Dysarthric Speech Transformer: A Sequence-to-Sequence Dysarthric Speech Recognition System

Seyed Reza Shahamiri, *Senior Member, IEEE*, Vanshika Lal, and Dhvani Shah

***Abstract*— Automatic Speech Recognition (ASR) technologies can be life-changing for individuals who suffer from dysarthria, a speech impairment that affects articulatory muscles and results in incomprehensive speech. Nevertheless, the performance of the current dysarthric ASR systems is unsatisfactory, especially for speakers with severe dysarthria who most benefit from this technology. While transformer and neural attention-base sequences-to-sequence ASR systems achieved state-of-the-art results in converting healthy speech to text, their applications as a Dysarthric ASR remain unexplored due to the complexities of dysarthric speech and the lack of extensive training data. In this study, we addressed this gap and proposed our Dysarthric Speech Transformer that uses a customized deep transformer architecture. To deal with the data scarcity problem, we designed a two-phase transfer learning pipeline to leverage healthy speech, investigated neural freezing configurations, and utilized audio data augmentation. Overall, we trained 45 speaker-adaptive dysarthric ASR in our investigations. Results indicate the effectiveness of the transfer learning pipeline and data augmentation, and emphasize the significance of deeper transformer architectures. The proposed ASR outperformed the state-of-the-art and delivered better accuracies for 73% of the dysarthric subjects whose speech samples were employed in this study, in which up to 23% of improvements were achieved.**

***Index Terms*— Dysarthria, dysarthric speech recognition, deep learning, transformers.**

I. Introduction

DYSARTHRIA occurs when the muscles responsible for articulation become weak or difficult to control. The impairment typically results in speech that is difficult to understand, and it is commonly caused by nervous system anomalies and illnesses that produce facial and articulatory muscle paralysis. Depending on the severity and underlying cause, the signs and symptoms of dysarthria may be slurred or slow speech, inability to whisper, rapid speech, or monotone speech [1]. As the impairment progresses, speech produced by dysarthric individuals becomes unintelligible due to the intensive muscle paralysis making phone production different from normal (aka healthy) speakers [2]. Hence, dysarthric individuals struggle to communicate with others.

People with dysarthria and other speech disorders can benefit from Automatic Speech Recognition systems (ASR) since ASR can enable computers to hear them and talk on their behalf. The technology can be life-changing for people suffering from severe dysarthria as computers can talk on their behalf and enable them to interact with digital devices [3]. ASR is the technique by which a computer recognizes spoken language or utterances. Recent advancements in ASR technologies have resulted in the widespread usage of ASR systems in various devices, including smartphones and smart home devices, to provide an automated assistant system that can accurately transcribe spoken words. However, even the best-performing ASR systems are ineffective for speakers with the speech impairment, especially those who could benefit from the technology most, such as severe dysarthria sufferers. Lee et al. [4] indicated that the deficiency might be due to the difficulties of getting sufficiently diversified impaired speech training samples. From a clinical standpoint, they uncovered how having variety might affect ASR performance, which might aid in developing a generalized system that difficult-to-recognize speakers can use.

Thus, automated recognition of dysarthric speech remains a challenge because of the features associated with the impairment. The irregularities of dysarthric speech negatively impact phone generation and articulation, resulting in high complexities in automatically processing and recognizing dysarthric speech. For example, the significant variations of dysarthric speech require modern ASR technologies to process a large amount of the impaired speech from many speakers to capture inter-speaker variability. Nonetheless, the availability of public dysarthric speech samples is very limited since it is difficult to capture a large amount of speech from such individuals due to muscle fatigue caused by the impairment [5]. As such, the scarcity of dysarthric speech samples is one of the major challenges preventing the successful development of dysarthric speech recognition systems.

Amongst the most successful ASR methods are Transformers and neural self-attention mechanisms [6] that have resulted in state-of-the-art ASR achievements. Transformers

Manuscript received 20 April 2023; revised 13 August 2023; accepted 17 August 2023. Date of publication 21 August 2023; date of current version 29 August 2023. *(Corresponding author: Seyed Reza Shahamiri.)*

The authors are with the Department of Electrical, Computer, and Software Engineering, Faculty of Engineering, The University of Auckland, Auckland 1010, New Zealand (e-mail: admin@rezanet.com; vlal080@aucklanduni.ac.nz; dsha439@aucklanduni.ac.nz).

Digital Object Identifier 10.1109/TNSRE.2023.3307020

have proven highly effective for sequence-to-sequence transduction tasks because neural attention modulates token representations using the interpretations of correlated tokens in the sequence and increasing the learning effectiveness of long-range dependencies. Nevertheless, end-to-end transformer self-attention-based ASR systems are yet to be properly investigated to recognize dysarthric speech since such architectures often require a significant amount of training data that are not publicly available for dysarthric speech.

To address this gap, we propose our Dysarthric Speech Transformer (DST), a speaker-adaptive, end-to-end dysarthric ASR that utilizes state-of-the-art transformer architectures composed of multiple encoder and decoder modules. To address the scarcity of dysarthric data and be able to train such deep architectures, we designed a training pipeline to apply two phases of transfer learning by utilizing two speech corpora. Two steps of neural freezing and parameter adjustments enabled the DST to learn and map healthy speech signals to character sequences before fine-tuning to recognize dysarthric speech. The proposed pipeline also applies audio data augmentation techniques to enhance the limited availability of dysarthric speech samples. The DST has been verified via speech samples collected from multiple dysarthric speakers with different speech intelligibility levels and compared in detail with state-of-the-art dysarthric speech recognition systems evaluated using the same speakers' utterances.

The contributions of this study can be summarized as follows:

1. The proposition of a seq-to-seq transformer ASR tailored for dysarthric speech,
2. Transformer architectural selection and configuration for dysarthric speech recognition,
3. Investigating the effects of depths over performance,
4. Tackling the dysarthric data scarcity issue by designing a two-phase transfer learning and neural freezing, and investigating the best transfer learning architectural configurations,
5. Studying the effects of audio data augmentation on the proposed Dysarthric Speech Transformer's performance, and
6. Detailed per-speaker performance comparison with the state-of-the-art.

The rest of this paper is organized as follows. The next section briefly explains transformers in the speech recognition context. The research methodology is explained in the following section, in which the proposed DST, the transfer learning pipeline, datasets, and data augmentation are described. The third section provides further information on the experimental setup, followed by the results and discussion sections. This section also presents the comparative and benchmarking study. The paper concludes with the conclusion section and recommendations for future studies.

