audio and articulatory data, E2E acoustic-articulatory ASR models are under-explored.

Recently, large-scale models trained on vast amounts of speech data, such as Whisper [2] and XLSR-53 [18], [19] have revolutionised ASR by offering robust, pretrained models that can generalise across different domains and speaker variations with minimal adaptation. These models extract rich, high-level acoustic features that have proven effective for a variety of speech recognition tasks. However, these features have only been applied to the single-modal ASR system [20], [21] and little attention has been given to integrating these powerful features with articulatory data, particularly in the context of dysarthric speech recognition.

In this paper, we addressed the research gaps mentioned above by building multi-modal acoustic-articulatory ADSR systems using Conformer-based E2E models. To the best of our knowledge, this is the first attempt to leverage the strengths of large-scale models such as Whisper and XLSR-53 for acoustic feature extraction, while integrating the real recorded articulatory data to improve the recognition performance for dysarthric speech. We compared this approach to models using conventional acoustic features (i.e., FBank). We also

explored the effectiveness of Whisper features extracted from the fine-tuned Whisper model on the in-domain (i.e., target) dataset. In addition, we compared the E2E model with the state-of-the-art hybrid (time delay neural network) TDNNF model [22], [23] and employed the pretrained Whisper-large model for zero-shot testing and fine-tuning. We presented results across various acoustic features, severity levels and ASR models. Experimental results show up to 3.0% absolute (22.2% relative) WER reduction for dysarthric speech using fine-tuned Whisper features (Whisper-FT+EMA vs. Whisper-FT) when integrating articulatory features.

## II. Methodology

### A. Feature extraction

We employed three types of features for the ADSR task in this paper: conventional acoustic features (i.e., FBank), large-scale pretrained acoustic features (i.e., XLSR-53 features, Whisper features, fine-tuned Whisper (Whisper-FT) features), and articulatory features.

*1) Acoustic features:* FBank features are commonly used in standard ASR systems [24], [25] and in acoustic-articulatory speech recognition systems [13]. In our experiments, we used 80-D FBank features with a frame shift of 10 ms and a frame length of 25 ms, as a baseline. We extracted large-scale pretrained features from the XLSR-53 and Whisper[1] models. In particular, given the previous state-of-the-art (SOTA) performance of XLSR-53 features achieved in the ADSR task [21], we extracted 1024-D XLSR-53 features [19], with 20 ms frame rate from the model's 20th[2] encoder layer. In addition, we extracted 1280-D Whisper features with a 20 ms frame rate from the output of the Whisper model encoder [2].

Considering both XLSR-53 and Whisper models were trained on large-scale multilingual typical speech, we hypothesise that their extracted features may still mismatch with the dysarthric speech data. While XLSR-53 is a self-supervised model that learns robust speech representations without a decoder, Whisper is a supervised ASR model trained with a strong decoder. To address this potential mismatch during feature extraction, we fine-tuned the Whisper model on the TORGO dataset and used the fine-tuned model to extract Whisper-FT features.

*2) Articulatory features:* Each articulatory data sample is measured by 12 sensors capturing articulatory movements in three dimensions, returning sensor positions in Cartesian coordinates (x, y, z) along with the spatial orientation angles [26], [27]. The sensors are attached to the tongue back (TB), tongue middle (TM), tongue tip (TT), forehead, bridge of the nose, upper lip (UL), lower lip (LL), lower incisor, left and right mouth, left and right ear. The articulatory data is downsampled and aligned with the acoustic features to create synchronised feature sequences. We pre-process the articulatory data following the process outlined in [14]. According to previous work and our testing, we used the pairwise Euclidean Distance (ED) between the UL and LL as articulatory features, computed from their respective Cartesian coordinates (UL_x,UL_y,UL_z) and (LL_x,LL_y,LL_z).

