JKAU: Comp. IT. Sci., **Vol**. **13** No. 2, pp: 36 – 57 (2024 A.D.)
DOI: 10.4197/Comp.13-2.3

Recent Advances in Dysarthric Speech Recognition: Approaches and Datasets

Tahani Alrajhi[1,2], Mourad Ykhlef[1], and Ahmed Alsanad[1]
[1]*Department of Information Systems, King Saud University, Riyadh, Saudi Arabia*
[2]*Department of Information Science, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia*
taalrajhi@iau.edu.sa, ykhlef@ksu.edu.sa, aasanad@ksu.edu.sa

***Abstract*—Dysarthria is a neuromotor speech disorder that results from physical disability and limits speech intelligibility. Dysarthric speakers can make use of speech recognition systems to help them communicate more effectively with others. This paper surveys the latest works conducted on dysarthric speech recognition that was carried out in a span of five years, specifically from 2018 until 2023. These works are categorized according to the approach that was followed to improve dysarthric speech recognition. The approaches include data augmentation, enhancement of dysarthric speech, speech and acoustic features, adaptation, and hybridization of multiple approaches.**

Keywords—Dysarthria; ASR; Dysarthric Speech Recognition; Automatic Speech Recognition; Speech Impairment

I. INTRODUCTION

Dysarthria is "a set of motor disorders resulting from general physical disabilities that limit speech intelligibility" [1]. Dysarthric speakers may also suffer from physical disabilities that might hinder them from communication through typing or using electronic devices and computers and hence, speech would be considered more convenient to them [2]. Therefore, a speech recognition system would be beneficial for them as it can take their speech as an input and convert it into text that can be used to communicate with others or with assistive technology.

Reviewing the literature shows that only a limited number of studies surveyed dysarthric speech recognition and discussed the efforts in this field. A study that aimed to review dysarthric speech recognition traced its development from 1990 to 2022 highlighting works conducted before and during deep learning era [3]. The authors summarized the development of ASR for dysarthric speech according to four aspects: acoustic models, acoustic features, language-lexical models, and End-to-End ASR. Other studies reviewed dysarthric speech recognition through comparing machine learning with deep learning techniques [4] or from a clinical perspective [5]. The significance of the present work stems from its focus on the recent advances in dysarthric speech recognition as it covers the works conducted recently in a span of five years, specifically from 2018 until 2023. In addition, it provides a comprehensive overview of dysarthric speech recognition research with a focus on the approach used to improve the recognition accuracy rather than the model used. Approaches mainly include data augmentation, enhancement of dysarthric speech, speech and acoustic features improvement, adaptation to dysarthric speakers or speech, and a hybrid of these approaches. Consequently, this paper aims to complement existing literature by providing a recent overview of different approaches that can be utilized to enhance dysarthric speech recognition along with the techniques used, word error rates achieved and the available datasets. This in turn can reflect on future research and help researchers achieve better results using this holistic view.

II. APPROACHES TO DYSARTHRIC SPEECH RECOGNITION

A. *Data Augmentation*

Data augmentation is a process used to artificially generate additional training data to support automatic speech recognition [6] [7]. This could be accomplished through many strategies such as time warping, time masking, frequency masking, synthesizing dysarthric speech or using multiple databases [7] [8] [9] [10]. The rationale behind this approach is to support the scarcity of dysarthric speech datasets that are used for training as it is

difficult to collect large datasets from dysarthric speakers because they struggle to produce speech due to their health conditions.

Synthesis of dysarthric speech has been employed in several works to augment data. Vachhani et. al. explored the effect of data augmentation on dysarthric ASR [6]. The authors produced synthetic dysarthric speech through performing temporal and speed modifications on normal speech. They also used a Random Forest Classifier (RFC) that was trained on actual dysarthric speech to classify synthetically generated dysarthric speech according to severity levels. After that, a DNN-HMM based ASR system was trained using normal speech and augmented dysarthric speech and then evaluated using UASpeech corpus. The results showed that tempo-based and speed-based data augmentation led to an absolute improvement of 4.24% and 2% respectively when compared to an ASR system trained only on normal speech.

Takashima et al. proposed an end-to-end ASR framework that is trained using multiple multilingual datasets instead of focusing only on one to augment data and overcome the scarcity of dysarthric speech data [9]. The first dataset contained speech data of Japanese persons with an articulation disorder resulting from athetoid cerebral palsy. The second contained speech data of non-Japanese persons with an articulation disorder. The third contained speech data of a physically unimpaired Japanese person. The main reason for using multiple datasets is that impaired speech data is limited and not easy to collect because of its large burden on users due to the strain put on the speech muscles. In addition, training the model on dysarthric speech from multiple languages can boost the training model and capture a better high-level representation. The framework consisted of two models: an acoustic model for dysarthric speech and a Language Model (LM) for each language regardless of dysarthria. The model was based on Listen, Attend, and Spell (LAS) with two listeners and two spellers. One listener is for dysarthric speech, and the other is for unimpaired speech. The spellers are one for Japanese and one for English where each one will get data according to language regardless of impairments. The results showed that the proposed model is promising and achieved a better character error rate (CER) when compared to the same model trained only on unimpaired speech or using impaired speech of one language.

Xiong et. al. studied speech tempo analysis at the phonetic level to reduce the mismatch between typical and atypical speech [11]. The authors non-linearly modified speech tempo and performed speech tempo analysis at the phonetic level using a forced alignment process from the traditional GMM-HMM ASR system. They considered two approaches. The first was to modify dysarthric speech into normal speech and use it as an input to an ASR system trained on normal speech. The second was to modify normal speech into dysarthric speech to augment data in personalized dysarthric ASR training. The authors found that the second approach was more effective and resulted in an absolute improvement of 7% in comparison to baseline speaker-dependent trained systems that are evaluated using UASpeech corpus especially for moderate to severe dysarthric speakers.

Misbullah et. al. employed time delay deep neural networks for dysarthric ASR and investigated its performance for this task [10]. The collected data included English and Mandarin dysarthric speech. The authors performed data augmentation to increase the dataset available to support the training process through changing audio speed to produce different data. In addition, they used the time-warp and Voice Track Length Normalization (VTLN) warp with different warping factors after performing speed augmentation. Moreover, they combined the English dysarthric training corpus they collected with normal speech from Common Voice dataset to increase the English training data as it was not enough for training. The results showed that well-tuned hyperparameters gave promising results and could lead to a stable network structure for English and Mandarin dysarthric speech. Also, data combination with normal speech and well-tuned hyperparameters could significantly improve the performance of ASR systems for dysarthric speakers.