II. Sequence-to-Sequence ASR Systems and Transformers

Seq-to-seq ASR is often referred to as the speech recognition approach that uses deep learning-based encoder-decoder models that map sequences of speech frames to sequences of characters. Both encoder and decoder modules are trained together with the same loss function, in contrast to traditional ASR models, where the acoustic, language, and pronunciation models were usually trained separately, each with a separate loss function. Seq-to-seq models disregard the frame-independence conjecture made by Hidden Markov Models and Connectionist Temporal Classification. This means their language model is implicit, and they can optimize error rates more efficiently, resulting in better overall ASR performance. The encoder module in seq-to-seq ASR converts the speech frames presented either via traditional acoustic features extraction methods or visually as spectrograms (aka voicegrams) to hidden representations. The decoder then converts these representations to a character sequence, character by character.

Initial seq-to-seq ASR systems were commonly built using Recurrent Neural Networks (RNNs). An example is Bahdanau et al. [7], who employed a deep bi-directional RNN to encode the speech signal into a suitable feature representation, and an attention-based Recurrent Sequence Generator RNN to decode this representation into a sequence of characters.

Nonetheless, using RNNs as the primary algorithm imposes limitations that prevent seq-to-seq ASR systems from unlocking their full potential [8]. While RNNs work well for short statements and prompt, their ability to learn larger contexts are limited as they have a shorter window to reference from. Additionally, RNNs' sequential nature makes them slow to train. To overcome these limitations, Transformers and self-attention were proposed by Vaswani et al. [6], [9], in which the attention network intuitively learns to pay attention to important features and ignores the rest, making the features context-aware. With self-attention, the network can generate representations for characters based on other characters surrounding it, modulating token representations. Attention acts as an interface between the encoder and decoder to provide the decoder with information from the decoder's hidden states. With multi-head attention, the attention operation can be done multiple times for each attention layer. Transformers are deep neural networks that leverage the self-attention concept, commonly used in modern seq-to-seq tasks [10]. A study compared transformers and RNNs in text-to-speech context and observed a larger minibatch resulted in better validation L1 loss [12] for transformers with faster training but reported a negative influence on the L1 loss for RNNs [11].

Transformer and attention-based ASR was first introduced by Dong et al. [13], where a 2-D attention mechanism was proposed and evaluated on the Wall Street Journal normal speech corpus. This study culminated with significantly lower training costs and achieved an excellent Word Error Rate (WER) demonstrating the Speech Transformer's efficiency and efficacy. Since their introduction, transformers have been utilized in various ASR systems.

With respect to seq-to-seq dysarthric ASR, Google's Project Euphonia researchers employed a Recurrent Neural Network Transducers (RNN-T) [14] architecture composed of an

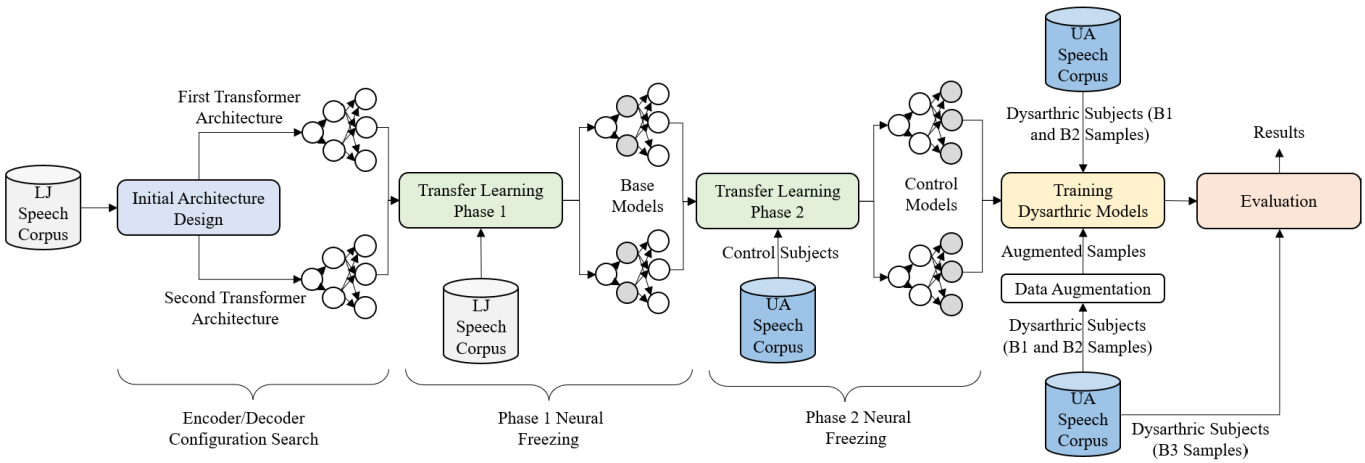

Fig. 1. Methodology overview.

encoder with eight Long Short-Term Memory (LSTM) layers and a language model with two LSTM layers. This ASR was evaluated on the Euphonia corpus [15], which includes 294 dysarthria participants among other speech-impaired subjects. While the authors did not report dysarthric-specific results, they indicated that dysarthric speech was particularly difficult to model and hence was classified among high WER subsets.

Despite the success of transformer normal speech ASR systems, their application for dysarthric speech is yet to be properly investigated. Since training transformer ASRs requires large datasets containing many labeled speech samples, the training pipeline and architecture design process need to carefully utilize the limited dysarthric speech data available, exploring all available venues to transfer knowledge across datasets and augment the available data. The following section explains our Dysarthric Speech Transformer and training pipeline to fill this gap and reports how we overcame the challenges mentioned before.

III. Methodology

Figure 1 portrays the overall methodology and the training pipeline. We began conducting experiments with healthy speech to design the initial transformer architecture. We experimented with various attention-based transformer encoder and decoder architectures to identify the optimal base model structures. From these trials, we chose the two transformer architectures presented in section III-B as the base models since they demonstrated superior performance compared to the others. These base models were first trained on healthy connected speech in the first phase of transfer learning. Next, we investigated the best neural freezing [16] architectures that best preserved the knowledge acquired from healthy speech to fine-tune the models as normal-speech control models. Once the control models were trained, we initiated the second phase of neural freezing and investigated the best configurations that retained the previous base and control knowledge but unlocked enough synaptic weights to learn dysarthric speaker-specific acoustic features. We also applied audio data augmentation to reinforce the limited dysarthric speech available, then trained the control models for each dysarthric participant independently. In particular, we refer to the initial models, trained on connected, healthy speech, as 'base models'. Subsequently, the base models trained on control subjects supplied by the dysarthric corpus are termed 'control models', while the control models trained on dysarthric subjects are designated 'dysarthric models'.