### B. ASR models

We employed two types of ASR models in this paper: an E2E ASR model and a hybrid ASR model. We employed the Conformer-based E2E model, which has demonstrated a strong performance in dysarthric speech [21], using the ESPnet toolkit [28]. We tuned the model to achieve the best performance for this task[3]. When using large pretrained models as a frontend for acoustic feature extraction, we added a linear projection layer with an output size of 80 [21]. For the hybrid model, we implemented the TDNNF architecture following the SOTA approach for hybrid ADSR system [23], using the Kaldi toolkit [29].

The purpose of comparing the E2E and hybrid models with the same input features is to assess how these different architectures handle identical acoustic and articulatory data. Since E2E models typically require large amounts of data for training, we are particularly interested in evaluating their performance in scenarios where only limited paired data is available. Especially when using large-scale pretrained acoustic features, we aim to determine which model—E2E or hybrid—performs better under these constraints, both with or without articulatory information.

## III. Experiment

### A. Dataset

The TORGO [26] dataset is the only publicly available resource containing 13127 aligned audio and articulatory recordings collected from 14 speakers[4]. Each articulatory recording is paired with two sets of audio recordings captured using a head-mounted microphone and an array microphone. The articulatory measurements were collected using a 3D AG500 electromagnetic articulography (EMA) system. The advantages of using real articulatory data, as opposed to synthetic articulatory data generated through acoustic-to-articulatory mapping [30]–[32], have been discussed in [14].

We used 70% of the dataset for training, 10% for validation, and 20% for testing [33]. It is important to ensure that the same utterance recorded using the head and array microphones (e.g., F03-Session1-array-0001 and F03-Session1-head-0001), was included in the same subset-either training or test. Table I presents the detailed data split including the number of hours, speakers and utterances allocated to various subsets. The

[1]Throughout this paper, "Whisper" refers to the Whisper-large-V2 model.

[2]We determined that the 20th layer provided optimal embeddings after testing various layers of the XLSR-53 model.

[3]The model consists of 12 Conformer encoder layers and 6 Transformer decoder layers, both with output dimensions of 256. The attention mechanism uses 4 attention heads. The feed-forward layers have 1024 units in the encoder and 2048 units in the decoder. The model was trained with a CTC weight of 0.3, an attention weight of 0.7, and 500 BPE units.

[4]Although the TORGO dataset includes 15 speakers (Eight of the speakers (5 males, 3 females) of the speakers have dysarthria ranging from mild to severe, while the remaining seven (4 males, 3 females) are typical speakers.), articulatory data is not available for speaker F01.

test set was further divided according to different dysarthria severity levels and different types of speech (e.g., dysarthric and typical).

TABLE I
DATA SPLIT FOR TRAINING AND TEST.

| Name | Dur (h) | Spk | Utterances |
|---|---|---|---|
| Train | 7.27 | 14 | 8546 |
| Valid | 1.07 | 14 | 1232 |
| Test | 2.13 | 14 | 2459 |
| Test subsets 1 | | | |
| Dys | 0.79 | 7 | 735 |
| Typ | 1.34 | 7 | 1724 |
| Test subsets 2 | | | |
| Severe | 0.27 | 3 | 212 |
| M/S | 0.17 | 1 | 122 |
| Moderate | 0.11 | 1 | 141 |
| Mild | 0.23 | 2 | 260 |

### B. Experiment Setup

All experiments apply two-fold speed perturbation [34] to the training data with factors of 0.9 and 1.1 using SOX [35]. For the Conformer and the Whisper fine-tuning model, we used a batch bin size of 800,000 [28], with up to 50 epochs and 2000 iterations per epoch. Early stopping (patience: 3) was based on validation loss, and the final model was averaged from the top 5 with the highest validation accuracy [36]. Decoding used a beam size of 10, and for Whisper zero-shot testing, the temperature was set to 0 additionally. For the hybrid TDNNF models, following [23], we trained for 10 epochs with a beam size of 15 and a lattice beam size of 8. Due to the prompt overlap issue in the TORGO dataset, we used the out-of-domain Librispeech language model (LM) [37] for decoding, ensuring fair results [38].

