Speaker-Independent Automatic Speech Recognition System for Dysarthric Speech using wav2vec 2.0

Samyaka Patil, Eshan Srirambhatla and Kavish Kumar

Abstract

Automatic Speech Recognition Systems are commonly characterized by articulatory blurring, imprecise production of consonants, distortion of vowels and poor articulation of phenomes. Dysarthric speakers might need assistive technology to make it easier to communicate with others. Past attempts to create an Automatic Speech Recognition system for the TORGO database include speaker dependent GMM- HMM models, DNN models and more recently speaker dependent and speaker independent model that use wav2vec 2.0 as a feature extractor. This paper presents a speaker independent LSTM model that is pre-trained on LibriSpeech, a non-dysarthric speech dataset on which the model produces a WER of 1.3%. The model is then fine tuned on TORGO, a dysarthric speech dataset where the model produces a WER of 48% and CER of 27%.

Introduction

Automatic Speech Recognition (ASR) systems and their applications have been widely explored in the field of computational linguistics, with new models being developed at a growing frequency.

ASR, complimented by Augmentative and Alternative Communication (AAC), offer convenient human-computer interactions for individuals facing communication barriers due to factors such an irregularity in pronunciation and fatigue caused by speaking, unveiling scope for positive societal impact. However, general ASR models have high word error rates (WER) for disordered speech as compared to healthy speech as shown by a study stating that severely dysarthric subjects may have a word-error rate of 97.5% on modern systems against 15.5% for the general population [1].

Dysarthria is one such motor speech disorder, characterized by slurred speech and incorrect phoneme pronunciation. People affected by dysarthria can benefit greatly from ASR technologies, however there is still a gap in ASR accuracy for this community as compared to the healthy population.

Recent years have seen the rise of exceptional speech models such as wav2vec 2.0 that have demonstrated notable proficiency in feature extraction for speech [2]. However, ASR models for dysarthric speech so far have largely been speaker- dependent, limiting their accessibility for a wider audience. Exploration into speaker- independent models has been sparse, with models not achieving low WER. This is in part due to lack of available dysarthric speech data, which is a common challenge faced while developing such systems.

In this paper, we propose a speaker- independent ASR model built on wav2vec 2.0 [3] and further fine-tuned using a Bi- Directional three layered Long Short Term Memory (LSTM) model. Performance for the model is evaluated via WER and character error rate (CER) and benchmarked against previously built wav2vec 2.0 models for dysarthric speech using the TORGO dataset. Experiments demonstrate that the proposed architecture can reduce the overall WER for speaker-independent dysarthric speech recognition by reaching an overall WER of 48.75% and CER of 26.62%.

Background and Significance

Dysarthria is a speech sound disorder, resulting from neurological impairment to the motor component of the speech system. Commonly characterized by articulatory blurring, imprecise production of consonants, distortion of vowels and poor articulation of phenomes, it is a condition in which problems occur with the muscles responsible for the physical production of speech [4]. Dysarthria by itself does not cause problems with understanding language, and a person affected by dysarthria can formulate syntactically correct sentences, it is the difficulty in pronunciation that may render the speech un-intelligible.

This impairment in functional communication can severely decrease the quality of life of affected persons, making it difficult for both speaker and listener to converse. Furthermore, traditional translators for dysarthric patients also keep them from experiencing the same independence as their counterparts with healthy speech [5]. Severity and characteristics of dysarthric speech vary greatly between individuals and are heavily influenced by the severity of neurological damage/disease. Problems in different nerve regions manifest in different ways such as a hoarse voice, hypernasality, reduced vocal dexterity, slurred speech and inconsistent stress patterns all further complicate the intelligibility of speech, resulting in strained communication for such speakers and listeners [6]. Dysarthric speech is also accompanied by slowness and disfluencies such as pauses and stuttering, which worsen over time and as the speaker gets increasingly fatigues, making comprehension of sentences challenging.