Our experiments were concluded by evaluating both transformer architectures per dysarthric participant, including how the model performed with and without the augmented data, and comparing both transformer architectures with each other and the state-of-the-art. The rest of this section explains this process in detail.

A. Materials and Participants

There are very few speech corpora available publicly that include dysarthric speech samples. Among them are Nemours [17], TORGO [18], and UA-Speech [19]. Google also has the Euphonia dataset [15] reported to have dysarthric samples, but this dataset was not publicly available at the time this study was conducted, and our request to access it was unsuccessful. Among the remaining datasets, UA-Speech is the largest, has more dysarthric participants, and has been the most widely used corpus in the literature for dysarthric speech recognition research [31]. As such, it was used in this study.

UA-Speech was developed by the University of Illinois researchers and features speech samples obtained from 19 dysarthric individuals with speech intelligibility levels ranging from 2% to 95%. The intelligibility levels of dysarthric speakers range from very low (0-25%) intelligibility to low (25-50%), mild (50-75%), and high (75-100%) intelligibility. Speech intelligibility can be defined as to what extent speech is comprehensible by a typical listener and is one of the mechanisms to define the severity of dysarthria [20]. The corpus overall provides speech samples collected from 28 speakers, including 15 dysarthric speakers and 13 healthy control speakers – the speech samples of the other four dysarthric participants are not publicly available (speakers M02, M03, F01, M06).

TABLE I
PARTICIPANTS

Number	Participants	Sex	Age	Speech Intelligibility (%)	Intelligibility Level
1	M04	Male	>18	2	Very Low
2	F03	Female	51	6	
3	M12	Male	19	7	
4	M01	Male	>18	17	
5	M07	Male	58	28	Low
6	F02	Female	30	29	
7	M16	Male	40	43	
8	M05	Male	21	58	Mild
9	M11	Male	48	62	
10	F04	Female	18	62	
11	M09	Male	18	86	High
12	M14	Male	44	90	
13	M10	Male	21	93	
14	M08	Male	28	95	
15	F05	Female	22	95	
16-27	UA-Speech Control Speakers	Four female and seven male	Not Provided	Not Applicable	Healthy Speech
28	LJ Speech Speaker	Male	Not Provided	Not Applicable	Healthy Speech

UA-Speech data is subdivided into two subsets per speaker of common and uncommon words, where the uncommon words are different per speaker. The common word samples are divided into B1, B2, and B3 blocks and are the same for all speakers. Each block provides utterances recorded in different sessions.

Here, we used the 155 common words to build and evaluate the dysarthric models and trained a separate model per dysarthric speaker. We used B1 and B2 utterances for training, but B3 samples were withheld and only used for testing to ensure they were unforeseen by the models. Each speaker model was trained using audio samples from blocks B1 and B2 and then tested using audio samples from block B3. The vocabulary comprised ten numerals, 19 computer instructions, 26 radio alphabets, and one hundred common words.

Additionally, to enable the two-phase transfer learning and neural freezing, we employed normal speech samples from eleven UA-Speech control subjects and the LJ Speech Dataset [21]. LJ Speech is a connected, normal speech corpus that comprises around 24 hours of labeled audio data. It includes 13,100 short audio recordings of a single speaker reading passages from seven non-fiction books. The total number of words is 225,715, with 13,821 distinct words. Each audio file is a single-channel 16-bit PCM WAV with a sampling rate of 22 KHz. However, LJ Speech utterances were resampled to 16KHz to make them consistent with UA-Speech samples. Table I provides the participants' information.

B. The Proposed Dysarthric Speech Transformer Architectures

In this study, we experimented with two transformer architectures selected during our initial architecture design experiments. Inspired by [6] and [13], the architecture of the first transformer model is presented in Fig.2, comprising four encoder modules and one decoder. Dysarthric speech samples were provided to the model via voicegrams (aka spectrograms) with 200ms frames sliding 80ms while Fast Fourier Transform of size 256 was applied. Before the voicegrams were provided to the first encoder, down-sampling via three convolutional layers was applied to utilize voicegrams' structural locality. Each convolution layer applied 64 filters of 1×11 dimension with two strides.

The speech features were then provided to the encoder components to provide hidden representations that the decoder would use. The encoder employed a multi-head attention layer with two heads followed by a dropout layer. Before the output was given to the feed-forward network, a residual connection was used to re-insert the initial encoder input into the data stream, and a normalization layer was applied. Instead of batch normalization, the transformers used layer normalization that normalized each input voicegram independently of others due to the sequential nature of speech data. Finally, the feed-forward network processed the data, applied dropout and another layer normalization, and then passed the encoded output to the next encoder module. The transformer was composed of four encoders stacked on top of each other. However, only the first encoder received the initial speech features, and the following encoders received the encoded output of the previous encoder.

The decoder was composed of three primary components. The initial attention and the feed-forward blocks were similar to the encoder architecture, but a middle attention component was added that also received the hidden representations, the output of the last encoder. The encoder received the text corresponding to the given dysarthric speech sample, vectorized and character position information inserted. Nonetheless, the input text was masked to include only the first N characters, but the last character was offset to force the model to learn to predict the $N+1$ character. The decoder used attention layers to recognize which tokens in the hidden representations supplied by the encoders most likely correspond to the text token it was attempting to estimate. Finally, the decoder output was given to a dense output layer with softmax activation to produce the next character probabilities. The most probable character was found by applying an argmax function to the softmax result.

During the inception phase, the initial input to the decoder was an empty text with "[start]" token. Then, after the first character was predicted, it was added to the initial text and fed back to the decoder to predict the next character. This process continued until the predicted character was token "[end]".

For the second transformer, a deeper encoder architecture depicted in Figure 3 was designed. Compared to the previous encoder, the second transformer's encoder was modified to 1) add a second attention block and 2) replace the feed-forward network block with two Depthwise Separable Convolution blocks to add more depth yet promote faster training. Additionally, while the decoder architecture remained the same as in Figure 2, we increased the number of encoder blocks to five and decoders to three. The rest of the configurations remained unchanged.