The original articulatory (EMA) features (5 ms frame rate) were resampled to 10 ms and 20 ms to align with the FBank and acoustic features extracted from XLSR-53, Whisper, and Whisper-FT. We applied utterance-level mean-variance normalization to both the EMA and acoustic features before concatenating them along the time dimension.

## IV. RESULTS AND DISCUSSION

The first block in Table II presents the WERs for different acoustic features, with and without EMA features, across various severity groups of dysarthric speech and averaged for both dysarthric (Dys) and typical (Typ) speech, using the Conformer ASR model. The *Whisper-FT+EMA* model achieves the best performance (10.5% WER) for dysarthric speech, particularly for the severe (20.8%) and moderate-to-severe (M/S) (7.6%) groups[5].

Fig. 1 illustrates the performance gain achieved by integrating EMA features. The left figure in Fig. 1 compares the performance gain for dysarthric (Dys) and typical (Typ) speech. For dysarthric speech, combining EMA features with Whisper-FT yields the highest improvement (+22.22%). In contrast,

[5]We performed statistical significance tests using the Matched Pairs Sentence-Segment Word Error (MAPSSWE) method [39], following [40].

TABLE II
WERS(%) OF THE ACOUSTIC AND ACOUSTIC-EMA FEATURES EXPERIMENTS ON TORGO TEST SETS.

| Feature | Severity group | | | | Average | |
|---|---|---|---|---|---|---|
| | *Severe* | *M/S* | *Mod* | *Mild* | *Dys* | *Typ* |
| **Conformer model** | | | | | | |
| FBank | 66.1 | 62.3 | 17.6 | 8.4 | 37.1 | 17.7 |
| XLSR-53 | 44.6 | 34.2 | 10.2 | 3.2 | 22.6 | 11.2 |
| Whisper | 60.0 | 26.0 | 13.1 | 3.5 | 26.6 | 12 |
| Whisper-FT | 32.5 | 8.8 | **8.2** | **1.2** | 13.5 | **10.1** |
| FBank+EMA | 66.8 | 75.7 | 22.9 | 10.6 | 41.5 | 17.1 |
| XLSR-53+EMA | 50.5 | 33.9 | 13.4 | 8.8 | 26.7 | 16.1 |
| Whisper+EMA | 49.2* | 27.5 | 16.9 | 9.5 | 26.2 | 17.1 |
| Whisper-FT+EMA | **20.8**** | **7.6** | 10.9 | 1.7 | **10.5*** | 11.3 |
| **Hybrid-TDNNF model** | | | | | | |
| Whisper | 88.5 | 74.0 | 80.6 | 52.2 | 73.8 | 65.4 |
| Whisper-FT | 47.1 | 25.2 | 37.6 | 18.1 | 33.3 | 16.6 |
| Whisper+EMA | 71.7 | 40.6 | 43.9 | 21.5 | 45.8 | 19.5 |
| Whisper-FT+EMA | 54.4 | 29.0 | 35.2 | 18.4 | 35.3 | 16.5 |
| **Whisper model** | | | | | | |
| Zero-shot test | 136.6 | 42.3 | 34.4 | 10.0 | 59.0 | 10.8 |
| FT-Whisper | 19.0 | 7.2 | 10.7 | 1.5 | 9.8 | 2.8 |

* $p < .05$; ** $p < .001$; *** $p < .001$.
Statistical significance tests were performed between the acoustic-only and acoustic + EMA experiments.

FBank shows a moderate performance drop (-11.86%), indicating that conventional acoustic features do not benefit from EMA data in the E2E model, unlike the hybrid ASR models in [38]. XLSR-53 features also show a decline (-18.14%) while and Whisper features show a slight increase (+1.50%), suggesting that without fine-tuning, large-scale models struggle to leverage EMA information efficiently. For typical speech, integrating EMA features results in performance declines across most features, except for FBank (+3.39%). This indicates that typical speech is already well-represented by large-scale pretrained acoustic features, whereas FBank features need to benefit from additional complementary EMA features to enhance performance.