Mariya Celin et. al. performed a two-level data augmentation through the employment of Virtual Microphone array synthesis (VM) that is followed by Multi-Resolution Feature Extraction (MRFE) in order to increase the training data [12]. In the first

step, the authors synthesized six virtual microphone array signals from the first microphone signal that was recorded using UASpeech corpus. This was attained through altering the phase parameter. In the second step, the authors used different window sizes to artificially produce multiple examples for a given utterance. This is because the features that are extracted from a single speech signal with different window sizes have different frequency resolutions. After that, and with the use of augmented speech data, the authors trained an isolated word hybrid DNN-HMM based ASR system using UASpeech corpus along with Tamil speech corpus which they have developed. The results showed a reduction in Word Error Rate (WER) up to 32.79% for low intelligible dysarthric speakers and up to 35.75% for very low intelligible dysarthric speakers.

The use of sequence discriminative training, specifically Lattice-Free Maximum Mutual Information (LF-MMI) was studied by Hermann and Doss to improve dysarthric ASR [13]. They employed frame subsampling and speed perturbation techniques to improve dysarthric ASR and augment the training data. Using these techniques with LF-MMI exhibited great results on the TORGO dataset. The average WER for isolated words was 42.9% and 25.9% for sentences.

Another work by Hermann and Doss attempted to develop a speech recognizer that can fit a wider audience and perform well for both dysarthric and control speakers [14]. They investigated the effect of the acoustic variability of dysarthric speech on speech recognition systems, and proposed a solution to mitigate this problem through combining multiple LF-MMI acoustic models that are trained on different subsets of speakers. The combination of the trained acoustic models was undertaken by computing the union of the decoding lattices with subsequent Minimum Bayes Risk (MBR) decoding. Speed perturbation was used as a form of data augmentation to add two additional copies of the training data in all the LF-MMI acoustic models. The results showed improvements for both control and dysarthric speech recognition.

Harvill et al. proposed a data augmentation method through synthesizing new words that are used to train a CTC- based ASR system and thus expand the vocabulary and increase the accuracy of the ASR system [15]. This was accomplished by using the available dysarthric speech to capture the vocal characteristics of a dysarthric speaker through a parallel voice conversion system and then synthesize dysarthric speech that can be used to augment data for training the ASR system. The results showed that their proposed method outperformed practical baselines.

Matsuzaka et al. used a Text-To-Speech (TTS) synthesis to augment data [8]. They trained a Deep Neural Network (DNN)-based TTS model using dysarthric speech recorded from one dysarthric speaker. Then, this trained TTS model was used to generate synthesized dysarthric speech. After that, both the dysarthric speaker recordings and the synthesized dysarthric speech were used to train the ASR system. The results showed an improvement in speech recognition error rate.

Soleymanpour et al. worked also on the synthesis of dysarthric speech to improve the training of DNN-HMM ASR system [16]. They improved a multi-speaker end-to-end text-to-speech (TTS) system that they used to synthesize dysarthric speech. This improvement was achieved by adding a dysarthria severity level coefficient and a pause insertion model in order to be able to synthesize dysarthric speech for varying severity levels. Moreover, they used other prosody coefficients such as energy, pitch, and duration. The results indicated that synthesis of dysarthric speech for training has a significant impact on ASR systems and that the use of prosody coefficients helped in reducing WER.

Yue et al. proposed a multi-stream model which consists of convolutional and recurrent layers. They used raw magnitude spectra of the source and filter components [17]. The authors separated the vocal tract and excitation elements through cepstral processing and recombined them using multi-stream CNNs. They also used speed perturbation to augment data. The authors showed that multi-stream processing makes use of the two information streams, the vocal tract and excitation, and assists in normalizing speaking style and speaking attributes. This can be beneficial in handling dysarthric speech which is known for its large inter-speaker and intra-speaker variability. The results showed that the proposed model reduced the absolute WER by up to 1.7% compared to MFCC baseline.

Afterwards, the authors worked with Heidi Christensen and Jon Barker to further explore the effectiveness of using raw waveform acoustic modeling, which is task-specific, instead of hand-crafted features [18]. This was done to include all task-relevant information and avoid discarding useful information. They also examined the parametric CNNs that require less training data, in comparison to nonparametric CNNs, which can compensate for the scarcity of dysarthric data. In addition, the authors studied the effectiveness of data augmentation and multi-stream acoustic modeling through the combination of parametric and non-parametric CNNs that are fed by raw waveform and hand-crafted features. They used speed perturbation to increase the amount of training data by three folds: slower, original, and faster. The results showed the effectiveness of using parametric models with data augmentation to deal with the data scarcity problem. Moreover, the parametric CNNs significantly outperformed the non-parametric CNNs in the experiments conducted using the TORGO dataset. Also, multi-stream acoustic modeling was able to further improve the model's performance. Table I summarizes the reviewed papers in data augmentation approach, the techniques used, and error rates. Error rates are calculated using words, characters, or phonemes i.e., Word Error Rate (WER), Character Error Rate (CER), or Phoneme Error Rate (PER).

TABLE I. SUMMARY OF DATA AUGMENTATION APPROACH PAPERS

Ref.	Year	Method	Techniques	Datasets	Error Rate
[6]	2018	Speed & temporal augmentation	-DNN-HMM ASR -RFC	UASpeech	Lowest overall WER 24.82% (using tempo augmentation)
[9]	2019	Multiple multilingual databases	LAS model (2 listeners & 2 spellers)	- Dysarthric speech (2 Japanese speakers + TORGO (English)) -Non-dysarthric (ATR Japanese speech database)	Avg. top-1 error 26.6% (CER)* Avg. top-3 error 22.05% (CER)*
[11]	2019	Speech tempo augmentation	-GMM-HMM ASR -Hybrid DNN-HMM with TDNN	UASpeech	Lowest avg. WER 30% (augmentation+ speaker-based speech tempo adjustment)
[10]	2020	Speed augmentation, VTLN warp & time warp + data combination (normal & dysarthric)	-Time delay deep neural network factorization (TDNN-F) -VTLN warp -Time warp	-English and Mandarin dysarthric speech -Common voice English dataset	Lowest WER for English 4.30% (using combined dysarthric +common voice) Lowest WER for Mandarin 6.08% (using dysarthric speech only)
[12]	2020	VM-MRFE	Hybrid DNN-HMM based ASR system	-UASpeech -Tamil corpus by authors	Lowest avg. WER for English is 18.33% (using MRFE) * Lowest avg. WER for Tamil is 30.15% (using MRFE) *
[13]	2020	Speed perturbation	subspace GMM - HMM/DNN (TDNN model trained with LF-MMI objective function)	-Pre-trained on LibriSpeech (only for HMM/DNN models) -TORGO (training and testing)	Lowest avg. WER 42.9% (isolated words) 25.9% (sentences) using LF-MMI with 10 ms frame shift
[14]	2021	Speed perturbation	-LF-MMI acoustic models -subspace GMM -DTW distance	-UASpeech + TORGO (both control and dysarthric speakers)	Lowest avg. WER 42.2% for isolated words (combination of 3 models) Lowest avg.WER 25.9% for sentences (model trained on both dysarthric +control speech)
[15]	2021	Synthesis of dysarthric speech	Attention-Based Voice Conversion - DTW (for alignment) -CTC-based ASR system	UASpeech	Lowest avg.WER 29.3% (using attention +LM)
[8]	2022	Synthesis of dysarthric speech	-DNN based TTS model -CTC-based ASR model	Text from ATR dataset – recorded speech by authors - JSUT corpus (non-dysarthric)	Lowest PER 46.22 (using recoded data + synthetic data)

