2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) | 979-8-3503-2797-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/SpeD59241.2023.10314881

Dysarthric Speech Recognition: A Comparative Study

Dhvani Shah, Vanshika Lal, Zihan Zhong, Qianli Wang, Seyed Reza Shahamiri*
Department of Electrical, Computer, and Software Engineering
Faculty of Engineering
The University of Auckland
Auckland, New Zealand
{dsha439, vlal080, zzho680, qwan121}@aucklanduni.ac.nz, *admin@rezanet.com

***Abstract*—Automatic Speech Recognition (ASR) systems are designed to convert spoken words into written text. These systems have the potential to greatly benefit individuals with speech impairments like dysarthria, improving their overall quality of life. However, there are significant challenges that researchers must address in order to develop an effective ASR system for atypical speech, specifically dysarthric speech. Among these challenges are the limited availability of dysarthric speech data and the need for an accurate interpretation model. In this study, we explored the development of dysarthric ASR systems, starting from traditional approaches that were specifically designed to recognize dysarthric speech. We then delve into deep-learning-based ASR systems and modern ASR engines that utilize transformers. Additionally, we highlighted the notable dysarthric speech collections and identified the key factors that impede the performance of dysarthric ASR systems. The findings indicate that although researchers have achieved gradual improvements in performance, further research is required to achieve practical performance levels. Likewise, it is important to establish guidelines and standards to accurately measure and benchmark the performance of dysarthric ASR systems in order to ensure their usefulness and effectiveness in real-world applications.**

Keywords—Dysarthria, dysarthric speech recognition, transformers, machine learning

I. INTRODUCTION

Dysarthria is a speech disorder characterized by difficulty in articulating words and sounds due to weak, imprecise, or uncoordinated movements of the muscles involved in speech production. It is caused by damage or impairment to the nerves, muscles, or brain structures that control speech production, such as the tongue and lips. Individuals with dysarthria may exhibit various symptoms depending on the underlying cause and severity of the condition. These symptoms can include slurred speech, unclear pronunciation, monotone or abnormal pitch variations, slow or rapid speech rate, and difficulty controlling volume or intensity of speech. The intelligibility and comprehensibility of speech can be significantly affected, making it challenging for others to understand the individual's spoken communication.

Automatic Speech Recognition (ASR) systems transcribe speech signals to the corresponding text. They have been utilized to develop technologies for aiding people with speech impairments such as dysarthria. By offering a medium of engagement, ASR systems can significantly improve the quality of their life. However, researchers need to deal with multiple challenges in designing Dysarthric Speech Recognition (DSR) systems. Among the most prominent challenges are the difficulty in obtaining a large enough speech dataset, aka data scarcity, and developing an accurate interpretation model for such speakers, given the large phone inaccuracies and speech variations among dysarthric speakers [1].

An overview of how ASR systems perform for dysarthric speakers is provided by Rosen et al. [2]. The authors suggested that speaker-adaptive (SA) systems equipped with extensive vocabularies might be advantageous for individuals with moderate dysarthria but indicated further research is still needed. Another review conducted in [3] compared different acoustic modeling strategies such as generative Hidden Markov Models (HMM), Artificial Neural Networks (ANN), Multi-Layer Perceptron (MLP), and hybrid approaches. The study concluded that researchers who combined MLPs with HMMs yielded better recognition rates, yet it was not clear whether this conclusion was made by considering DSRs or only ASRs designed for healthy speakers. Furthermore, this study was conducted before deep learning-based ASR approaches obtained state-of-the-art performances and outperformed traditional acoustic models.

Despite substantial advances in ASR technologies, even the best-performing normal speech ASR systems are ineffective for speakers with impaired speech. A study conducted by Rowe et al. [4] indicated that the deficiency may be due to the difficulties of getting sufficiently diversified training samples of impaired speech. From a clinical standpoint, they uncovered how having variety may affect ASR performance. Many techniques need large-scale data collection activities to support the development of a generalized ASR system for difficult-to-recognize speakers. Given a sufficiently diversified dataset, end-to-end (E2E) machine learning models may eliminate the requirement to comprehend the underlying pathophysiological processes.