The right figure in Fig. 1 helps to better understand how EMA features impact the performance across different levels of dysarthric speech severity when using various acoustic features. It demonstrates that the integration of EMA features benefits the most for more severely impaired dysarthric speech. reducing WER by 36% for severe cases when using Whisper-FT. However, for moderate and mild dysarthria, EMA features reduce performance across all acoustic features, with Whisper showing a 171.43% WER increase. Interestingly, the performance decline is less pronounced with Whisper-FT (41.67% vs. 171.43%), further demonstrating the value of fine-tuning Whisper for feature extraction.

Overall, EMA features are particularly useful for recognising more severe dysarthric speech, where the acoustic signal alone may be insufficient. However, as the severity of the speech impairment decreases, the benefits of EMA features diminish, and in some cases, their inclusion can reduce model performance. This highlights the need for selective integration based on severity.

Fig. 2 shows the training dynamics in terms of CTC-

attention loss vs. epoch on the validation set for various acoustic feature sets, with and without EMA features. Whisper-FT features exhibit the fastest convergence and achieve the lowest validation loss, both with and without EMA features. The integration of EMA features tends to benefit conventional acoustic features (i.e., FBank) more noticeably, leading to faster convergence and improved validation performance[6]. However, for large-scale models such as Whisper and XLSR-53, EMA features have negative impacts, as these models already capture extensive acoustic information, reducing the need for complementary articulatory information. When applying Whisper-FT features, fine-tuning allows the integration of EMA features to provide slightly additional benefits.

For a fair comparison as mentioned in Section II-B, we trained four hybrid TDNNF ADSR systems using the following inputs: (1) Whisper, (2) Whisper-FT, (3) Whisper+EMA, and (4) Whisper-FT+EMA. As shown in the middle block of Table II, although integrating EMA features with Whisper features significantly improves performance compared to Whisper features alone (45.8% vs. 73.8%), the hybrid TDNNF systems consistently perform worse than the corresponding Conformer models across all severity groups. This further highlights the effectiveness of the Conformer architecture, especially when using large-scale pretrained features and integrating EMA features, making it the more effective choice for ADSR.

In addition to primary experiments, we implemented zero-shot testing using Whisper and a fully fine-tuned Whisper model (FT-Whisper (Note that FT-Whisper refers to the fine-tuned Whisper ASR model, while Whisper-FT refers to the acoustic features extracted from the fine-tuned Whisper model, which are then used in either the Conformer or TDNNF ASR models.), where only speech data can be used during fine-tuning. As shown in the bottom block of Table II, the zero-shot testing results are much worse while FT-Whisper achieves a significantly lower WER across all dysarthria severity levels. Although these results outperform our proposed *Whisper-FT+EMA Conformer* model, we attribute this to the strong language modelling capabilities of the decoder in the Whisper model [2] which was fine-tuned on the TORGO training set's prompts (less-variant and easy-learned-read prompts). While our Conformer models do not use LMs and the TDNNF models leverage a Librispeech LM for decoding, to avoid unfair decoding advantages caused by the overlapped training and test prompts [38]. In addition, Whisper-FT benefits from a large amount of out-of-domain data in the acoustic model, while the Whisper-FT+EMA Conformer is trained solely on in-domain data. Despite this, the Whisper-FT+EMA Conformer system achieves comparable results (10.5% vs. 9.8%). Therefore, despite the slightly lower WER achieved by the FT-Whisper model, we believe our proposed Whisper-FT+EMA Conformer model remains the best-performing system when evaluated on a fair and consistent basis.

[6]Note that the validation loss is averaged across both Dys and Typ speech. While WER increases for Dys, it decreases for Typ speech, leading to overall performance improvement.