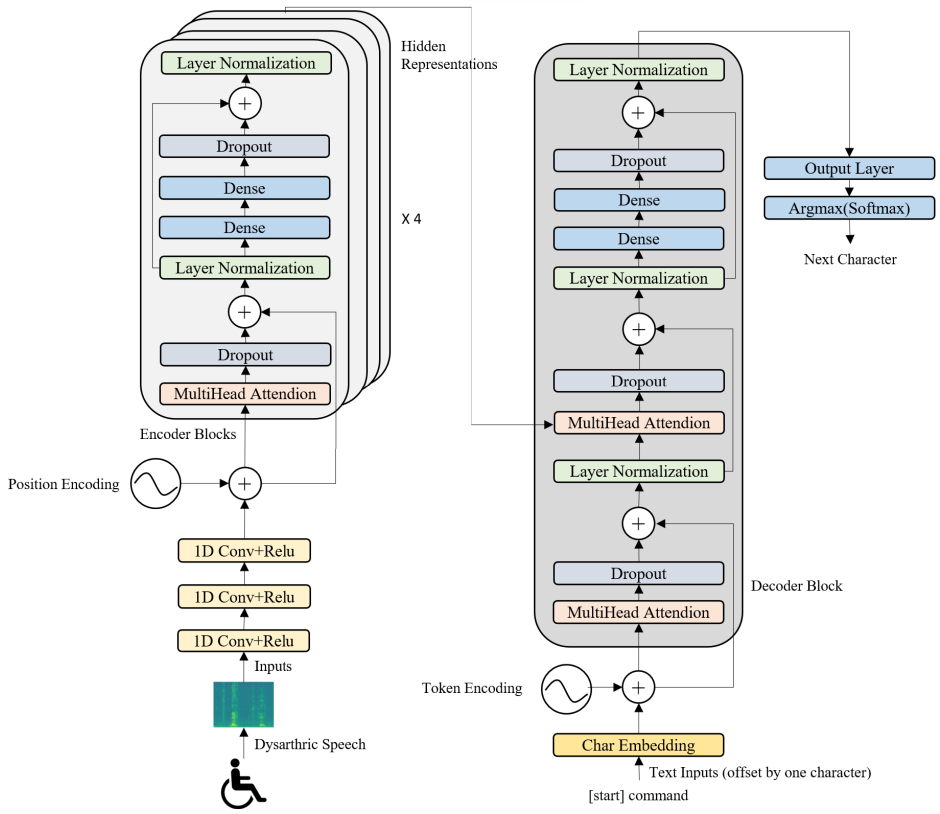

Fig. 2. The first transformer architecture.

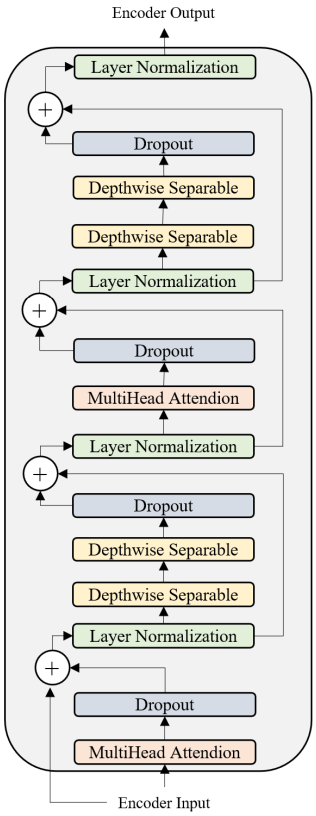

Fig. 3. The modified encoder used in the second transformer.

In order to identify the best-performing architectures and hyperparameters, LJ Speech samples were divided into 80% training and 20% validation, and Bayesian Optimization tuning with the Gaussian process [22] was used. Amongst the hyperparameters that were trailed and adjusted during this process were the number of hidden layers, number of attention heads, number of feed-forward blocks, number of encoder and decoder layers, etc. The presented architectures achieved the best initial performances on LJ Speech validation samples. All activations functions were set to ReLU (instead of the output layer), the optimizer was Adam, and categorical cross-entropy was the loss function. The training data was given to each network in batches of 64 samples.

C. Transfer Learning

A two-phased transfer learning pipeline via neural freezing was designed to leverage healthy and dysarthric speech samples. To build the base models, both transformer architectures were initially trained on the LJ Speech speaker with a 99:1 training-to-validation ratio. Then, we investigated which layers to freeze and continued training with UA-Speech control samples before dysarthric models were trained. While building the control models, we omitted one of the UA-Speech control speakers to evaluate the transfer learning performance. This approach was used to maximize the advantages of the available normal speech data to overcome the scarcity of dysarthric data.

Neural freezing locks the weights assigned to the frozen neurons and forces the learning algorithm to converge on the new training data by only adjusting the unfrozen synaptic weights. Here, we applied neural freezing in two steps: once after the LJ Speech models were trained and once before the

dysarthric models were trained. In each step, we investigated freezing which layers delivered the best performance. The details of this investigation and which model components were frozen during each neural freezing phase are provided in section IV.

D. Data Augmentation

Data augmentation is a technique commonly used in machine learning tasks that applies random oscillations and perturbations to augment the training data without affecting the class labels. The primary goal of data augmentation is to improve model generalizability, especially when data is scarce. Adding additional data exposes the model to more data variations, which often leads to improving the training efficacy.

We utilized audio data augmentation to create extra voicegrams based on the available dysarthric speech data, increasing the number of training samples. The new voicegrams were created by shifting, noise injection, and speed and pitch changing, similar to [23]. Pitch and speed modifications were done by a factor of 10, white noise was altered by a random factor ranging from 0 to the length of the audio, and the shift was divided by a factor of 10. Each B1 and B2 UA-Speech dysarthric utterance was modified via the above augmentation techniques. The augmented samples were then added to the original dysarthric training data and used to train the dysarthric models.

E. Dysarthric Models Training and Evaluation

Once both transfer learning phases were applied and control models were ready, two dysarthric models were trained for each dysarthric speaker from Table I, one model per transformer architecture presented before. During training, we applied both B1 and B2 original and augmented utterances. However, different neural freezing configurations stated in the next section were studied per speaker to find the best match.

Since UA-Speech only provides isolated words and not connected speech, the performance of the models was measured using Word Recognition Accuracy (WRA) instead of WER. WRA is the most widely used metric in the literature to measure and benchmark UA-Speech, and defined as the percentage of the number of words the model could correctly identify to the number of words attempted.