In this study we compared the performances of notable DSR systems ranging from traditional DSR approaches to those with deep learning engines. We have also provided an overview of dysarthric speech corpora that have been used widely by DSR researchers and utilized for benchmarking commonly.

II. DYSARTHRIC SPEECH CORPORA

This section explains a few English dysarthric speech datasets that have been used to develop DSR systems. There are also datasets for languages other than English that are not listed below as our study only surveyed English DSRs.

A. UA-Speech

The University of Illinois established the UA-Speech corpus [5], designed for ASR applications for patients with neuromotor impairments. The dataset comprises recordings of a 7-channel microphone array and a video camera from 19 speakers with cerebral palsy (speech samples from 15 speakers are available publicly) and 13 age-matched healthy controls. Each speaker was recorded uttering 765 isolated words, including ten numeral digits, 26 radio alphabets, 19 computer commands, 100 common words, and 155 uncommon words. The uncommon words were chosen using a greedy algorithm to increase uncommon phonemes.

The intelligibility of UA-Speech speakers was assessed by five listeners and provided in the dataset. Previous studies commonly classified the dysarthric severity of the participants based on the intelligibility level of the speaker [6] into very low, low, mid, and high intelligibility, corresponding to very severe, severe, moderate, and mild dysarthria, respectively. This is an important notation since the performance of DSR systems varies significantly across different dysarthria severity levels.

UA-Speech has been the most widely used dysarthric speech corpus in the literature [7] at the time of this writing since it had participants from a more diverse dysarthria spectrum compared to the other publicly available corpora, enabling researchers to measure the generalizability of their models better.

B. TORGO

The University of Toronto's TORGO corpus [8] includes dysarthric speech caused by cerebral palsy (CP) and Amyotrophic Lateral Sclerosis (ALS). The speech prompts were in the forms of isolated words, restricted sentences, and unrestricted sentences from eight dysarthric participants (three females, five males) and seven healthy participants. Non-words prompts were also included to test the speakers' articulatory control relating to plosives and prosody. Unlike UA-Speech, this corpus did not clearly indicate the speech intelligibility of the dysarthric participants, making it more challenging to measure performance per dysarthric intelligibility class.

C. Nemours

The Nemours database [9] contains 814 short nonsensical phrases in the format of "The *X* is *Y*ing the *Z*," collected from eleven male dysarthric speakers. *X* and *Z* were never the same and selected without replacement from a set of 74 monosyllabic nouns. *Y* was chosen without replacement from a set of 37 disyllabic verbs. Each speaker had varied degrees of dysarthria and recorded 74 sentences and two connected-speech paragraphs. Compared to TORGO and UA-Speech, Nemours has been the least used corpus among DSR researchers.

D. Project Euphoria Database

The Project Euphonia [10] databases collected by Google comprise two categories of non-standard speech databases. The first was accented speech from the L2 Arctic dataset. The other was the data gathered in collaboration with ALS Therapy development, which contains 36 hours of audio from 67 people with ALS. Using their home computers, the participants self-recorded reading phrases from a relatively limited linguistic domain. Many phrases were simple prompts with simple grammatical structures (for example, "When is the basketball game tonight?"). However, many typical filler words in normal speech, such as "um" and "uh," were not included in the recordings. Unlike the previous corpora, this corpus is not publicly available at the time of this writing.