[16]	2022	Synthesis of dysarthric speech	-modified FastSpeech2 (multi talker TTS as a voice conversion system) -DNN-HMM ASR model	TORGO	Lowest avg.WER 39.2% (Using the second experiment)
[17]	2022	Speed perturbation	CNN with recurrent layers	TORGO	Lowest avg.WER 40.6% (speed perturbation +Mag feature)
[18]	2022	Speed perturbation	-Parametric CNNs -Non-parametric CNNs	TORGO	Lowest WER 33.1% (FBank(CNN)+Raw(Parz) with Concat-2)

* Indicates that the average WER was not directly provided in the paper and therefore it was calculated using the formula: Avg. = sum of error rates reported/number of error rates.

B. Enhancement of dysarthric speech

Enhancement of dysarthric speech is another approach that works on dysarthric speech signals to make it more intelligible and thus can be recognized by listeners and automatic speech recognition systems [19]. Bhat et al. investigated the enhancement of dysarthric speech features to match the features of normal speech and thus enable ASR systems to recognize it [20]. A Time-Delay Neural Network based Denoising Autoencoder (TDNN-DAE) was used in this study to enhance dysarthric speech. Then, a DNN-HMM ASR system was used to recognize the enhanced speech. The authors evaluated the proposed method for speaker-independent and speaker-adaptive based ASR systems. The results revealed that the enhancement of dysarthric speech led to an absolute improvement of 13% in the performance of the speaker-independent ASR system and 3% in the performance of the speaker-adaptive ASR system. Moreover, the analysis showed that the ASR performance significantly improved at all severity levels of dysarthria.

Wang et al. suggested the use of voice conversion method for a dysarthric speech reconstruction task [21]. The proposed method included three steps. First, a Text-To-Speech system (TTS) was trained with transcribed normal speech. Second, the text-encoder of this trained TTS system (teacher) was used to train a speech-encoder (student) to extract linguistic representations from transcribed dysarthric speech through a cross-modal knowledge distillation process (teacher-student framework). Third, the trained speech-encoder was concatenated with the attention and the decoder of the TTS system in the first step to carry out the dysarthric speech reconstruction task through mapping dysarthric speech to normal speech. The findings indicate that the proposed method significantly improved speech quality, generating a highly natural and intelligible speech, especially for speakers with severe dysarthria. A comparison between the original dysarthric speech and the reconstructed speech revealed a reduction in WER by 35.4% and 48.7% for speakers with low and very low intelligibility levels, respectively.

Sidi Yakoub et al. proposed a speech enhancement technique that aims to improve the quality of dysarthric speech as a preprocessing step prior to its recognition [19]. The authors used Empirical Mode Decomposition and Hurst-based mode selection (EMDH) as an enhancement technique with deep learning using convolutional neural networks. First, the dysarthric speech is enhanced using EMDH. Then, the Mel-frequency cepstral coefficients are extracted and used as input to the CNN recognizer. The results indicated that the proposed approach of using EMDH-CNN increased the accuracy by 20.72% when compared to HMM-GMMs baseline systems. Also, it increased the accuracy by 9.95% when compared to a CNN without a prior enhancement step.

Another work that tackled the enhancement of dysarthric speech as a first step prior to the recognition process was conducted by Rajeswari et al. [22]. The enhancement was carried out using Variational Mode Decomposition (VMD) and wavelet thresholding. Then, the enhanced and reconstructed signals were fed to CNNs. This, in turn, enabled these networks to learn the specific features of dysarthric speech and, therefore, the speech model can support dysarthric speech recognition. The results showed that this method improved the recognition accuracy when compared to currently used methods based on generative models and artificial neural networks. Moreover, it was able to achieve an average accuracy of 95.95% with VMD

based enhancement and 91.80% without enhancement.

Ding et al. proposed a multi-task Transformer for dysarthric ASR [23]. This Transformer performs two tasks: an auxiliary task that involves input feature reconstruction, and a main task for dysarthric speech recognition. The auxiliary task attempts to perform two reconstruction methods: a cross-domain reconstruction which reconstructs clear speech features from dysarthric speech and an intra-domain reconstruction that reconstructs clear speech features from corrupted normal speech. Both the auxiliary task and the main task of the Transformer share the same encoder network. Moreover, the authors designed an adaptive rebalance sampling scheme to optimize the utterance sampling frequency. This was done to mitigate the imbalance distribution of dysarthria datasets. The results showed that the multi-task Transformer outperformed other baseline systems across all dysarthric speakers.

Prananta investigated in his master thesis some methods to improve the intelligibility of dysarthric speech for ASR systems [24]. Three experiments were conducted. The first one inspected the use of Cycle-consistent Generative Adversarial Network for Voice Conversion (CycleGan-VC) to convert dysarthric speech to normal speech. The second aimed at training CycleGan-VC with parallel data processing and Dynamic Time Warping (DTW) as a speech enhancement technique to improve the performance of the proposed method. The third experiment tackled the adjustment of speech rate of dysarthric speech using Time Stretching (TS) to improve the performance of the ASR system. The findings indicated that the use of CycleGan-VC did not improve the performance of dysarthric ASR in terms of Phone Error Rate (PER). Moreover, training CycleGan-VC with DTW and parallel data provided minor improvements and did not improve much compared to dysarthric speech baseline. However, using time stretching for the adjustment of speech rate of dysarthric speech improved the ASR performance by 19.8% for female speakers and by 5.5% for male speakers.

Prananta also worked with Halpern, Feng, and Scharenborg on a comparison between several Generative Adversarial Network-based (GAN) voice conversion methods [25]. The authors investigated the effectiveness of these methods on the enhancement of dysarthric speech in order to improve dysarthric ASR. A rigorous ablation study was carried out as an attempt to find the most effective solution to enhance dysarthric speech recognition. The results showed that signal processing methods that are straightforward like time stretching and denoising gave comparable results to state-of-the-art GAN-based voice conversion methods using a phoneme recognition task. Moreover, the researchers proposed the application of MaskCycleGAN-VC that is used for voice conversion on time stretched speech as a solution to improve the recognition. This combination provided results that are somewhat better than pure time stretching for dysarthric speakers with mid to high severity. Table II summarizes the reviewed papers in enhancement approach, the techniques used, and error rates.