However, it is important to note the varying severity of dysarthria. Some speakers with mild severity can maintain a relatively normal rate of speech and have relatively less articulatory discrepancies [7]. Speakers may also face increased speaking fatigue, making speech more difficult for extended durations of time. Dysarthria not only causes physical discomfort to a speaker, but also negatively affects their social interaction and overall quality of life [5].

Various neurological and motor disorders may give rise to dysarthria. These include Parkinson's disease, Huntington's disease, Amyotrophic Lateral Sclerosis (ALS), Lyme disease and Cerebral Palsy (CP) among others. These ailments cause lesions to areas of the brain involved in planning, executing or regulation motor operations in skeletal muscles such as those of the head or neck, which can result in the dysfunction or failure of the nervous system's ability to activate motor units and effect the correct range and strength of movements. Due to these physical limitations that often accompany dysarthria, affected persons may have limited ability to interact with computers, keyboards and the environment [8].

Therefore, people with dysarthria can benefit greatly from ASR systems, offering an interface to better communicate with a general audience. This paper explores how ASR models can help people with dysarthria better communicate, specifically employing a novel architecture using wav2vec 2.0 and Bi-Directional LSTM on the TORGO database, pre-trained on LibriSpeech.

While dysarthria is a highly variable condition, in this paper we have chosen not to treat it based on several categories that it could be divided into, but have rather trained a single model on all the different variabilities of the condition. This decision was guided by a few practical constraints of available data and our overarching goal of building a generalised, speaker-independent model. The TORGO dataset, which was our main dataset for this research, consists of a limited number of speech samples with mixed severity levels. Given this small sample size, further dividing dysarthria into subtypes would not provide a statistically meaningful subset for evaluation and could cause overfitting of the model.

The primary focus of this research lies in computational modelling rather than clinical subtyping. The study aims to evaluate whether deep learning architectures, such as wav2vec 2.0, can generalise effectively across the variable landscape of dysarthric speech. Designing a single, inclusive ASR system for all dysarthric speakers is also more scalable than developing subtype-specific models.

Literature Review

Dysarthric speech recognition falls under the domain of computational linguistics. Computational linguistics involves the way people pronounce and articulate their words, and how these features can be processed computationally to classify, recognize, transcribe or perform other similar automation processes to it. In the case of dysarthria, the phonemes and lexical patterns differ vastly from the standard conventions of human speech, making it increasingly difficult to effectively translate their sentences and words as done by linguists in the past. Prior to the advent of machine learning technology, translating dysarthric speech heavily relied on human interpretation and contextual

understanding. A dysarthric patient could also be assigned a personal translator, one who could effectively understand their pronunciation and articulation. However, finding such a person was not easy, and most people would continue to struggle to interact with the world [9].

With the progress of speech technology, computational linguistics has aided in the recognition and transcription of dysarthric speech. The goal of this was to reduce the need for personalized translators and to automate the speech recognition process itself. Dysarthric speech recognition and its research have shifted from early machine learning-based approaches to the more recent approaches involving deep learning and multi-modal techniques. Multi-modal techniques utilize audio-visual cues, and more recently give emphasis to the integration of visual information like lip movement alongside acoustic features. Past discussions in literature focus on effective methods for fusing audio and visual modalities and addressing challenges like the lack of data for dysarthric speech, and voice quality degradation in severe dysarthria [10, 11].

Initial research explored artificial neural networks (ANN) for simple acoustic modelling and tested them on dysarthric speech [12]. This early school of thought focused on adapting general ASR techniques using machine learning paradigms. Further research during this time looked into some articulatory differences in phoneme pronunciation between dysarthric speakers and healthy speakers, and how that impacted ASR accuracy in order to understand the specific challenges posed while building models for dysarthric speech recognition [13].

Studies during this time compared various acoustic features by using Fast Fourier Transform, Linear Predictive Coding, and Mel-Frequency Cepstral Coefficients (MFCC) within Hidden Markov Model (HMM) frameworks to identify efficient representations for dysarthric speech [14]. This was in an effort to refine the input to ASR systems based on specific characteristics of dysarthric speech.