IV. Experiments

All experiments were conducted on our specialized deep learning workstation with an AMD Ryzen Threadripper 3990X 64-Core CPU, 256Gb RAM, and four NVIDIA RTX 6000 GPUs.

Once both initial transformer architectures were finalized, they were trained for 200 epochs on the LJ Speech speaker's data as base models to prepare for the first phase of transfer learning. Next, for the first transformer model based on the architecture shown in Figure 2, the last two encoders were frozen, and then the model was trained on UA-Speech control speakers for another 100 epochs, yielding the first control model. For Transformer 2 (based on the modified architecture and encoder shown in Figure 3), the best results were obtained when all three decoders were frozen, resulting in the second control model. The training of the control models was done via a speaker-independent paradigm in which all speech samples from speaker CM06 were withheld during training and only used for testing the models, while utterances of the remaining UA-Speech control speakers were used to train the control models. The first transformer architecture delivered a WRA of 89% for the testing speaker, while the second architecture delivered a 92% WRA. For testing purposes, we also trained the second transformer only using the UA-Speech control data bypassing the first transfer learning phase. This model delivered the best WRA of 87% for CM06, indicating the effectiveness of the first transfer learning phase since a 5% WRA improvement was obtained.

After the control models were trained, they were saved and employed as the pre-trained networks for the second transfer learning phase and the training of dysarthric models. We trained and evaluated 30 speaker-adaptive dysarthric models, two models per dysarthric speaker based on both transformer architectures. All B1 and B2 UA-Speech dysarthric blocks plus the augmented training utterances were used to train the dysarthric models, and B3 speech samples were used for testing. Additionally, to measure the effectiveness of the audio data augmentation process explained before, we trained another 15 dysarthric models based on Transformer 1 control model by only using the original dysarthric data, excluding the augmented utterances.

In the second phase of transfer learning, another set of neural freezing was applied before the dysarthric models were trained, as stated before. Here, since the dysarthric speakers belonged to different severity classes, and to enable the models to better adapt to the variations of dysarthric speech per severity class, Phase 2 neural freezing was done differently, as shown in Table II.

V. Results and Discussion

Table III shows the testing results with all 45 dysarthric models with both transformer architectures. The inclusion of augmented audio data improved the first transformer's accuracy for 12 out of 15 speakers. The most significant increase in WRA was for speaker F03, where an improvement of 17% was achieved, while M08 showed the highest WRA decrease of 4% when augmented data was used during training. The average improvements for each intelligibility level with augmented data were 9%, 12%, and 3% for very low, low, and mild intelligibility subjects, but no improvement was gained with high intelligibility subjects. The augmented dysarthric voicegrams delivered an absolute average WRA improvement of 5% across all speakers, which shows data augmentation was overall effective. Data augmentation was particularly effective for very low and low intelligibility subjects, where significant improvements of up to 12% were achieved. The speakers where no improvement was seen via data augmentation were M09, M08, F05, and F04, among which the first three speakers' utterances were highly comprehensible and almost unrecognizable from normal speech. Hence, data augmentation was ineffective and even decreased performance for speakers whose speech was close to healthy.

TABLE II
TRANSFER LEARNING PHASE 2 NEURAL FREEZING CONFIGURATIONS

Intelligibility Level	Dysarthric Models	Transformer 1 (no data augmentation)	Transformer 1 (with data augmentation)	Transformer 2 (with data augmentation)
Very Low	M04, F03, M12, M12, M01	Decoder	Decoder	Last decoder
Low	M07, F02, M16	Last encoder's feed-forward component	Last encoder's feed-forward component	Last decoder
Mild	M05, M11, F04	Decoder's feed-forward component	Decoder	Last decoder's feed-forward component
High	M09, M14, M10, M08, F05	Decoder's feed-forward component	Second decoder's dense layer	Last decoder's feed-forward component

TABLE III
DYSARTHRIC MODELS TESTING RESULTS

Intelligibility Level	Dysarthric Models	Transformer 1 (no data augmentation) WRA	Transformer 1 (with data augmentation) WRA	Transformer 2 (with data augmentation) WRA
Very Low	M04	7%	9%	7%
	F03	34%	51%	51%
	M12	24%	34%	53%
	M01	31%	38%	51%
Very Low Intelligibility Average WRA (%)		*24%*	*33%*	*41%*
Low	M07	56%	72%	76%
	F02	52%	67%	74%
	M16	51%	56%	70%
Low Intelligibility Average WRA (%)		*53%*	*65%*	*73%*
Mild	M05	55%	62%	66%
	M11	45%	54%	58%
	F04	63%	56%	72%
Mild Intelligibility Average WRA (%)		*54%*	*57%*	*65%*
High	M09	79%	79%	84%
	M14	76%	79%	85%
	M10	89%	90%	91%
	M08	83%	79%	88%
	F05	89%	88%	93%
High Intelligibility Average WRA (%)		*83%*	*83%*	*88%*
Absolute Average WRA (%)		**53%**	**60%**	**67%**

Between the two transformer architectures trained with original and augmented data, Transformer 2 delivered better WRAs for 87% of the dysarthric participants. The biggest improvement was for severe dysarthric subject M12, with a 7% intelligibility level, for which Transformer 2 provided 19% better WRA, followed by mild intelligibility subject F04, with a 16% WRA increase. Nonetheless, the second transformer architecture did not improve WRA for the very low intelligibility subject F03 and obtained a 2% lower WRA for subject M04. Having a speech intelligibility score of 2%, M04 exhibited the lowest level of intelligibility among all dysarthric speakers. Although it is challenging to provide a concrete explanation for the subpar performance of both Transformers on this particular speaker, we hypothesize that the extremely limited intelligibility of M04 speech, combined with the speaker's distinct speech characteristics, led to substantial divergence from the speech features found in the healthy and control speech data used for pre-training the models. Consequently, the models struggled to establish a strong and reliable mapping of M04's speech signals to the designated vocabulary.