TABLE II. SUMMARY OF ENHANCEMENT APPROACH PAPERS

Ref.	Year	Method	Techniques	Datasets	Error Rate
[20]	2018	Features enhancement using deep denoising autoencoders	-TDNN-DAE (enhance dysarthric speech) -DNN-HMM ASR system	UASpeech	Lowest WER 18.54% using best configuration ***
[21]	2020	Dysarthric speech reconstruction using knowledge distillation (KD)	-Tacotron (TTS model) -WaveRNN (synthesize waveform) E2E dysarthric speech reconstruction system	-LJSpeech dataset (normal speech) -UASpeech (dysarthric & normal speech)	Avg. WER 33% *
[19]	2020	Spectral subtraction, Wiener filtering and EMDH	-EMDH (enhancement technique) -CNN system	Nemours	Lowest global WER 35.14% (using EMDH-CNN with10-fold cross-validation) **
[22]	2022	Denoising	-VMD & wavelet thresholding (enhancement technique)	UASpeech	Overall avg. WER 4.05% (VMD+CNN) **

			-CNN model		
[23]	2021	Speech feature reconstruction (intra & cross-domain reconstruction)	-Multi-task Transformer (hybrid CTC/attention E2E ASR architecture)	-LibriSpeech -TORGO	Lowest avg. WER 15.88% (proposed model with rebalance sampling)*
[24]	2021	DTW, parallel data processing, TS, denoising, VC	-CycleGAN-VC + MaskCycleGAN-VC (voice conversion models) -Pre-trained HMM-based ASR	-UASpeech -Model pre-trained on TIMIT	Lowest avg. PER for males 70.3% & for females 76.1% (using Dysarthric &TS)
[25]	2022	DTW, parallel data processing, TS, 2-step adversarial loss, denoising, VC)	-GAN architectures (CycleGAN-VC, DiscoGAN, MaskCycleGAN-VC) -Pre-trained HMM-based ASR	-UASpeech - Model pre-trained on TIMIT	Lowest avg. PER: males 66.4% (dysarthric+TS) - females 73.2% (MaskCycleGAN-VC+TS)

* Indicates that the average WER was not directly provided in the paper and therefore it was calculated using the formula: Avg = sum of error rates reported/number of error rates.

** Indicates that the results were given in terms of recognition accuracy, and it was converted to error rate for comparison reasons using the formula: Error rate = 100-accuracy.

*** Trained on normal +TDNN-DAE enhanced dysarthric speech & tested on Temporally adapted+TDNN-DAE enhanced dysarthric speech in speaker adaptive scenario.

C. Speech and acoustic features

Speech and acoustic features can be utilized, improved, or normalized to support dysarthric ASR as dysarthric speech exhibits large inter-speaker and intra-speaker variabilities [7] [26]. Mathew et al. conducted a study to compare different feature extraction methods and investigated which features are most suitable to dysarthric ASR tasks [27]. The experiment was carried out using an HMM-based recognition system. The features that were considered in the comparison are Perceptual Linear Prediction (PLP), Mel-Frequency Cepstral Coefficients (MFCC), reflection coefficients and filter bank feature sets. The TORGO dataset was used to study and compare the performance of these features. The results showed that PLP is the most suitable feature for a dysarthric ASR task. Also, the researchers found that MFCC and PLP were able to provide better results than reflection coefficients and filter bank feature sets.

Kim et al. studied the effect of using Convolutional Long Short-Term Memory Recurrent Neural Networks (CLSTM-RNNs) on dysarthric ASR [28]. They hypothesized that using CLSTM-RNNs can capture the distinct characteristics of dysarthric speech where CNNs can be used to extract effective local features and LSTM-RNNs can be used to model features temporal dependencies. The experiment included four types of CLSTM-RNNs: Time domain CNN with LSTM-RNN (T-CLSTM-RNN), Frequency domain CNN with LSTM-RNN (F-CLSTM-RNN), Time Frequency CNN with LSTM-RNN (TF-CLSTM-RNN), and Parallel Time Frequency CNN with LSTM-RNN (PTF-CLSTM-RNN). The results showed that CLSTM-RNNs were able to provide a substantial improvement when compared to using only CNN or only LSTM-RNN. Of the four types in the experiment, TF-CLSTM-RNN achieved the best overall performance.

Hu et al. presented two dysarthric speech recognition systems for Cantonese and English [29]. They used a Gated Neural Network (GNN) modeling technique for both systems to integrate acoustic features with visual features and optionally with prosody features that are based on pitch. A novel Bayesian GNN Audio-Visual Speech Recognition (AVSR) architecture was employed in the English recognition task to obtain a robust integration of acoustic and visual features. As for the Cantonese recognition task, they used pitch features to assist acoustic features. The performance of the proposed systems was compared with Google speech recognition API and human recognition results. The findings showed that both systems outperformed Google's speech recognition API, and the English system outperformed human recognition for all speakers.

Zaidi et al. investigated the concatenation of several variants of Jitter and Shimmer with Mel-Frequency Cepstral Coefficients (MFCC) and Power Normalized Cepstral Coefficients (PNCC) which are speech parameterization coefficients to improve dysarthric ASR [26]. Jitter represents a quantification of small deviations of true periodicity i.e., cycle-to-cycle F0 perturbation while Shimmer is the analogue of Jitter and is calculated by the amplitude A0 contour instead of F0 contour. The authors developed an automatic acknowledgment of continuous

pathological speech system to help dysarthric speakers and help doctors make a primary diagnosis. The results indicated that the combination of PNCCs coefficients with the Shimmer Ampl PQ3 classical Baken or with the Shimmer CV yielded the best results compared to their basic system.

Another work by Zaidi et al. investigated DNNs ability to improve dysarthric ASR using CNNs and LSTM neural networks [30]. First, they compared the use of different input features with dysarthric speech recognition systems. These features were Mel-Frequency Spectral Coefficients (MFSCs), MFCCs, and PLPs. Then, they compared the performance of CNN and LSTM architectures with HMMs and GMMs models to find the best dysarthric speech recognizer. The findings showed that the best result was achieved by a speaker-dependent CNN using PLP with an accuracy of 82%. This result constitutes an improvement of 32% and 11% when compared to GMM-HMM and LSTM based systems' performance, respectively.

Chandrakala presented a review and an analysis of different approaches including discriminative, generative, hybrid model-based approaches and unsupervised approaches [31]. The author also proposed generative model-driven feature learning approaches for dysarthric ASR. She compared the results of the proposed model with two different types of discriminative classifiers: Transition Embedding Support Vector Machine (TE-SVM) and Likelihood Embedding Support Vector Machine (LE-SVM). The results showed that even though TE-SVM gave a good performance due to the increase in the number of features, it was not able to improve the discrimination when it was compared to LE-SVM. This is because LE-SVM was able to capture discriminative information that can lead to higher accuracy. Moreover, the effective fixed dimensional representation that can be formed using log likelihood probabilities provided by HMMs gave better performance.