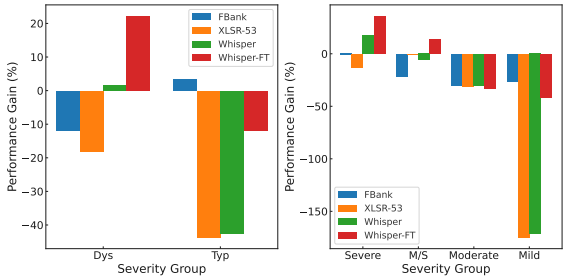


Fig. 1. Performance gain achieved by integrating EMA features across different severity groups.

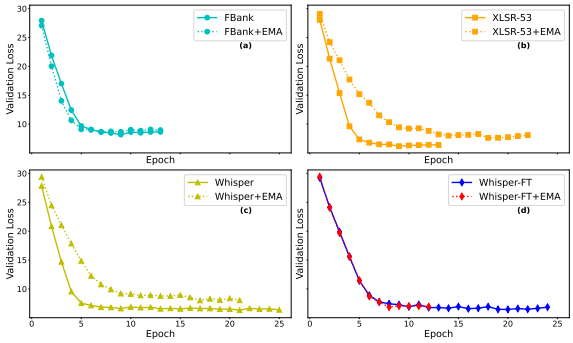


Fig. 2. Training dynamics in terms of CTC-attention loss vs. epoch on the validation set for different features. (a) Fbank w/o EMA, (b) XLSR-53 w/o EMA, (c) Whisper w/o EMA, (d) Whisper-FT w/o EMA.

## V. Conclusion

This paper explored the integration of articulatory features with acoustic features extracted from large-scale models (Whisper, XLSR-53 and fine-tuned Whisper) using Conformer-based E2E models for ADSR. We also compared the performance with hybrid TDNNF models and the Whisper ASR models. The results demonstrate that the choice of acoustic features and the integration of articulatory data improve performance, particularly for more severely impaired speech. The *Whisper-FT+EMA Conformer* model achieves the best performance, with a WER of 10.5% for dysarthric speech and a 36% improvement for the severe group, significantly outperforming conventional FBank features.

Articulatory features were most beneficial for severely impaired speech, while their impact on mild dysarthria and typical speech was limited or negative. Large-scale pretrained acoustic features such as Whisper and XLSR-53 performed well, but fine-tuning is important for fully leveraging the articulatory data. This highlights the importance of using large-scale pretrained acoustic features and fine-tuning to enhance model performance. These findings emphasise the importance of selectively integrating articulatory features based on speech severity and acoustic features, and perhaps acoustic models. Future work includes exploring the usefulness of other modalities and further optimising the integration of articulatory features with large-scare pretrained acoustic features in E2E models.

## REFERENCES


[1] W. Gowers, "Clinical speech syndromes of the motor systems," *Neurology for the Speech-Language Pathologist. Fifth edition. Philadelphia: Butter worth_ Heinemenn*, pp. 196–203, 2001.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[3] I. Calvo, P. Tropea, M. Viganò, M. Scialla, A. B. Cavalcante, M. Grajzer, M. Gilardone, and M. Corbo, "Evaluation of an automatic speech recognition platform for dysarthric speech," *Folia Phoniatrica et Logopaedica*, vol. 73, no. 5, pp. 432–441, 2021.

[4] H. Christensen, C. Fox, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech." 2012.

[5] E. Salama, R. El-Khoribi, and M. Shoman, "Audio-visual speech recognition for people with speech disorders," *International Journal of Computer Applications*, vol. 96, no. 2, 2014.

[6] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech," in *NIPS*, 2010, pp. 70–78.

[7] F. Xiong, J. Barker, and H. Christensen, "Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018.

[8] O. Fujimura, "Relative invariance of articulatory movements, in invariance and variability in speech processes," *Lawrence Erlbaum*, pp. 226–242, 1986.