III. Traditional Dysarthric ASR

The literature shows many attempts have been undertaken to develop ASR technologies using dysarthric speech datasets. Among the earlier attempts were generative machine learning approaches to model time-varying spectral sequences, and HMMs were widely used to capture the variability in dysarthric speech, considering the prevalence of HMM-based ASR systems prior to the emergence of neural networks and deep learning. In this regard, we can refer to Deller et al. [11] who used a vector quantized HMM approach to model isolated dysarthric speech to deal with the enormous variability of dysarthric speech, validated with a 44-year-old male speaker with mixed spastic athetoid CP. According to the Computerized Assessment of Intelligibility of Dysarthric Speech (CAIDS) index, the participant's intelligibility score was 59.3%. Two different sets of vocabulary were used to test the recognition performance. Likewise, the same individual and vocabulary were used to assess the improvement of the transition clipping HMM procedure and compared results with a standard HMM. The authors discovered that combining a fully structured HMM with transitional region clipping enhanced recognition. This study was also extended to two individuals with cerebral palsy considered moderately intelligible and truly nonverbal. Even with speech unintelligible to an untrained listener, this ASR delivered a recognition rate of 78%. The study indicated employing Mel Frequency Cepstrum Coefficients (MFCCs) could aid in getting higher results.

Nevertheless, generative model-based classifiers such as HMMs perform poorly due to overlapping classes and insufficient training data. Dysarthric speech is typically partial or incomplete, resulting in incorrect temporal dynamics learning. To address these challenges, Rajeswari and Chandrakala [12] studied fixed dimensional vector representations that provided better discrimination for dysarthric speech than the conventional HMMs. They focused on learning features for dysarthric speech recognition, which required recognizing consecutive patterns of six different duration utterances. The authors experimented with various models, such as log-likelihood Support Vector Machine (LL-SVM) and transition SVM, and achieved an accuracy of 87.91% and 73.68% on UA-Speech, respectively, outperforming the traditional HMM-based approaches.

Furthermore, HMMs suffer from other limitations that impede their performance. For instance, HMMs assume that all probabilities depend solely on the current state, resulting in a limited ability to capture correlations between neighboring input frames. Moreover, the probability density models used in HMMs exhibit suboptimal modeling accuracy, and the requirement of Maximum Likelihood training leads to inadequate discrimination among acoustic models when insufficient training data is available [13]. These limitations, coupled with the rise of neural networks, have motivated DSR

researchers to shift their focus toward exploring deep learning-based approaches [14].

IV. Deep Learning-based dysarthric ASR

Speech recognition using neural networks can help to overcome the aforementioned HMM constraints. During training, neural networks progressively update information learned from prior knowledge, allowing them to deal with changes and irregularities in the speech signal as well as fragmented or missing information [15]. Amongst the earliest studies in this context is Jayaram et al. [16], which used ANNs to recognize isolated word dysarthric speech. While the overall accuracy was not disclosed in this study, experiments were conducted with one subject with severe dysarthria with an intelligibility score between 10%-20%.

Another example using ANNs and MLPs is [17] in which the authors initially studied the different MFCC configurations to identify the best acoustic features presenting dysarthric speech. Once the best acoustic features were identified, the authors developed a dysarthric multi-networks speech recognizer (DM-NSR) based on the Enhanced Multi-Learner active theory. The DM-NSR leverages ensemble learning and employs an array of neural networks to learn the acoustic features associated with individual words in the vocabulary. Evaluating on speech samples provided by all UA-Speech participants via both speaker-dependent and independent paradigms, recognition rate improvement by up to 24.67% was reported for a small 25 words vocabulary.

Another example is Elman Backpropagation Network (EBN) adopted in [18], in which a DSR was proposed and evaluated on nine randomly selected subjects from UA-Speech using a 100-word vocabulary, achieving a mean accuracy of 85.05%. Glottal to noise excitation (GNE) ratio was used in this study, which indicates whether a given voice signal originates from vibrations of the vocal folds or turbulent noise generated in the vocal tract. It is related to breathiness measure and was applied after speech pre-processing. Comparing GNE and MFCCs, the study found GNE's superiority.

As stated before, hybrid models that combine neural networks with other machine learning algorithms were also investigated in the context of DSRs. An example is the HMM/ANN DSR proposed by Polur and Miller [13], which was evaluated on speech samples obtained from three moderately intelligible cerebral palsy speakers, obtaining a mean recognition rate of 97% on a 25-word vocabulary.