Researchers have also worked on various methods to go about ASR for dysarthric speech, namely speaker-dependent, speaker- adaptive and speaker-independent models, with most historical research being in the speaker-dependent field. Speaker-dependent ASR systems have been developed with limited vocabulary. This approach focused on individual adaptation rather than generalized models. The developments of ASR systems based on context-dependant HMMs (CD-HMM) to recognize audio commands from speakers ranging from mild to severe dysarthria and methods to handle pronunciation variations in lexical models represent further advancements within the same machine learning era [15].

Comparisons have been drawn between different machine learning methods like HMMs and Support vector machines (SVM) for recognizing dysarthric speech, with various strengths and weaknesses of each being highlighted in the process [16]. Metrics used commonly include WER and Accuracy Rates.

The scarcity of dysarthric speech data was also recognized as a significant challenge during this time due to the lack of large- scale databases in the public domain. Dysarthric speech datasets are hard to procure, as getting a lot of people with dysarthria to speak for long durations of time is not only uncomfortable and painful for them but also expensive. A common workaround for this was data augmentation, which modified healthy speech into dysarthric speech using automated models and research on data augmentation techniques such as using deep auto-encoders to modify and perturb healthy speech to simulate dysarthric speech has continued [17].

Despite these efforts, traditional machine learning models struggle with data variability which is inherent in dysarthric speech, especially in cases with severe dysarthria or aphasia. As the world moved on from traditional ML methods to deep learning, general ASR models showed a huge jump in accuracy which naturally led to research on using deep learning for dysarthric speech recognition. Research has involved incorporating articulatory information into deep neural networks which has better captured the variability in dysarthric speech.

Techniques like Kullback-Leibler (KL)-HMM that model state emission probabilities based on DNN-

estimated posterior probabilities and speaker adaptation methods using regularization represent more accurate jumps in ASR systems for dysarthric speech [18]. One method for this was through employing non-linear approaches to modify speech rhythm to bridge the gap between typical and atypical speech. Text-to-speech (TTS) systems could generate synthetic dysarthric speech, and could incorporate factors like dysarthria severity [19]. Techniques like virtual microphone array synthesis and multiresolution feature extraction (VM-MRFE) were applied, along with exploring prosodic transformation and time-feature masking [20].

While these methods helped solve the problem of scarcity, they often resulted in a depletion of quality, with discussions on whether methods like auto-encoders or TTS systems can accurately and adequately represent real-world variability. Simulated dysarthric speech also lacked dataset diversity, which is critical for the training of speaker-independent models [21].

Traditional ASR systems rely on identifying phoneme sequences which aim to map speech signals to basic units of language first. However, in dysarthric speech these phonemes can be largely imprecise. So, researchers have also experimented with extracting voice grams instead of spectrograms which offer a potential advantage, by shifting the focus from discrete phoneme recognition to holistic visual representation of entire words. This approach bypasses the need for precise phoneme labelling, which is challenging and often inaccurate for dysarthric speech. By learning to recognize the visual "shape" of words, speech vision becomes more robust to the altered and imprecise phonemes characteristic of dysarthria [22].

More recently, advancements in transformer-based models have allowed for further research in dysarthric ASR. Encoder models such as HuBERT have been used as the feature backbone for dysarthric ASR. One method employs a feature extractor trained with HuBERT to produce per-word prototypes that encapsulate the characteristics of previously unseen speakers. By enhancing representation quality, dysarthric speech recognition performance can be further improved [33]. Furthermore, architectures such as whisper have been used as end-to-end models for dysarthric ASR. Whisper is a supervised, end-to-end encoder–decoder Transformer trained on hundreds of thousands of hours of labelled audio for ASR. Most notably, “CBA-Whisper” [34] fine-tunes Whisper for dysarthric ASR through a curriculum- learning schedule. The system integrates WhisperX-style preprocessing for long-form segmentation/alignment and rule-based postprocessing to mitigate stuttering and hallucinations.