Hernandez et al. explored the effectiveness of using Wav2Vec, Hubert, and multilingual XLSR self-supervised speech representations as features for training an acoustic model for dysarthric speech recognition [32]. They used three corpora representing different types of dysarthria from different languages: UASpeech (English), PC-GITA (Spanish), and EasyCall (Italian). The findings showed that using the extracted features from the multilingual XLSR model provided the lowest WERs for all datasets: English, Italian, and Spanish. It was observed that the features extracted from the multilingual model XLSR were able to provide lower WERs in comparison to other models although these models were trained on larger amounts of data but from English only. This can be attributed to the fact that a multilingual model contains more variations and thus is more suitable for dysarthric speech known for its variation as well. Table III summarizes the reviewed papers in speech and acoustic features approach, the techniques used, and error rates.

TABLE III. SUMMARY OF SPEECH AND ACOUSTIC FEATURES APPROACH PAPERS

Ref.	Year	Method	Techniques	Datasets	Error Rate
[27]	2018	MFCC, PLP, filter bank & reflection coefficients	HMM-based recognition system (isolated word recognizer)	TORGO	Lowest WER 36.73% (using PLP) **
[28]	2018	Features in spectral, temporal, & spectro-temporal domains	HMM-based dysarthric ASR systems (GMM-HMM, DNN-HMM, CNN-HMM & CLSTM-RNN-HMM)	9 dysarthric patients	Lowest avg. PER 30.6% (average of all testing sessions using TF-CLSTM-RNN)
[29]	2019	Acoustic features (Mel-scale log filter bank (FBK)), visual features & prosody based on pitch features	-GNN ASR (Cantonese system) -Bayesian GNN AVSR (English system)	-UASpeech (English) -CUDYS (Cantonese)	Avg. WER 31.2% (English)* Avg. CER 39.33% (Cantonese)*
[26]	2019	MFCC, PNCC, JITTER & SHIMMER coefficients	Hidden Models of Markov (HMM) (Basic system)	Nemours	Lowest WER 51.81% (with PNCC_0_SHIMMER CV) **

[30]	2021	MFCCs, MFSCs, PLPs	-CNN system -LSTM system	Nemours (dysarthric speech)	Lowest global PER 37.31% (CNN-based system with PLPs+ Poly1ReLU activation function) **
[31]	2020	MFCC	Left to right HMM with 2 types of discriminative classifiers TE-SVM & LE-SVM	UA-Speech (isolated words utterances)	Overall WER 12.09 % (using LE-SVM) **
[32]	2022	Self-supervised speech representations as features	-Acoustic models (E2E with a conformer encoder & transformer decoder) -wav2vec, Hubert & Multilingual XLSR (for speech representations)	-Pre-training: LibriVox (wav2vec, Hubert) & Common Voice, Babel & Multilingual Libri-Speech (XLSR) -UASpeech, PC-GITA (Spanish), EasyCall (Italian)	Lowest WER for English: XLSR model: 26.1% (speaker dependent) 47.3% (speaker independent) WER 12.9% (PC-GITA) & 16.5% (EasyCall) using XLSR-PD model

*Indicates that the average WER was not directly provided in the paper and therefore it was calculated using the formula: Avg = sum of error rates reported/number of error rates.

**indicates that the results were given in terms of recognition accuracy, and it was converted to error rate for comparison reasons using the formula: Error rate = 100-accuracy.

D. Adaptation

Researchers who employed adaptation mainly worked on the adaptation of the model itself, acoustic model or the pronunciation dictionary (lexicon) which are used in speech recognition to adapt the system to dysarthric speech/speakers [33] [34]. Moreover, models can be adapted to each dysarthric speaker to personalize the model to the targeted speaker [35]. In addition, Transfer Learning (TL) and fine-tuning can be utilized to adapt the weights of the model to improve dysarthric ASR [35] [36] [37]. This approach can be employed to alleviate the problem of data scarcity and to speed up the training process. In such works, researchers would transfer the learning of unimpaired speech or impaired speech of another language then fine-tune the model using the limited dysarthric speech available [38]. Others worked on domain or cross-domain adaptation and severity-based speaker adaptation [33] [39] [40].

The Chinese University of Hong Kong (CUHK) developed an ASR system for dysarthric speakers using the UASpeech database [40]. The authors constructed a number of DNN acoustic models. First, they constructed these models with a deep and stacked architecture. Then, they developed some of the models' advanced variants using LSTM-RNNs and Time Delayed Neural Networks (TDNNs). These variants were explored to study the benefit of longer-range context modeling. In addition, they utilized Learning Hidden Unit Contributions (LHUC) to perform speaker adaptation in order to handle inter-speaker variability. Moreover, to deal with feature extraction bottleneck with stacked DNN systems, the researchers used a semi-supervised Complementary Auto Encoder (CAE). Furthermore, to improve the recognition performance, cross domain adaptation was employed. Cross domain adaptation transforms the mean and variance of out of domain systems to be able to describe the distribution of dysarthric speech. This is done under the supervision of the recognition outputs from the proposed UASpeech stacked hybrid DNN system. The authors utilized two out of domain systems which were trained separately on broadcast news and switchboard data and adapted towards the UASpeech data and then adopted in a combination of six systems. The results showed that the final combined system provided the best performance with an overall word accuracy of 69.4% using a test set of 16 speakers.

Among the significant works in relation to ASR of non-standard speech is a project known as Euphonia which was conducted by a group of researchers from Google and Amyotrophic Lateral Sclerosis (ALS) therapy development institute [41]. In this study, two types of non-standard speech were included: dysarthric and accented. Two models were adopted: RNN Transducer (RNN-T) and LAS. These models were fine-tuned using the collected data to achieve state-of-the-art results for dysarthric and heavily accented speech. The results showed that fine-tuning on small amounts of non-standard speech can yield good results. The researchers indicated that through using approximately one hour of data, a personalized ASR model that outperforms cloud-based services can be created. Regarding WER, there was a 70% improvement over the base model for dysarthric speech and 35.1% improvement for accented speech. For the RNN-T model, fine-tuning only the first layer of the encoder and the joint layer, which usually happens within the first 5-10 minutes of training, achieved 90% of total relative improvement.

Takashima et al. employed transfer learning for dysarthric ASR [38]. This was accomplished through transferring two types of knowledge: the phonetic and linguistic characteristics of unimpaired Japanese speech, and the dysarthric characteristics of dysarthric non-Japanese speech. The rationale behind this transfer is to enable the use of deep learning with the limited dysarthric Japanese speech data available. After transferring this knowledge, the authors fine-tuned the model using Japanese dysarthric speech. The results showed that the use of additional speech data and transfer learning can significantly improve speech recognition performance. Nearly a year later, Takashima et al. presented another work on dysarthric ASR based on deep metric learning [36]. The motivation behind this work is that dysarthric speech considerably fluctuates even if the person was repeating the same sentence. Therefore, dysarthric speech tends to have great variation even within recognition classes. The proposed system learns an embedded representation where the distance between sentences within the same class is small and between sentences of different classes is large enabling the system to distinguish dysarthric speech easily. Moreover, the results showed that using deep metric learning consistently improved word-recognition accuracy. Furthermore, they evaluated the proposed system in combination with transfer learning using additional speech data from an unimpaired person which provided further improvement in performance.