The introduction of Deep Neural Networks (DNN) was prompted by the need to address the sensitivity of traditional machine learning algorithms to discrepancies between training and test data caused by the greater variability among speakers and the heterogeneous nature of dysarthric datasets. This sensitivity hampers the ability of traditional algorithms to effectively generalize on unseen speech patterns and handle larger vocabularies. The spectral transition is important in capturing the local temporal–dimensional properties of dysarthric speech, in which signals vary more noticeably than those produced by a physically unimpaired person. Among studies that investigated this point and compared DNN based DSRs and traditional models are Convolutive Bottleneck Network (CBN) speaker-dependent DSR [19], and España-Bonet et al. [20] that compared classical and neural network architectures on the TORGO corpus. Likewise, Zaidi et al. [21] investigated how Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs) could improve DSR performance on the Nemours dataset. The results showed that the CNN-based system achieved a better accuracy of up to 82% for one mild-severity dysarthric subject.

Transfer learning is a machine learning technique by which knowledge acquired in a model for one task is reused as the foundation for another task [22], considering the task context [23]. Given the scarcity of dysarthric speech, transfer learning has gained popularity in training deeper models. For example, Speech Vision [24] is an innovative DSR system that adopts a novel acoustic modeling approach mapping the spectrogram shapes to words. This study applied transfer learning via partial neuron freezing to retain general knowledge obtained from training the model on healthy speech. Additionally, visual data augmentation techniques and synthetically generated dysarthric speech [25] were applied to expand the dysarthric training samples. The model achieved an absolute average word recognition accuracy of 64.71% from evaluating the model on UA-Speech participants with a 155-word vocabulary, achieving state-of-the-art results on UA-Speech.

V. Sequence-to-sequence Models

A sequence-to-sequence ASR employs a sequence-to-sequence model architecture to convert spoken language into written text. This model architecture is commonly used in machine translation tasks but has also been adapted for speech recognition. In a sequence-to-sequence speech recognition system, the input sequence is the acoustic frames or spectrogram representation frames of the speech signal, while the output sequence is the corresponding sequence of words, characters, or phones. The model usually consists of an encoder network that processes the input sequence and captures the relevant acoustic features, and a decoder network that generates the output sequence by predicting the most likely tokens. Unlike the previous DSRs, encoder and decoder networks are trained in an end-to-end manner in contrast to separate training of acoustic, language, and pronunciation models (where applicable). These systems have shown promising results and are particularly effective in handling tasks where the input and output sequences have variable lengths. While seq-to-seq DSRs are yet to be widely studied, this section reviews some of the previous works with related applications.

Parrotron is an end-to-end speech-to-speech conversion model developed by Google Research [26] that employs an autoregressive Recurrent Neural Network as a spectrogram decoder, alongside an ASR decoder for multitask training. This model can normalize speech regardless of accent, prosody, or background noise, into the voice of a predefined speaker. This model can adapt to enhance the intelligibility and naturalness of atypical speech, such as speech from a deaf speaker. The normalized speech was evaluated with a state-of-the-art ASR engine, using the Word Error Rate (WER) as a metric to present speech intelligibility. The spectrogram decoder is complemented by a pre-net and attention mechanism that processes the output from each decoder time step. This setup

involves a stack of two LSTM layers and a convolutional post-net, generating the target spectrogram prediction and enhancing it with residual information.

Taking advantage of the representations learned by Parrotron's encoder, an additional text decoder was added to recognize speech without a significant increase in the number of model's hyperparameters [27]. This means the model transforms input spectrograms to output spectrograms in the voice of a predetermined target speaker while generating hypotheses in a target vocabulary. The system was reported to reduce WER by 77% in atypical speech with as little as an hour of training. Furthermore, by using a customized synthesizer built on atypical speech for data augmentation, the model achieves an additional 10% improvement. The authors demonstrated the method's generalization across eight types of atypical speech for various speech impairment severities.