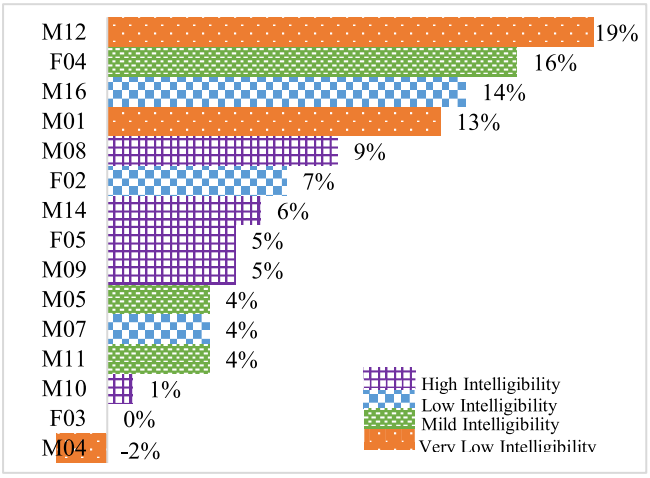

Fig. 4. Transformer 2 WRA improvements over Transformer 1 (with data augmentation).

TABLE IV
INTELLIGIBILITY LEVEL IMPROVEMENTS ACHIEVED BY TRANSFORMER 2 OVER TRANSFORMER 1

Intelligibility Level	Average WRA Improvements
Very Low Intelligibility	8%
Low Intelligibility	8%
Mild Intelligibility	8%
High Intelligibility	5%
Overall	**7%**

Figure 4 portrays Transformer 2 improvements over Transformer 1, and Table IV summarizes average improvements across intelligibility levels. With respect to the intelligibility levels, Transformer 2 obtained better results across all levels with an 8% average improvement across the spectrum instead of mild dysarthric subjects with high intelligibility, where on average, 5% better WRA was obtained. Overall, Transformer 2, which employed the deeper encoder architecture utilizing depthwise separable convolutions over feed-forward layers and an increased number of encoders and decoders, yielded an overall 7% better WRA over the traditional transformer architecture. The second transformer's better performance indicates that the increasing depth and the modifications made to the architecture were effective and enabled the model to capture the complexity of dysarthric speech better. Although this increased depth might seemingly lead to slower training and model inception, we observed no significant disparities in training or inception times between the two transformer architectures. This lack of distinction can be attributed to the fact that Transformer 2's depthwise convolution replaced the feed-forward blocks in Transformer 1, resulting in improved efficiency in the encoders of Transformer 2.

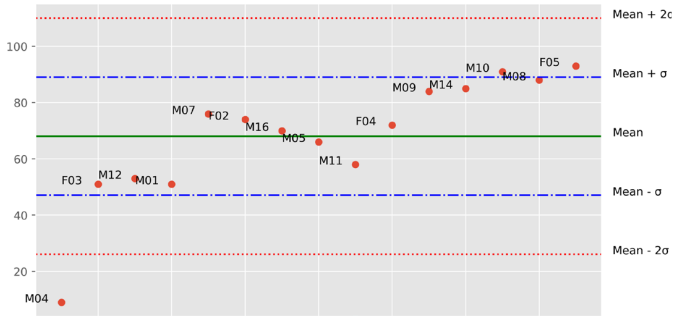

Fig. 5. Dysarthric speech transformer statistical analysis (WRA%).

Overall, the proposed Dysarthric Speech Transformer delivered a mean WRA of 68%, considering the best per-speaker results obtained from all three sets of experimental results shown in Table III with standard deviation σ=*21%*. Figure 5 portrays the statistical analysis of the proposed DST by plotting the best WRA achieved per speaker concerning mean and $\pm 1\sigma$ and $\pm 2\sigma$ above and below the mean. As can be seen, 80% of the WRAs obtained are within $\pm 1\sigma$ of the mean, and the remaining WRAs are very close to *mean*$+\sigma$ and below $+2\sigma$, except M04, which was an outlier. The WRA distribution shown in Figure 5 indicates a normal distribution, which further establishes confidence in the results obtained from the proposed DST.

A. Performance Comparative Study and Benchmarking

Table V provides a comparative study between the proposed DST and state-of-the-art WRA results reported in the literature for the UA-Speech corpus. A fair comparison should pay special attention to the number of dysarthric participants in the study, the intelligibility class of the subjects, and the vocabulary size. These criteria are important because increasing the vocabulary size adds to the complexity of ASR, and including more dysarthric subjects indicates better generalizability of the results and higher statistical significance.

The highest WRA reported on UA-Speech is [24]. However, in this study, all dysarthric utterances were mixed, then divided with a 75:25 train/test ratio. Given UA-Speech presents each utterance eight times, once for each different microphone array setup, this strategy of defining train and test speech utterances should be avoided. This strategy likely employs identical utterances for training and testing, as all microphone samples were derived from the same recording source. This means the results presented did not indicate the model's generalizability but memorability and training performance. Additionally, the results reported in [24] were achieved over a small vocabulary of 29 words.

Likewise, while the authors of study [25] indicated that each participant's speech samples were divided into three categories, they did not indicate whether the data was divided based on UA-Speech block categorization (B1, B2, and B3 utterances) or microphone data; hence their results may have the same limitation as [24]. Besides, not all UA-Speech dysarthric speakers were employed in their study. The next top WRA was reported in [26], but the vocabulary size was only 25 words, and the speech samples of only seven dysarthric participants were used in this study.

TABLE V
UA-SPEECH WRA PERFORMANCE COMPARISON

Reference	ASR Paradigm	# UA-Speech Participants	Best WRA Reported	Comments
[27]	Speaker Dependent	7	Maximum Average WRA 30.8%	PLP + MAP Adaptation and HMM
[27]	Speaker Adaptive	7	Maximum Average WRA 36.8%	PLP + MAP Adaptation and HMM
[26]	Speaker Dependent	7	Absolute Average WRA 81%	MFCCs+MLPs with MVML architecture
[26]	Speaker Independent	All	Absolute Average WRA 75%	MFCCs+MLPs with MVML architecture
[24]	Speaker Dependent	All	Absolute Average WRA 88%	MFCCs+LL-SVM
[25]	Speaker Dependent	9	Absolute Average WRA 85%	GNE+RNN
[28]	Speaker Dependent	All	Absolute Average WRA 59%	The model was a hybrid MAP-MLLR-HMM with MFCC
[29]	Speaker Adaptive	All	Absolute Average WRA 54.16%	PLP features +HMMs
[5]	Speaker Adaptive	All	Absolute Average WRA 61% with no synthetic data used	The model was based on Spatial Convolutional Network to recognize word shapes presented as voicegrams
The Proposed DST	Speaker Adaptive	All	Absolute Average WRA 68%	Transformer and self-attention