Xiong et al. investigated the application of an improved transfer learning framework to personalized ASR models for dysarthric speakers [35]. A CNN-TDNN-F ASR acoustic model which was trained on source domain data was utilized to transfer and adapt the neural network weights using the limited data from dysarthric speakers in the target domain. The evaluation was based on UASpeech. The findings indicated that linear weights in neural layers played a significant role in improving dysarthric speech modeling. In comparison to speaker-dependent training, the proposed model achieved an average of 11.6% relative recognition improvement. Moreover, it achieved an average of 7.6% relative recognition improvement when compared to data combination training. Also, further recognition performance improvement that reached 2% when compared to transfer learning baseline was gained for speakers with moderate to severe dysarthria. This was accomplished through a selection of utterance-based data from the source domain instead of speaker-based data selection resulting in more accurate selection of the most beneficial data from the source domain. The authors offered two ways to do that; using incremental transfer learning through constructing an intermediate domain or increasing the training pool of the target domain. Based on the results, incremental learning outperformed data combination except for very severe dysarthric speech.

A two-step acoustic model adaptation approach for dysarthric ASR was proposed by Takashima et al. [37]. The researchers utilized transfer learning with adaptation to improve the recognition of dysarthric speech. They focused on athetoid cerebral palsy that causes involuntary muscle movements. The motivation behind using a two-step adaptation approach stems from the fact that dysarthric speakers generally have different speaking styles when compared to non-dysarthric speakers. Also, each dysarthric speaker can benefit from this adaptation approach due to their unique differences and causes of disability. Therefore, having two steps can assist in achieving better results. The first step is used to train a speaker-independent non-dysarthric model using dysarthric speech from many speakers to get the general characteristics of dysarthric speech. This was done using a baseline model that was pre-trained on non-dysarthric speech. Then, and by utilizing transfer learning, they retrained the model on the dysarthric speech of multiple speakers to get a speaker independent dysarthric model. The second step uses the resulting speaker-independent dysarthric model and trains it on dysarthric speech of the target speaker to adapt to that specific speaker. The results gained through these two steps are better than adapting the non-dysarthric model immediately to the targeted speaker in one step.

Wang et al. proposed a way to improve dysarthric ASR through transfer learning and model re-initialization [42]. Instead of directly fine-tuning a pre-trained base model for dysarthric speech, the authors suggested reinitializing the base model via meta-learning. They explained that the mismatching nature of statistic distribution between dysarthric and normal speech can limit the adaptation performance

of the base model. Thus, a meta-learning model that reinitializes the base model to learn dysarthric speech knowledge can adapt faster to unseen dysarthric speakers. The Model-Agnostic Meta Learning (MAML) and Reptile algorithms were utilized and extended to meta update the base model through a repeated simulating adaptation to different dysarthric speakers. The findings showed that the enhanced model performed better and adapted faster to unseen dysarthric speech. Using UASpeech, the best model was able to achieve a reduction in WER of 54.2% compared to the base model without fine-tuning. When compared to the base model that was directly fine-tuned, the proposed model achieved 7.6% relative WER reduction. These results are comparable to those of the state-of-the-art hybrid DNN-HMM model.

Al Qatab et al. employed several adaptation techniques to determine the adequate amount of adaptation data needed for speaker adaptation [33]. Due to the difficulty of generating speech by dysarthric speakers, identifying a saturation point where additional data will not result in an increase in the recognition accuracy can save their efforts and time. The authors investigated the use of two adaptation techniques: Maximum Likelihood Linear Regression (MLLR) and Maximum A Posterior (MAP). They experimented with each technique individually and with a combination of both in different sequences; MAP+MLLR sequence and MLLR+MAP sequence. Linear regression between the recognition accuracy and the data size was used to determine the saturation point. Moreover, the adaptation and test data were categorized according to severity level to yield a severity-based speaker adaptation. The results showed that the saturation point of MAP is lower in general than MLLR when used individually which means that more adaptation data is needed by MLLR to reach the lowest WER. Also, the sequence MLLR+MAP increases the effectiveness of the adaptation as the accuracy increases with each additional adaptation data. This shows that when adaptation techniques are combined, the order of the sequence affects the saturation point.

Sawa et al. proposed a two-step method to adapt the pronunciation dictionary to improve dysarthric ASR [34]. First, they carried out a phoneme recognition task using the target speaker's speech to identify the actual pronunciation of words by the dysarthric speaker and use this information later to perform a pronunciation analysis. Second, they extracted rules based on analyzing misrecognition patterns found in the first step and then adapted the dictionary by adding these pronunciations to it. After that, the adapted dictionary was used to train a speaker-dependent model for the targeted dysarthric speaker as well as to recognize their speech. The evaluation was conducted on a large vocabulary continuous speech recognition task. The results showed that the adapted dictionary was able to decrease the WER, and that consonants (mainly unvoiced) tend to be misrecognized. Table IV summarizes the reviewed papers in adaptation approach, the techniques used, and error rates.

TABLE IV. SUMMARY OF ADAPTATION APPROACH PAPERS

Ref.	Year	Method	Techniques	Datasets	Error Rate
[40]	2018	Speaker adaptation & cross-domain adaptation	-GMM-HMM, DNN, TDNN, LSTM-RNN & Systems Combinations. -CAE (for feature extraction bottleneck) -LHUC (speaker adaptation) -Cross domain adaptation	-Switchboard & broadcast news dataset (out of domain) -UASpeech	Lowest WER 30.6 % (combination of six systems)**
[41]	2019	Transfer Learning (TL) & fine-tuning	E2E sequence-to-sequence models: Bidirectional RNN-T model LAS model	-Training: RNN-T (Google voice-search traffic), LAS (LibriSpeech) -Fine-tuning (recorded dysarthric speech)	Lowest avg. WER 20.9% for severe dysarthria &10.8% for mild dysarthria (Using RNN-T)
[38]	2019	Transfer learning	LAS model	- 5 Japanese dysarthric speakers -ATR (unimpaired Japanese) -TORGO (English dysarthric)	Lowest avg. PER 25.69% (trans. 3 – two-decoder LAS)
[36]	2020	Transfer learning	-GMM-HMM (base model) -DNN-based model with TL	- 5 Japanese dysarthric speakers -ATR dataset	Lowest avg. WER 10.21% (proposed method+ TL with updating last layer +ArcFace)**