Conformer Parrotron [28] improves over Parrotron by enhancing its accuracy, voice conversion quality, training speed, and performance over atypical speech. This new model builds upon the Conformer encoder by leveraging Transformers. The Conformer adapts the relative sinusoidal positional encoding scheme derived from Transformer-XL [29]. This relative positional encoding scheme enables the self-attention module [30] to improve generalization over the lengths of input sequences, allowing the entire model to better accommodate the diverse lengths of input waveforms. The Conformer architecture for joint voice conversion and recognition resulted in a relative reduction in WER of between 24% and 32% for normal speech, as observed in the Librispeech dataset and Google's in-house Voice Search traffic data. These architectures reportedly could lead to a significant WER reduction of up to 53% while reducing half of the training time for atypical speech, particularly dysarthric speech.

Lidan et al. [31] introduced a contrastive learning framework to interpret dysarthric speech. The model uses a projection head to retain critical task-specific information and employs data augmentation techniques to mitigate data scarcity. It applies an attentive temporal pyramid network for faster, parallel computation while incorporating data augmentation methods such as time warping, frequency masking, and time masking [32]. This approach was integrated with a Connectionist Temporal Classification (CTC) attention-based decoder, which validated the efficiency of the learned speech representations, supporting flexible model fine-tuning, and enhanced the system's robustness and adaptability to various dysarthric speech patterns. In both frozen and fine-tuning tests, the model consistently achieved superior performance across speakers with varying degrees of dysarthria, thus confirming the robustness of their framework.

Further advancements were made by Ding et al. [33], who introduced a multi-task Transformer with an auxiliary task of input feature reconstruction to address the complexities of dysarthric speech recognition. The model consisted of a primary dysarthric speech recognition task and an auxiliary speech feature reconstruction task, both sharing the encoder network. By employing intra-domain and cross-domain reconstruction methods, the model effectively handled the intra and inter-speech variability often encountered in dysarthric speech, enhancing the robustness and generalization of speech representations. The proposed model surpassed all other methods examined in the study in terms of both WER and Character Error Rate (CER) across all dysarthric speakers, highlighting its potential for significant advancements.

While some of the previous studies leveraged Transformers, none of them were based on the state-of-the-art seq-to-seq attention-based ASR architectures or mainly focused on dysarthria. In this context, a recent study proposed Dysarthric Speech Transformer, as the first DSR based on the modern ASR architectures that delivered state-of-the-art results on UA-Speech [34].

VI. Discussion

ASR systems designed for healthy speakers are often ineffective in accurately transcribing dysarthric speech, particularly in cases of severe dysarthria, due to substantial deviations from normal speech patterns. Therefore, specialized speech recognition systems need to be developed specifically to handle dysarthric speech. Multiple efforts have been made to create DSR technologies, and Table 1 provides a summary and comparison of the notable DSR systems published in the literature. However, comparing these DSR systems is a complex task, as various factors need to be considered to ensure a direct and fair comparison. The complexities of comparing DSR systems stem from several factors, including the number of participants, vocabulary size, ASR paradigm (speaker-dependent (SD), speaker-adaptive (SA), or speaker-independent (SI)), dysarthric severity, and intelligibility level. The performance of DSR systems is significantly influenced by these factors.

It is important to note that DSR performances reported on a limited number of speakers or studies that do not report per-severity/intelligibility levels may not generalize well to broader contexts. Additionally, increasing the vocabulary size can substantially increase the complexity of the ASR task, especially when considering isolated word DSR systems. Furthermore, not all studies clearly specified how they defined the training and testing data to ensure that training samples did not unintentionally influence the testing data, thus affecting model generalization. Cross-validation was not conducted in all studies either. Therefore, readers need to consider these limitations when interpreting the performances provided in Table 1.