GNE: Glottal to Noise Excitation, PLP: Perceptual Linear Prediction, MAP: Maximum A Posteriori, HMM: Hidden Markov Model, MFCC: Mel-Frequency Cepstral Coefficients, MLP: Multilayer Perceptron ANN, MVML: Multi-View Multi Learner, SVM: Support Vector Machine, RNN: Recurrent Neural Network

Thus, the average WRA reported in [5] is currently the highest reported in the literature for UA-Speech that not only obtained from all 15 UA-Speech participants but also applied the same train/test split strategy we considered in our study in which dysarthric B1 and B2 samples were used for training and B3 for testing. Similarly, the vocabulary used in [5] is identical to ours. As such, we selected the dysarthric ASR reported in [5] to benchmark our proposed DST since it enables us to perform a direct, fair comparison. Accordingly, per-speaker comparisons between the best WRAs achieved by the proposed Dysarthric Speech Transformer and the benchmark system are depicted in Figure 6. Notably, study [5] reported two sets of results, one when their model was trained

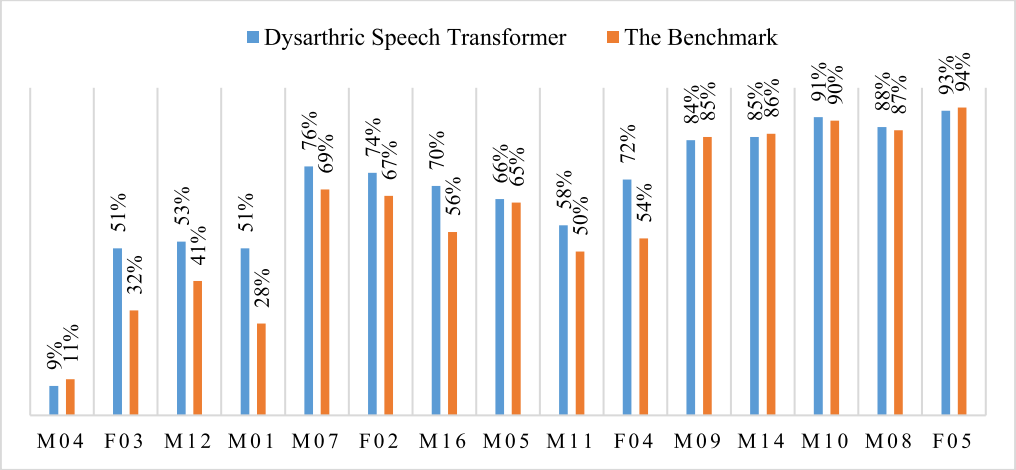

Fig. 6. The proposed dysarthric speech transformer word recognition accuracies vs [5].

with the original and visually augmented data and another set 6when synthetically generated dysarthria speech samples were also added. Since we did not use synthetic data, our comparison is based on the results reported for the benchmark model trained with the original and augmented dysarthric data, excluding experiments with synthetic data.

The proposed DST achieved better performances for 11 out of 15 dysarthric speakers. We can see significant improvements for speakers M01, F03, and F04, with 23%, 19%, and 18% better WRAs, respectively. DST's improvements are more substantial for severe dysarthria, in which the DST obtained, on average, 13% better results than the benchmark, followed by an average of 9% better WRAs for low and mild intelligibility speakers. On the other hand, there was no noticeable change for high intelligibility speakers where both the proposed DST and the benchmark ASR performed similarly. Overall, the DST improved WRA over [5] by an average of 7%. Nonetheless, the benchmark dysarthric ASR delivered slightly better WRA for M04.

VI. Conclusion

In this study, we experimented with transformer and attention-based architectures and proposed two seq-to-seq architectures to develop a dysarthric ASR. We experimented and measured how increasing the depth and number of transformer encoders and decoders could lead to better performances, and how using depthwise separable convolution instead of fully-connected encoder components could improve word recognition accuracies. To tackle the scarcity of dysarthric speech data to train deep transformers, we developed a two-phase transfer learning pipeline and investigated the best neural freezing configurations that best retain the knowledge acquired from healthy speakers. We have also studied how applying audio data augmentation could help further address the data scarcity issues. Overall, we trained and evaluated 45 dysarthric models based on two transformer architectures we designed. Our proposed Dysarthric Speech Transformer outperformed the state-of-the-art and delivered up to 23% better accuracies. Specifically, the DST was more capable of recognizing severe dysarthria with very low speech intelligibility compared to the benchmark ASR.

We recommend the following avenues for future researchers to investigate and further improve the DST:

- We used B1 and B2 UA-Speech samples during our experiments for training and B3 for testing. However, this strategy did not use uncommon words provided by the dataset. We recommend future researchers include all available UA-Speech data during training but only withhold one of the B blocks for testing with unforeseen utterances. Including more dysarthric training samples may result in further DST performance improvements. We did not apply this strategy to maintain consistency with the previous notable studies and enable fair benchmarking.
- Alternatively, common words could be used for training and uncommon words for testing, and vice versa, as this strategy still keeps training data from leaking into the testing set, providing objective measures of generalizability.
- Since the models were initially trained on a large corpus with an open-set vocabulary, we can consider DST an open-vocabulary ASR. However, because UA-Speech is a closed-set vocabulary, our evaluation was done as a fixed-vocabulary ASR, which is common in dysarthric ASR research on UA-Speech. For example, all studies reported in Table V were evaluated as fix-vocabulary ASRs. Nevertheless, given the open-vocabulary nature of the base model, cross-database evaluation could explore the openness of the proposed ASR.
- Likewise, even though our evaluations were based on isolated words, the proposed DST can recognize connected speech since it maps speech utterances to a sequence of characters. Other less widely used dysarthric datasets, such as TORGO, contain connected dysarthric speech. Hence, this can be another avenue for future studies to explore how well the DST performs as a connected speech dysarthric ASR. This strategy could be specifically of interest to mild and high intelligibility dysarthric speakers who are more capable of speaking continuously in contrast to severe dysarthric patients.
- The use of synthetically generated dysarthric speech has been investigated and shown effective in helping rectify the data scarcity issue [30]. This approach could produce

unlimited synthetic speech for a given dysarthric speaker, which could further augment dysarthric speech samples. We recommend future studies investigating whether this can lead to open-vocabulary ASR or performance improvements.