Though research in the field has been commendable, most dysarthric models that have been implemented so far are speaker- dependent. This means that for each dysarthric speaker, one would need to train a different model for it to be effective. Speaker-independent models are challenging to implement since dysarthric speech is typically associated with high variability in both phonematic and articulatory information. Unlike healthy speech, where it is relatively easy to detect speech patterns, dysarthric speakers have a different way of enunciating words depending on which muscle best allows them to articulate their words in an understandable manner [23].

This produces very different speech patterns for different speakers, even if they are trying to say the same word or sentence. This gap can effectively be addressed through our proposed solution of implementing a speaker independent model using the wav2vec 2.0 model for feature extraction along with 3 layered, Bi-Directional LSTM to map the features to text. This would effectively prevent the need of extensive individualised training data, which can not only be hard to procure, but infeasible for a majority of the population.

Methodology

To answer the question of how we can develop a model that is able to generalize wav2vec 2.0 features for a speaker independent ASR system, we propose the following machine learning architecture. The novel architecture used aims to build on previously successful speaker-dependent model architecture

while pushing forward the development of a speaker-independent model.

Dataset

The TORGO Database includes aligned acoustic and articulatory recording from 8 individuals with dysarthria caused by Cerebral Palsy or Amyotrophic Lateral Sclerosis along with 7 control speakers without any disorders [1]. Each dysarthric speaker had their degree of severity of the disorder evaluated by a speech-language pathologist in terms of clinical intelligibility and motor functions of the articulators per the Frenchay dysarthria assessment [24]. Over multiple sessions, each speaker recorded over 3 hours of speech which included Non Words - used to control for the baseline abilities of the dysarthric speakers, Short Words - useful for studying speech acoustics without the need for word boundary detection, and Restricted Sentences - to utilize lexical, syntactic, and semantic processing in ASR. The stimuli used in the dataset was sourced from various different materials including the TIMIT database [25] and more. Acoustic data in the TORGO Database was collected using both a head mounted and a directional, array microphone. The collection of movement data and time aligned acoustic data was carried out using the 3D AG500 electro-magnetic articulograph (EMA) with fully automated calibration which allowed for 3d recordings of articulatory movements inside and outside the vocal tract, providing a detailed window on the nature and direction of speech related activity.

The motor functions of each speaker in the TORGO dataset were evaluated using the standardized Frenchay Dysarthria Assessment (FDA) [32] by a speech- language pathologist. This assessment is designed to diagnose individuals with dysarthria while being applicable to therapy. FDA measures 28 relevant dimensions of speech grouped into 8 categories, namely reflex, respiration, lips, jaw, soft palate, laryngeal, tongue, and intelligibility. FDA allows the evaluation of the nature and severity of dysarthria by providing a detailed and objective measure of how dysarthria has impacted a person's ability to speak. LibriSpeech ASR Corpus or LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech, prepared by Vassil Panayotov. The dataset comprises of approximately 2484 unique speakers with 1283 of them being female, and 1201 male speakers. LibriSpeech also includes n-gram language models and corresponding texts excerpted from Project Guntenberg books, containing 803 million tokens and 977,000 unique words [30].

Feature Extraction

Dysarthric speech has a considerable amount of variation not just between different speakers but also in a particular speaker. To overcome this, high quality feature representations of human speech need to be extracted from the audio. In the past, GMM-HMM models used Mel- Frequency Cepstral Coefficients (MFCCs) [26]. However, these features aren't able to represent human speech well enough to develop speaker-independent models and hence, individual models were trained for each speaker. To overcome this, we use the wav2vec 2.0 model for feature extraction. This model, developed by Meta AI, excels in extracting robust speech representations, which are crucial for tasks like ASR, especially in low-resource settings [27].

The wav2vec 2.0 model is composed of a multi-layer convolutional feature encoder, a transformer and a quantisation module.