We recommend that the DSR research community establish standards and guidelines for validating DSR models and develop clear strategies for splitting training and test data in publicly available dysarthric speech datasets. Based on our observations, a common strategy used with the UA-Speech dataset for SD or SA paradigms involved utilizing two out of the three speech blocks per speaker for training and validation, reserving the third block exclusively for testing the model's performance. This strategy ensures the training utterances are not leaked into testing, and hence the results reported are proper indication of model' generalization. There were other studies that shuffled all available samples, and then divided them into train/validation and testing. Given each recording is provided multiple times in UA-Speech, this strategy increases the likelihood of the training utterances being used for testing,

TABLE I. COMPARATAIVE STUDY

Paper, Year	ASR Paradigm	Model	Database	Vocab	Subjects	Performance
[11], 1991	SI	HMM	W-10 and W-196	10 digits + 196 common words	Three participants with cerebral palsy: CJ 65%, LE 59%, RP 22% CAIDS index	LE: 92% digits, 88.3% words; CJ 95%, RP 78%
[12], 2016	SD	HMM-SVM	UA-Speech	10 digits, 26 radio alphabets, 10 computer commands, and 100 common words	11 male and 4 female by intelligibility 4 very low, 3 low, 3 mid and 5 high	Overall WRA accuracy of 87.91% for LL-SVM and 73.68% for TP-SVM
[16], 1995	SD	ANN	None	20 words	1 individual with cerebral palsy and	Peak WRR of 42.5%
[17], 2021	SD	EMAL-MLP	UA-Speech	25 Words	All UA-Speech Speakers	Average WRR of 80.83%
[17], 2021	SI	EMAL-MLP	UA-Speech	25 Words	13 UA-Speech Speakers	Average WRR of 75.41%
[18], 2016	SD	EBN	UA-Speech	100 words (digits, computer commands, common and uncommon words)	9 dysarthric speakers with varying speech intelligibility rates (from very low to mid)	Average WRA of 85.05%
[35], 2015	SA	DNN-HMM	TORGO	NA	15 subjects (8 dysarthric speakers with varying severity and 7 control speakers without any disorder)	Outperformed classical models improving Word Error Rate (WER) by 3% for control speakers and 13% for subjects with dysarthria.
[36], 2020	SD	EMDH-CNN	Nemours	NA	11 American male speakers with different degrees of dysarthria	Increased the overall accuracy by 20.72% and 9.95% compared to HMM-GMM and CNN systems, respectively
[24], 2021	SA	Spatial CNN	UA-Speech	155 words: 10 digits, 19 computer commands, 26 radio alphabets, and 100 common words	15 in total. by intelligibility: 4 very low, 3 low, 3 mild, 5 high	Average WRA of 64.71%
[27], 2021	SA	Transformer	LibriSpeech + Google voice search + Project Euphonia	NA	14 speakers with 8 types of etiologies, 7 of which are ALS	Average WER of 12.2%
[28], 2021	SA	Conformer	LibriSpeech + Google voice search + Project Euphonia	NA	10 speakers, 8 with ALS and another 2 speakers each with distinct etiology	Average WER of 9.9%
[31], 2021	SI	Contrastive Learning Framework	Pre-trained on LibriSpeech, fine-tuned on TORGO	NA	15 speakers with 8 dysarthric, 7 healthy	Average CER of 19%, WER of 19.11%
[33], 2021	SI	Multi-Task Transformer	Pre-trained on LibriSpeech, fine-tuned on TORGO	NA	15 speakers with 8 dysarthric, 7 healthy	Average CER of 16.22%, WER of 15.87%
[34], 2023	SA	Attention-based Transformer	UA-Speech	155 words: 10 digits, 19 computer commands, 26 radio alphabets, and 100 common words	15 in total. by intelligibility: 4 very low, 3 low, 3 mild, 5 high	Average WRA of 67%

which compromises any results reported. On the other hand, we did not notice any notable approach to split the training and testing data on TORGO, which makes it challenging to compare and benchmark on this dataset.

Furthermore, it is crucial for researchers to indicate the appropriate metrics to measure, depending on whether the DSR focuses on isolated speech or connected speech. We noticed instances where incorrect metrics were used before. For example, some studies reported WER on UA-Speech, despite the fact that the corpus primarily consists of isolated speech samples uttering single words. WER is not a proper performance indicator in this context. Instead, Word Recognition Rate (WRR), also known as Word Recognition Accuracy (WRA), should have been used. In the case of sequence-to-sequence DSRs targeting UA-Speech, Character Error Rate (CER) could be measured in addition to WRR to assess performance accurately.