REFERENCES

[1] T. Tamura, Y. Tanaka, Y. Watanabe, and K. Sato, "Relationships between maximum tongue pressure and second formant transition in speakers with different types of dysarthria," *PLoS ONE*, vol. 17, no. 3, Mar. 2022, Art. no. e0264995, doi: 10.1371/journal.pone.0264995.

[2] C. Tanchip et al., "Validating automatic diadochokinesis analysis methods across dysarthria severity and syllable task in amyotrophic lateral sclerosis," *J. Speech, Lang., Hearing Res.*, vol. 65, no. 3, pp. 940–953, Mar. 2022, doi: 10.1044/2021_JSLHR-21-00503.

[3] B. F. Zaidi, S. A. Selouani, M. Boudraa, and M. S. Yakoub, "Deep neural network architectures for dysarthric speech analysis and recognition," *Neural Comput. Appl.*, vol. 33, pp. 1–20, Jan. 2021, doi: 10.1007/s00521-020-05672-2.

[4] S. H. Lee, M. Kim, H. G. Seo, B.-M. Oh, G. Lee, and J.-H. Leigh, "Assessment of dysarthria using one-word speech recognition with hidden Markov models," *J. Korean Med. Sci.*, vol. 34, no. 13, p. e108, 2019, doi: 10.3346/jkms.2019.34.e108.

[5] S. R. Shahamiri, "Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 852–861, 2021, doi: 10.1109/TNSRE.2021.3076778.

[6] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Jun. 2017, doi: 10.48550/arxiv.1706.03762.

[7] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4945–4949, doi: 10.1109/ICASSP.2016.7472618.

[8] S. S. Tirumala and S. R. Shahamiri, "A deep autoencoder approach for speaker identification," in *Proc. 9th Int. Conf. Signal Process. Syst.*, Nov. 2017, pp. 175–179, doi: 10.1145/3163080.3163097.

[9] A. Vaswani et al., "Tensor2Tensor for neural machine translation," 2018, *arXiv:1803.07416*.

[10] F. Chollet, *Deep Learning With Python*, 2nd ed. NY, USA: Manning Publication, 2021. Accessed: Jan. 3, 2023. [Online]. Available: https://www.manning.com/books/deep-learning-with-python-second-edition

[11] S. Karita et al., "A comparative study on transformer vs RNN in speech applications," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 449–456, doi: 10.1109/ASRU46091.2019.9003750.

[12] S. R. Shahamiri, W. M. N. W. Kadir, and S. Ibrahim, "A single-network ANN-based oracle to verify logical software modules," in *Proc. 2nd Int. Conf. Softw. Technol. Eng.*, Oct. 2010, pp. V2-272–V2-276, doi: 10.1109/ICSTE.2010.5608808.

[13] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5884–5888, doi: 10.1109/ICASSP.2018.8462506.

[14] A. Graves, "Sequence transduction with recurrent neural networks," Nov. 2012, *arXiv:1211.3711*, doi: 10.48550/arxiv.1211.3711.

[15] R. L. MacDonald et al., "Disordered speech data collection: Lessons learned at 1 million utterances from project Euphonia," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2021, pp. 4833–4837, doi: 10.21437/Interspeech.2021-697.

[16] Z. D. Champiri, S. S. B. Salim, and S. R. Shahamiri, "The role of context for recommendations in digital libraries," *Int. J. Social Sci. Humanity*, vol. 5, no. 11, pp. 948–954, 2015, doi: 10.7763/ijssh.2015.v5.585.

[17] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "Nemours database of dysarthric speech," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, vol. 3, 1996, pp. 1962–1965.

[18] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, Dec. 2012.

[19] H. Kim et al., "Dysarthric speech database for universal access research," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Brisbane, QLD, Australia, 2008, pp. 1741–1744.

[20] M. C. Coppens-Hofman, H. Terband, A. F. M. Snik, and B. A. M. Maassen, "Speech characteristics and intelligibility in adults with mild and moderate intellectual disabilities," *Folia Phoniatrica et Logopaedica*, vol. 68, no. 4, pp. 175–182, 2016, doi: 10.1159/000450548.

[21] *The LJ Speech Dataset*. Accessed: Oct. 7, 2022. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[22] T. O'Malley et al. (2019). *KerasTuner*. [Online]. Available: https://github.com/keras-team/keras-tuner

[23] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH, ISCA)*, Sep. 2018, pp. 471–475, doi: 10.21437/Interspeech.2018-1751.

[24] N. Rajeswari and S. Chandrakala, "Generative model-driven feature learning for dysarthric speech recognition," *Biocybernetics Biomed. Eng.*, vol. 36, no. 4, pp. 553–561, Jan. 2016, doi: 10.1016/J.BBE.2016.05.003.

[25] S. S. Nidhyananthan, R. S. S. Kumari, and V. Shenbagalakshmi, "Assessment of dysarthric speech using Elman back propagation network (recurrent network) for speech recognition," *Int. J. Speech Technol.*, vol. 19, no. 3, pp. 577–583, Sep. 2016, doi: 10.1007/s10772-016-9349-1.

[26] S. R. Shahamiri, "Neural network-based multi-view enhanced multi-learner active learning: Theory and experiments," *J. Exp. Theor. Artif. Intell.*, vol. 34, no. 6, pp. 989–1009, Nov. 2022, doi: 10.1080/0952813X.2021.1948921.

[27] H. V. Sharma and M. Hasegawa-Johnson, "State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition," in *Proc. NAACL HLT Workshop Speech Lang. Process. Assistive Technol.*, Los Angeles, CA, USA, 2010, pp. 72–79.

[28] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *Proc. 6th Workshop Speech Lang. Process. Assistive Technol. (SLPAT)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 65–71, doi: 10.18653/v1/W15-5112.

[29] H. Christensen, S. P. Cunningham, C. Fox, P. D. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc. 13th Annu. Conf. Int. Speech Commun. Associatio (INTERSPEECH)*, Portland, OR, USA, 2012, pp. 1776–1779.

[30] A. Hu, D. Phadnis, and S. R. Shahamiri, "Generating synthetic dysarthric speech to overcome dysarthria acoustic data scarcity," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 6, pp. 6751–6768, Jun. 2023, doi: 10.1007/s12652-021-03542-w.

[31] S. Liu et al., "Recent progress in the CUHK dysarthric speech recognition system," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2267–2281, 2021, doi: 10.1109/TASLP.2021.3091805.