[35]	2020	Transfer learning	-Hybrid DNN-HMM ASR (training) -CNN-TDNN-F ASR acoustic model (baseline of TL)	UASpeech	Lowest avg. WER 30.76% (SD → CTL)
[37]	2020	Transfer learning	LF-MMI (baseline model) (TDNN) layers	- 4 Japanese dysarthric speakers -ATR dataset -CSJ (Corpus of Spontaneous Japanese (non-dysarthric))	Lowest avg. WER 53.7% (2-step adaptation+ LR factor=1)*
[42]	2021	Transfer learning & re-initialization	-MAML & Reptile algorithms for Meta-learning & Re-initialization. -Base models (LAS, QuartzNet)	-UASpeech (dysarthric speech) -LibriSpeech (normal speech)	Lowest overall WER 30.5% (base QuartzNet +Reptile re-initialization & adaptation)
[33]	2021	MLLR, MAP, MLLR+MAP, MAP+MLLR, severity-based speaker adaptation	-BSAM (Baseline Speech Acoustic Model) -Linear regression (for saturation point)	Training: normal speech (Wall Street Journal +TIMIT) + dysarthric (UASpeech,TORGO). Adaptation & testing: dysarthric (Nemours)	Lowest avg. WER 9.66% (MLLR + MAP sequence adaptation) *
[34]	2022	Pronunciation dictionary adaptation + transfer learning	-A hybrid CTC/attention model (E2E ASR for phoneme recognition task) -DNN-HMM hybrid model (word-recognition task) trained based on LF-MMI criterion	- 2 Japanese dysarthric speakers - CSJ (to construct baseline general dictionary, pre-train hybrid CTC/attention model, & train LM in word recognition model)	WER 46.32% (1st speaker) & 59.46% (2nd speaker) (dependent model, using the adapted dictionary)

*Indicates that the average WER was not directly provided in the paper and therefore it was calculated using the formula: Avg = sum of error rates reported/number of error rates.

**indicates that the results were given in terms of recognition accuracy, and it was converted to error rate for comparison reasons using the formula: Error rate = 100-accuracy.

E. Hybrid approaches in dysarthric speech recognition

Woszczyk et al. proposed the use of Domain Adversarial Neural Networks (DANN) for dysarthric ASR [43]. The proposed model is a speaker-independent speech recognition system that combines domain-invariant features with domain adversarial training to cope with the limitation of dysarthric speech data. The researchers used an End-to-End (E2E) CNN as a baseline system which takes raw audio as input to perform a classification task on ten spoken digits. UASpeech was used, and the results were compared to a speaker-dependent model, a speaker-adaptive model, and Multi-Task Learning (MTL) models. The speaker-adaptive model is the speaker-independent model that is fine-tuned on the speech of a certain speaker. The results showed that the proposed model outperformed the baseline CNN model by an absolute recognition rate of 12.18%. When compared to the speaker-adaptive model, it achieved comparable results. Though the model provided similar results to multi-task learning models with labeled dysarthric speech data, it performed better with unlabeled data.

Lin et al. proposed the use of E2E Automatic Speech Recognition (ASR) and Automatic Speech Attribute Transcription (ASAT) for patients with dysarthria [44]. The study presented a staged knowledge distillation method for dysarthric patients to deal with the low resource challenge in training ASR systems and provide an effective teacher-student learning approach. All the models were pre-trained using normal speech from LibriSpeech. As for retraining and due to limited dysarthric data, the researchers used all speech samples in TORGO (dysarthric and normal). They inspected the effectiveness of the proposed staged conditional teacher-student method together with four different systems. Different approaches were used in these systems to perform fine-tuning to adapt the system to dysarthric speech. The first system was fully fine-tuned on dysarthric speech from TORGO. The second adopted 100 hours of normal speech data from LibriSpeech for data augmentation and then the net was fully fine-tuned. In the third system, only the decoder was fine-tuned. As for the fourth system, the model was refactored where the layers of the decoder were shared and fine-tuned with speech perturbation. For evaluation, the models were tested using dysarthric speech from TORGO. The results showed that the accuracy of the proposed models significantly outperformed the traditional methods with a reduction of around 38.28% relative phone error rate and 48.33% relative attribute detection

error rate when compared to their baselines. Utilizing data augmentation through using additional normal speech and increasing TORGO dysarthric speech samples using speed perturbation yielded the lowest phone error rate of 29.84%. Moreover, the proposed method has potential to be used as a medical diagnostic aid and as a rehabilitation tool for patients with dysarthria.

Yue et al. investigated several methods to improve continuous dysarthric ASR systems [45]. They explored the effectiveness of using an AutoEncoder BottleNeck feature extractor (AE-BN) that is pre-trained on normal speech data from LibriSpeech and fine-tuned on dysarthric speech from TORGO. Furthermore, they studied the effect of combining acoustic features with the features extracted by AE-BN that was pre-trained on typical speech. Moreover, they employed speed perturbation to augment data during the training phase. Also, two multi-task optimization techniques were exploited: joint optimization and monophone regularization. The results showed that the addition of AE-BN features resulted in a reduction of WER by 2.63% compared to the baseline system. Applying joint optimization and monophone regularization techniques led to a further reduction of WER by 0.65% and 2.33%, respectively.

Xie et al. discussed the acoustic variability among dysarthric speakers which is difficult to be precisely modeled [46]. This issue motivated the researchers to present a Variational Auto-Encoder based Variability Encoder (VAEVE) that can be used to explicitly encode dysarthric speech variability. VAEVE reconstructs the input acoustic features using low dimensional latent variable and phoneme information so that the latent variable is forced to encode the variability information. The variability encodings are used as auxiliary features for acoustic modeling. The authors experimented with different systems to find the best solution to dysarthric speech recognition. In one of the systems, they applied the variability encodings to the system trained with data augmentation using speed perturbation. Then, LHUC adaptation is applied to test data. This system yielded the lowest overall WER. The results showed that applying variability encodings was able to improve the performance of the systems in comparison to the baseline system without them.

Lin et al. proposed a study to recognize Mandarin speech commands of dysarthric speakers using a CNN with a Phonetic PosteriorGram (PPG) speech feature system [47]. They compared their proposed model CNN-PPG to a CNN-MFCC model and an ASR-based system. The authors also employed data augmentation to obtain more training data that is needed for deep learning model training. The speech commands of the training set in the proposed system were converted to MFCC features and then to PPG features which were used to train the CNN model. The results indicated that the proposed Speech Command Recognition (SCR) system (CNN-PPG) provided better results than CNN-MFCC and ASR systems with a recognition accuracy of 93.49%. This shows that the PPG speech feature can achieve better recognition performance than MFCC. Also, the proposed system used a smaller model size with nearly half the number of parameters compared to the other models which can reduce the implementation cost for the users.