VII. CONCLUSION

This paper presents a survey and comparison of three categories of dysarthric ASR systems. Previous efforts using HMM and ANN models have shown notable advancements; however, they struggled to fully address the unique characteristics of dysarthric speech. More recent approaches utilizing end-to-end architectures and transformer models have exhibited promising outcomes. Nevertheless, enhancing the performance of ASR systems for dysarthric speech remains a substantial challenge due to the considerable variability both between and within speakers and the lack of publicly available large dysarthric speech corpora.

Future research directions involve incorporating auxiliary tasks within the multi-task framework, improving data augmentation methods, and devising more effective sampling schemes. These approaches aim to tackle the inherent

variability of dysarthric speech and enhance the robustness of DSR systems. Accomplishing these objectives has the potential to significantly advance the capabilities of DSR systems, making them more inclusive and valuable for a wider range of users. Moreover, procedural guidelines are needed to define proper evaluation methods, enabling direct comparison between DSR systems.

REFERENCES

[1] K. Bharti and P. K. Das, "A Survey on ASR Systems for Dysarthric Speech," *AIST 2022 - 4th International Conference on Artificial Intelligence and Speech Technology*, 2022, doi: 10.1109/AIST55798.2022.10065162.

[2] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *Augmentative and Alternative Communication*, vol. 16, no. 1, pp. 48–60, Jan. 2000, doi: 10.1080/07434610012331278904.

[3] M. Rughani and D. Shivakrishna, "A Review on Dysarthric Speech Recognition," *International Journal of Advanced Networking & Applications*, 2014.

[4] H. P. Rowe, S. E. Gutz, M. F. Maffei, K. Tomanek, and J. R. Green, "Characterizing Dysarthria Diversity for Automatic Speech Recognition: A Tutorial From the Clinical Perspective," *Front Comput Sci*, vol. 4, Apr. 2022, doi: 10.3389/fcomp.2022.770210.

[5] H. Kim *et al.*, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[6] S. D. S. Barreto and K. Z. Ortiz, "Speech Intelligibility in Dysarthrias: Influence of Utterance Length," *Folia Phoniatrica et Logopaedica*, vol. 72, no. 3, pp. 202–210, Apr. 2020, doi: 10.1159/000497178.

[7] S. Liu *et al.*, "Recent Progress in the CUHK Dysarthric Speech Recognition System," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, 2021, doi: 10.1109/TASLP.2021.3091805.

[8] 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, pp. 523–541, 2012.

[9] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The Nemours database of dysarthric speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 1996, pp. 1962–1965.

[10] B. MacDonald *et al.*, "Disordered speech data collection: Lessons learned at 1 million utterances from project euphonia," 2021.

[11] J. R. Deller Jr, D. Hsu, and L. J. Ferrier, "On the use of Hidden Markov Modelling for recognition of dysarthric speech," *Comput Methods Programs Biomed*, vol. 35, no. 2, pp. 125–139, 1991.

[12] N. Rajeswari and S. Chandrakala, "Generative model-driven feature learning for dysarthric speech recognition," *Biocybern Biomed Eng*, vol. 36, no. 4, pp. 553–561, 2016.

[13] P. D. Polur and G. E. Miller, "Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals," *Med Eng Phys*, vol. 28, no. 8, pp. 741–748, 2006.

[14] S. S. Tirumala and S. R. Shahamiri, "A deep autoencoder approach for speaker identification," in *9th International Conference on Signal Processing Systems (ICSPS)*, Auckland, New Zealand: ACM, Nov. 2017, pp. 175–179. doi: 10.1145/3163080.3163097.