The feature encoder $f : X \mapsto Z$ takes as input raw audio X and outputs latent speech representations $Z = (\mathbf{z}1, \ldots, \mathbf{z}T)$ for T time-steps. It consists of blocks containing a temporal convolution, followed by layer normalization and a Gaussian Error Linear Unit (GELU) activation function to make the transformation non-linear and normal.

The Transformer $g : Z \mapsto C$ the takes these latent speech representations to build representations $C = (\mathbf{c}1, \ldots, \mathbf{c}T)$ capturing information from the entire sequence. It builds context representations over the speech input and self-attention captures dependencies over the sequence of latent representations, which enhances the models understanding of the speech context.

The quantization module $Z \mapsto Q$ takes the output of the feature encoder and discretizes it to $\mathbf{q}t$ to represent the targets (Figure 1) in the self-supervised objective.

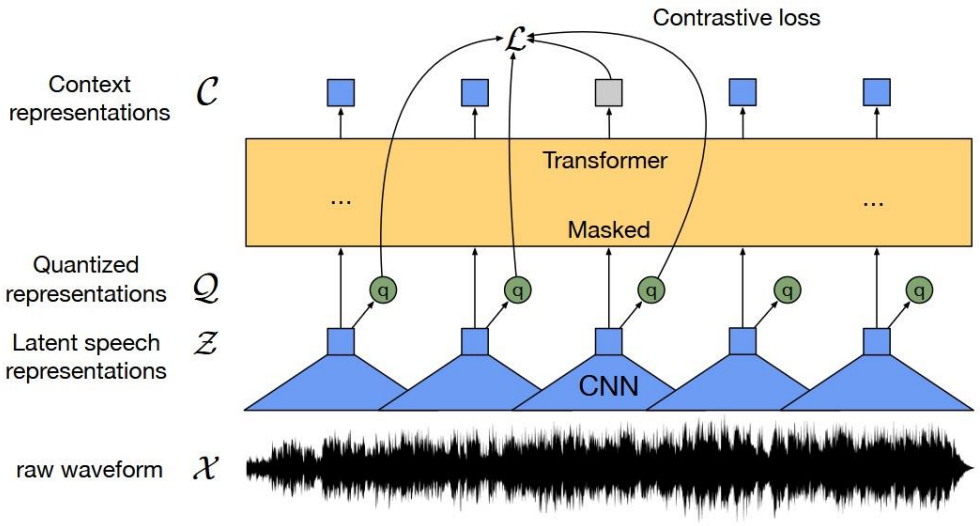

Figure 1: Overview of wav2vec 2.0 architecture. Reprinted from [3]

Data Pre-processing

Various types of pre-processing methods were experimented with such as silence removal, noise removal, and both. By removing the silence between words wav2vec 2.0 was not able to accurately represent the natural pauses and gaps humans make while speaking, resulting in higher error rates while training the model. Through further testing, the best pre- processing method found was cutting the silence at the beginning and end of the audio file before passing it on to the wav2vec 2.0 model. This resulted in more accurate speech representations due to the natural gap in speech being retained and no loss of speech information that occurs while running any noise reducing algorithm such as spectral subtraction.

All the audio labels or transcripts were pre- processed by first converting all the text to lowercase and removing all characters that weren't letters or blank spaces. All blank spaces at the beginning or end of the transcript were also removed. Any datapoint with a transcript that referenced an image was completely removed from the dataset since the speaker isn't uttering the image name and instead a description of the image.

The data was then split into train and test sets with 80% of the data in the train set of which 25% is used as the validation set. The speakers within this set were randomized, and not picked specifically.

LSTM

The wav2vec 2.0 model first converts each audio file into 1024 dimensional vectors which are padded with zeros to match the length of the longest vector sequence so that each sequence is then the same length. These vectors are then mapped to the text by first passing through an input projection layer which projects the 1024 dimensional vectors into 256 dimensions while applying normalization and dropout with a ReLU activation function.

These vectors are then passed into a three layered, bi-directional LSTM to process information in both past-to-future and future-to-past directions. The first layer converts the 256 dimensional input into 512 dimensions and each of the subsequent layers keep it at 512 dimensions. Residuals are added for skip connections between LSTM layers to help with gradient flow.