[9] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *ICSLP*, 2000.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[11] "Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition," in *Proceedings of Interspeech 2020*. International Speech Communication Association (ISCA), 2020, pp. 4581–4585.

[12] S. Liu, M. Geng, S. Hu, X. Xie, M. Cui, J. Yu, X. Liu, and H. Meng, "Recent progress in the cuhk dysarthric speech recognition system," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2267–2281, 2021.

[13] E. Yılmaz, V. Mitra, C. Bartels, and H. Franco, "Articulatory features for asr of pathological speech," in *Proc. INTERSPEECH*, 2018, pp. 2958–2962.

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

[15] A. Almadhor, R. Irfan, J. Gao, N. Saleem, H. T. Rauf, and S. Kadry, "E2e-dasr: End-to-end deep learning-based dysarthric automatic speech recognition," *Expert Systems with Applications*, vol. 222, p. 119797, 2023.

[16] S. R. Shahamiri, "Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 852–861, 2021.

[17] R. Doshi, Y. Chen, L. Jiang, X. Zhang, F. Biadsy, B. Ramabhadran, F. Chu, A. Rosenberg, and P. J. Moreno, "Extending parrotron: An end-to-end, speech conversion and speech recognition model for atypical speech," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6988–6992.

[18] 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.

[19] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. INTERSPEECH*, 2021, pp. 2426–2430.

[20] X. Chang, T. Maekaku, P. Guo, J. Shi, Y.-J. Lu, A. S. Subramanian, T. Wang, S.-w. Yang, Y. Tsao, H.-y. Lee *et al.*, "An exploration of self-supervised pretrained representations for end-to-end speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 228–235.

[21] A. Hernandez, P. A. Pérez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," *Proc. INTERSPEECH*, pp. 51–55, 2022.

[22] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," in *Backpropagation*. Psychology Press, 2013, pp. 35–61.

[23] E. Hermann and M. M. Doss, "Dysarthric speech recognition with lattice-free mmi," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6109–6113.

[24] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.

[25] Y. Zhang, A. Herygers, T. Patel, Z. Yue, and O. Scharenborg, "Exploring data augmentation in bias mitigation against non-native-accented speech," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[26] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, pp. 523–541, 2012.

[27] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, 2010.

[28] S. Watanabe *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[30] S. Hu, S. Liu, X. Xie, M. Geng, T. Wang, S. Hu, M. Cui, X. Liu, and H. Meng, "Exploiting cross domain acoustic-to-articulatory inverted features for disordered speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6747–6751.

[31] S. K. Maharana, A. Illa, R. Mannem, Y. Belur, P. Shetty, V. P. Kumar, S. Vengalil, K. Polavarapu, N. Atchayaram, and P. K. Ghosh, "Acoustic-to-articulatory inversion for dysarthric speech by using cross-corpus acoustic-articulatory data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6458–6462.

[32] S. K. Maharana, K. K. Adidam, S. Nandi, and A. Srivastava, "Acoustic-to-articulatory inversion for dysarthric speech: Are pre-trained self-supervised representations favorable?" in *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 408–412.

[33] Z. Yue, E. Loweimi, H. Christensen, J. Barker, and Z. Cvetkovic, "Acoustic modelling from raw source and filter components for dysarthric speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2968–2980, 2022.

[34] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[35] B. Chris, R. Mans, Robs, and K. Ulrich, "SoX - Sound eXchange," [online]. Available: http://sox.sourceforge.net.

[36] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get m for free," in *International Conference on Learning Representations*, 2016.

[37] "Librispeech language models, vocabulary and g2p models." [online]. Available: https://www.openslr.org/11/.

[38] Z. Yue, F. Xiong, H. Christensen, and J. Barker, "Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6094–6098.

[39] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. of IEEE Int. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, 1989, pp. 532–535.

[40] "WER statistical significance test." [online]. Available: https://github.com/talhanai/wer-sigtest.