[15] S. R. Shahamiri, W. M. N. W. Kadir, and S. Ibrahim, "A single-network ANN-based oracle to verify logical software modules," in *ICSTE 2010 - 2010 2nd International Conference on Software Technology and Engineering, Proceedings*, 2010. doi: 10.1109/ICSTE.2010.5608808.

[16] G. Jayaram and K. Abdelhamied, "Experiments in dysarthric speech recognition using artificial neural networks," *J Rehabil Res Dev*, vol. 32, p. 162, 1995.

[17] S. R. Shahamiri, "Neural network-based multi-view enhanced multi-learner active learning: theory and experiments," *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1–21, Jul. 2021, doi: 10.1080/0952813X.2021.1948921.

[18] S. Selva Nidhyananthan, R. Shantha Selva kumari, and V. Shenbagalakshmi, "Assessment of dysarthric speech using Elman back propagation network (recurrent network) for speech recognition," *Int J Speech Technol*, vol. 19, pp. 577–583, 2016.

[19] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Convolutive bottleneck network with dropout for dysarthric speech recognition," *Transactions on Machine Learning and Artificial Intelligence*, vol. 2, pp. 1–15, 2014.

[20] C. España-Bonet and J. A. R. Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, Nov. 2016, pp. 97–107. doi: 10.1007/978-3-319-49169-1_10.

[21] B. F. Zaidi, S. A. Selouani, M. Boudraa, and M. Sidi Yakoub, "Deep neural network architectures for dysarthric speech analysis and recognition," *Neural Comput Appl*, vol. 33, pp. 9089–9108, 2021.

[22] N. Ding *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nat Mach Intell*, vol. 5, no. 3, 2023, doi: 10.1038/s42256-023-00626-4.

[23] Z. D. Champiri, S. S. B. Salim, and S. R. Shahamiri, "The Role of Context for Recommendations in Digital Libraries," *International Journal of Social Science and Humanity*, vol. 5, no. 11, 2015, doi: 10.7763/ijssh.2015.v5.585.

[24] 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, doi: 10.1109/TNSRE.2021.3076778.

[25] A. Hu, D. Phadnis, and S. R. Shahamiri, "Generating synthetic dysarthric speech to overcome dysarthria acoustic data scarcity," *J Ambient Intell Humaniz Comput*, 2021, doi: 10.1007/s12652-021-03542-w.

[26] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Interspeech 2019*, 2019. doi: 10.21437/interspeech.2019-1789.

[27] R. Doshi *et al.*, "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)*, 2021, pp. 6988–6992. doi: 10.1109/ICASSP39728.2021.9414644.

[28] Z. et al. Chen, "Conformer Parrotron: A faster and stronger end-to-end speech conversion and recognition model for atypical speech," in *Interspeech 2021*, 2021. doi: 10.21437/interspeech.2021-676.

[29] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context," *ArXiv*, 2019.

[30] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[31] L. Wu, D. Zong, S. Sun, and J. Zhao, "A Sequential Contrastive Learning Framework for Robust Dysarthric Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7303–7307. doi: 10.1109/ICASSP39728.2021.9415017.

[32] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, and E. D. et al. Cubuk, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019, pp. 2613–2617.

[33] C. Ding, S. Sun, and J. Zhao, "Multi-Task Transformer with Input Feature Reconstruction for Dysarthric Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7318–7322. doi: 10.1109/ICASSP39728.2021.9414614.

[34] S. R. Shahamiri, V. Lal, and D. Shah, "Dysarthric Speech Transformer: A Sequence-to-Sequence Dysarthric Speech Recognition System," *IEEE Trans Neural Syst Rehabil Eng*, vol. 31, pp. 3407–3416, 2023, doi: 10.1109/TNSRE.2023.3307020.

[35] C. Espana-Bonet and J. A. R. Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in *Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH 2016, Lisbon, Portugal, November 23-25, 2016, Proceedings 3*, 2016, pp. 97–107.

[36] M. Sidi Yakoub, S. Selouani, B.-F. Zaidi, and A. Bouchair, "Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network," *EURASIP J Audio Speech Music Process*, vol. 2020, no. 1, pp. 1–7, 2020.