The attention layer then computes attention weights for each time step which are passed through a

softmax activation function that help the model focus on the most relevant parts of the sequence which is particularly helpful for dysarthric speech due to its irregular timing patterns.

The LSTM output is passed on to the linear layers, which for each timestep computes character probabilities for 28 characters which include 26 lowercase letters, blank and a blank space for the Connectionist Temporal Classification (CTC) Loss.

Experiment

All speech from the TORGO dataset and LibriSpeech was first passed the wav2vec 2.0 model for feature extraction, where they were converted to 1024-d vectors. The Bi- directional LSTM model was then trained on the LibriSpeech data which produced a WER of around 1.3%. The model was further fine-tuned on dysarthric speech data from TORGO which reached an overall WER of about 48% and CER of about 27%.

The model is trained using a CTC loss function for 30 epochs. After each epoch the validation loss is calculated to ensure the model doesn't overfit.

While training, early stopping is used if the validation loss doesn't decrease for 5 straight epochs in which case, the model with the best validation loss is used to evaluate the test set.

Around 22 different models were trained with different hyperparameters and the one with the best validation loss was finalised. The hyperparameters chosen are given in Table 1.

Batch Size	Learning Rate	Patience	Early-Stopping delta	Dropout Rate
64	5.00E-04	5	0.001	0.3

Table 1: Final chosen hyperparameters

Results

The evaluation of the model was done using both the WER and CER. WER takes into account three types of errors: Substitutions (S), where a word is replaced by another; Deletions (D), where a word in the reference is missed; and Insertions (I), where an extra word appears in the transcription and is given by the formula [28]:

$$WER = (S + I + D) / Number\ of\ Words\ Spoken$$

While, CER was used to provide a finer view of the model's performance, similar to WER, CER is calculated as [29]:

$$CER = (S + I + D) / Number\ of\ Characters$$

The model achieved an overall WER of 50.22% and CER of 30.48%. Examples of the reference transcript and model predictions are given below:

Example 1 (Speaker: M03):

Reference: 'slip'

Prediction: 'slip

Example 2 (Speaker: F03):

Reference: 'ride'

Prediction: 'side'

Example 3 (Speaker: FC03): Reference: 'the hotel owner shrugged'

Prediction: ‘the hotel owner shrugged’

Compared to previous speaker independent wav2vec 2.0 based models for dysarthric speakers such as by Su, 2024 [27]. This model achieves a similar WER and CER without any pre training required. This significantly cuts down on training time for the LSTM and the wav2vec 2.0 inferencing time for feature extraction. The overall and speaker specific error rates are given in Tables 2, 3, and 4.

Speaker	WER (%)	CER (%)
Overall	48.75	26.62

Table 2: Overall WER and CER

Speaker	WER (%)	CER (%)
F01	93.91	60.14
F03	72.25	43.22
F04	39.38	18.4
M01	100	63.58
M02	94.91	66.62
M03	34.95	18.04
M04	92.87	62.5
M05	99.12	72.22

Table 3: WER and CER for Dysarthric Speakers

Speaker	WER (%)	CER (%)
FC01	53.47	30.69
FC02	42.25	17.76
FC03	46.35	26.77
MC01	39.01	17.18
MC02	46.66	24.14
MC03	26.75	10.96
MC04	32.48	15.68

Table 4: WER and CER for Control Speakers

Conclusion

In this paper, we explore a new architectural approach to speaker independent models for dysarthric speakers. The LSTM architecture built is trained on 1024 dimensional feature representations of dysarthric and control speech. The proposed methodology includes extracting high level features using the wav2vec 2.0 model and training a 3 layered Bi-Directional LSTM to map the vectors to speech using the CTC loss function. The results show that building a speaker independent model using wav2vec 2.0 extracted features and a pre-trained LSTM can help lower WERs and CERs without much pre-training required. This study has potential limitations. Due to high strain on their speech muscles, speech data from dysarthric speakers is limited due to which speaker-independent models aren't able to generalize well to the whole set of dysarthric speakers. Future developments for the model will include pre-training on a large corpus of healthy speech and then fine tuning it on the TORGO dataset to increase vocabulary and further decrease the model's WER and CER.

Data Availability Statement

Results for this model and the source code used can be found on github under the following link: https://github.com/xshxn/mlpr-project

References

1. Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2011b). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation, 46(4), 523–541. https://doi.org/10.1007/s10579-011-9145-0
2. Dieck, T. T., Pérez-Toro, P. A., Arias, T., Noeth, E., & Klumpp, P. (2022). Wav2vec behind the Scenes: How end2end Models learn Phonetics. Interspeech 2022, 5130–5134. https://doi.org/10.21437/interspeech.2022-10865
3. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Neural Information Processing Systems, 33, 12449–12460 https://doi.org/10.48550/arXiv.2006.11477
4. Enderby, P. (2013). Disorders of communication. Handbook of Clinical Neurology, 273–281. https://doi.org/10.1016/b978-0-444-52901-5.00022-8
5. Hartelius, L., Elmberg, M., Holm, R., Lövberg, A., & Nikolaidis, S. (2007). Living with Dysarthria: Evaluation of a Self-Report Questionnaire. Folia Phoniatrica Et Logopaedica, 60(1), 11–19. https://doi.org/10.1159/000111799
6. Kent, R. D., & Rosen, K. (2004b). Motor Control Perspectives on Motor Speech Disorders. In Oxford University Press eBooks (pp. 285–311). https://doi.org/10.1093/oso/9780198526261.003.0012
7. Raghavendra, P., Rosengren, E., & Hunnicutt, S. (2001b). An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. Augmentative and Alternative Communication, Vol-17(4), 265–275. https://doi.org/10.1080/aac.17.4.265.275
8. Shahamiri, S. R. (2021b). Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29, 852–861. https://doi.org/10.1109/tnsre.2021.3076778
9. Vinotha, R., Hepsiba, D., Anand, L. D. V., Andrew, J., & Eunice, R. J. (2024). Enhancing dysarthric speech recognition through SepFormer and hierarchical attention network models with multistage transfer learning. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-80764-w

10. Yu, C., Su, X., & Qian, Z. (2023). Multi-Stage Audio-Visual fusion for dysarthric speech recognition with Pre-Trained models. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 1912–1921. https://doi.org/10.1109/tnsre.2023.3262001

11. Chen, X., Wang, Y., Wu, X., Wang, D.,Wu, Z., Liu, X., & Meng, H. (2024). Exploiting Audio-Visual Features with Pretrained AV-HuBERT for Multi-Modal Dysarthric Speech Reconstruction. arXiv (Cornell University). https://doi.org/10.48550/arXiv.2402.01234

12. Jayaram, G., & Abdelhamied, K. (1995). Experiments in dysarthric speech recognition using artificial neural networks. Journal of rehabilitation research and development, 32(2), 162–169. https://pubmed.ncbi.nlm.nih.gov/7602300/

13. Wilson, B. B. J. (2000). Acoustic variability in dysarthria and computer speech recognition. Clinical Linguistics & Phonetics, 14(4), 307–327. https://doi.org/10.1080/02699200050024001

14. Polur, P., & Miller, G. (2005). Experiments with fast Fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 13(4), 558–561. https://doi.org/10.1109/tnsre.2005.856074

15. Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M., & Parker, M. (2003). Automatic Speech recognition with sparse training data for dysarthric speakers. EUROSPEECH 2003 - INTERSPEECH 2003, 1189–1192. https://doi.org/10.21437/eurospeech.2003-384

16. Hasegawa-Johnson, M., Gunderson, J.,Perlman, A., & Huang, T. (2006). Hmm-Based and Svm-Based Recognition of the Speech of Talkers With Spastic Dysarthria. International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp.2006.1660840

17. Vachhani, B., Bhat, C., & Kopparapu, S. K. (2018). Data augmentation using healthy speech for dysarthric speech recognition. Interspeech 2022. https://doi.org/10.21437/interspeech.2018-1751

18. Kim, M., Kim, Y., Yoo, J., Wang, J., & Kim, H. (2017). Regularized speaker adaptation of KL-HMM for Dysarthric Speech Recognition. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25(9), 1581–1591. https://doi.org/10.1109/tnsre.2017.2681691

19. Soleymanpour, M., Johnson, M. T., Soleymanpour, R., & Berry, J. (2022). Synthesizing dysarthric speech using multi-talker TTS for dysarthric speech recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.11571

20. Soleymanpour, M., Johnson, M. T., & Berry, J. (2021). Dysarthric Speech Augmentation Using Prosodic Transformation and Masking for Subword End-to-end ASR. International Conference on Speech Technology and Human-Computer Dialogue, SpeD, 42–46. https://doi.org/10.1109/sped53181.2021.9587372

21. Soleymanpour, M., Johnson, M. T., Soleymanpour, R., & Berry, J. (2024). Accurate synthesis of dysarthric Speech for ASR data augmentation. Speech Communication, 164(103112). https://doi.org/10.1016/j.specom.2024.103112

22. Shahamiri, S. R. (2021). Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29, 852–861. https://doi.org/10.1109/TNSRE.2021.3076778

23. Rowe, H. P., Gutz, S. E., Maffei, M. F., Tomanek, K., & Green, J. R. (2022). Characterizing Dysarthria Diversity for Automatic Speech Recognition: A Tutorial From the Clinical

Perspective. Frontiers in Computer Science, 4. https://doi.org/10.3389/fcomp.2022.770210

24. Enderby, P. (1980). Frenchay dysarthria assessment. International Journal of Language & Communication Disorders, 15(3), 165–173. https://doi.org/10.3109/13682828009112541
25. Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., & Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus [Dataset]. https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/SWVENO
26. Joy, N. M., & Umesh, S. (2018). Improving acoustic models in TORGO Dysarthric Speech Database. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26(3), 637–645. https://doi.org/10.1109/tnsre.2018.2802914
27. Enhancing English Dysarthric Speech Recognition with Age-Matched Healthy Speech: A Fine-Tuning Approach Using wav2vec 2.0 - Studenttheses Campus Fryslan. (n.d.). Retrieved May 2, 2025, from https://campus-fryslan.studenttheses.ub.rug.nl/543
28. Park, Y., Patwardhan, S., Visweswariah, K., & Gates, S. C. (2008). An empirical analysis of word error rate and keyword error rate. Interspeech 2022. https://doi.org/10.21437/interspeech.2008-537
29. K, T. D., James, J., Gopinath, D. P., & K, M. A. (2024). Advocating character error rate for multilingual ASR evaluation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2410.07400
30. V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp.5206-5210 https://doi.org/10.1109/ICASSP.2015.7178964
31. Enderby, Pamela. (2011). The Frenchay Dysarthria Assessment. International Journal of Language & Communication Disorders. 15. 165 - 173.10.3109/13682828009112541. https://doi.org/10.3109/13682828009112541
32. Wang, Shiyao & Zhao, Shiwan & Zhou, Jiaming & Kong, Aobo & Qin, Yong. (2024). Enhancing Dysarthric Speech Recognition for Unseen Speakers via Prototype-Based Adaptation. 1305-1309.10.21437/Interspeech.2024-1360. https://doi.org/10.21437/Interspeech.2024-1360
33. Tan, T., Chen, X., Le, X., Fan, W., Xia, X., Huang, C., Lu, J. (2025) CBA-Whisper: Curriculum Learning-Based AdaLoRA Fine-Tuning on Whisper for Low-Resource Dysarthric Speech Recognition. Proc. Interspeech 2025,3309-3313, doi:10.21437/Interspeech.2025-1705 https://doi.org/10.21437/Interspeech.2025-1705