Green et al. investigated the performance of personalized automatic speech recognition systems for dysarthric speakers against the performance of Speaker Independent ASR models (SI) and human transcribers [48]. A group of 432 speakers with different speech impairment types, causes, and severity levels recorded their speech using a web-based application. The first independent ASR model (SI-1) that was used as a benchmark was Google's commercial ASR system accessed through speech-to-text API. The second one (SI-2) was an end-to-end ASR model based on RNN-T architecture. The researchers created a personalized ASR model based on (SI-2) for each participant using their own recordings. For the adaptation process, the authors worked on optimizing the fine-tuning procedure because the recorded data of each speaker was only between 15 minutes and 2 hours. They found that updating the first five encoder layers instead of the whole model prevented overfitting. Moreover, the authors utilized SpecAugment to increase the system's robustness and found that it worked best when they greatly increased the time masking and reduced the frequency masking settings. The results indicated that the personalized ASR models outperformed the speaker-independent models significantly and provided an accuracy that was similar or better than human listeners. The median

WER of the proposed personalized ASR models was 4.6%.

Shahamiri introduced Speech Vision, a dysarthric ASR system, that addressed some challenges usually faced by ASR systems [49]. These challenges are dysarthric speech data scarcity, phoneme labeling imprecision, and the alternation and inaccuracy of dysarthric phonemes. In Speech Vision, speech features are extracted visually to identify the shape of the words pronounced by dysarthric speakers which leads to the elimination of phoneme related challenges. As for the problem of data scarcity, Speech Vision employed three strategies: visual data augmentation, synthetic data generation, and transfer learning. The results showed that Speech Vision outperformed other dysarthric speech recognizers that use the same dysarthric data. Moreover, Speech Vision that used synthetic voicegrams delivered an average word recognition accuracy of 64.71%. Shahamiri further explored the generation of synthetic data that was employed in Speech Vision with Hu and Phadnis in [50]. They proposed the idea of utilizing and adapting speech generation systems that are designed to narrate normal speech in order to generate dysarthric speech. This synthesized dysarthric speech can be subsequently used in training an ASR system and thus increases the recognition accuracy. In addition, and to alleviate the limitation of dysarthric speech used to train the speech generation system, transfer learning was employed. The authors pre-trained the speech generation system on normal speech and then adapted this knowledge by training the model on dysarthric speech. After that, this generation system was used to synthesize dysarthric speech to be used in training their dysarthric ASR system (Speech Vision). The results showed that using synthetic dysarthric speech during training has improved the performance of ASR systems.

Liu et al. revealed the latest efforts of the Chinese University of Hong Kong (CUHK) to improve the performance of dysarthric ASR systems [51]. The authors experimented with different novel modeling techniques including spectra-temporal perturbation for data augmentation, Neural Architectural Search (NAS), and model-based speaker adaptation. In addition, they employed a cross-domain generation of visual features to be used within an Audio-Visual Speech Recognition (AVSR) that they developed. They found that the proposed speaker adaptation techniques were able to model the great variability among dysarthric speakers and allowed fast adaptation to each dysarthric speaker that can be performed using as little as 3.06 seconds of speech. The results showed that the combination of the proposed techniques yielded an average WER of 25.21%, which is the lowest WER on the UASpeech test set. Moreover, this combination was able to reduce the overall WER by an absolute 5.39% over the previously proposed CUHK system [37]. The proposed AVSR was able to achieve an average WER of 15.79% excluding very low intelligible speakers, which is close to normal speech recognition WERs. The authors obtained similar improvements when they utilized these techniques on a Chinese dysarthric ASR task using CUDYS dataset.

Deng et al. proposed a Bayesian parametric and neural architectural domain adaptation approach for dysarthric ASR instead of using the conventional adaptation approach that considers only parameter fine-tuning on limited data [39]. Both the standard parameters and hyper parameters of a lattice-free MMI factored TDNN system were trained on large quantities of normal speech obtained from two corpora: English LibriSpeech and Cantonese SpeechOcean. Then, the system was domain adapted to CUDYS dysarthric speech corpus. The results showed an absolute reduction in recognition error rate by 1.82% compared to the baseline systems that perform model parameter fine-tuning only. Also, continuous performance improvements were gained when the authors performed data augmentation and LHUC based speaker adaptation. The experimental results revealed that Bayesian adaptation was able to lessen the risk of overfitting that might occur when directly fine-tuning systems with large numbers of parameters. Moreover, they found that architectural adaptation was able to improve the generalization of systems with the use of parameter adaptation only.

A sequential contrastive learning framework was proposed by Wu et al. [7]. They explored several data augmentation methods to alleviate data scarcity as well as form two branches of the framework and support contrastive learning. These methods include time warping, frequency masking, and time masking. The authors utilized transfer learning as the model

was pre-trained on non-dysarthric speech from LibriSpeech and then it was fine-tuned on dysarthric speech from TORGO. The results demonstrated the effectiveness of the framework as it provided results better than or comparable to the supervised baseline. Moreover, combining data augmentation strategies provided better results than using a single one.

A recent study by Revathi et al. analyzed speech enhancement techniques and the use of multiple features in a cluster-based dysarthric ASR system [52]. The authors developed an isolated digit recognition system for dysarthric speech and presented a comparative analysis of dysarthric ASR using six features, seven enhancement techniques, and a Vector Quantization (VQ) based modeling technique. This was evaluated using test utterances of two female speakers with an intelligibility level of 6% and 95%. They performed two types of assessments: an experimental evaluation and a subjective assessment to test the utterances of dysarthric speakers. Regarding the dysarthric speaker with a 6% intelligibility level, the experimental evaluation using the automated system integrating all the features and speech enhancement techniques outperformed the subjective assessment with a 4% WER. As for the dysarthric speaker with a 95% intelligibility level, both experimental evaluation and manual recognition yielded the same results; 0% WER for the subjective assessment and experimental evaluation using a system that integrates GFE features and speech enhancement techniques.

Yue et al. investigated combining acoustic features with articulatory features to improve dysarthric ASR [53]. They proposed multi-stream architecture where the streams of acoustic and articulatory features are first pre-processed and then fused using different schemes to find the optimal fusion level and training dynamics. After fusion and before the output layer, fused streams are post-processed. Data augmentation was employed to augment the training data using speed perturbation. In addition, monophone regularization was used as an auxiliary task for optimization. The results showed that fusing articulatory and acoustic features using the optimal fusion scheme yielded a substantial reduction in absolute WER by up to 4.6% where the best improvement was for severe dysarthric speakers.

Mariya Celin et. al. incorporated transfer learning with data augmentation to support dysarthric ASR [54]. First, the authors trained a speaker-independent model on normal speech and performed transfer learning to three speaker-dependent ASR systems to compare various speech data augmentation techniques. The first was trained using speed and volume perturbed speech data. The second was trained using speech data augmented through Virtual Microphone array synthesis and Multi-Resolution Feature Extraction (VM-MRFE) which was previously used by the authors in [12]. The third was trained on speech data augmented using both techniques. All these systems were trained on UAspeech, TORGO, and SSN-Tamil dysarthric speech corpus developed by the authors. Moreover, they considered both continuous speech and isolated words for comparison purposes. The results showed that for isolated words, the combination of data augmentation techniques outperformed the stand-alone augmentation techniques with an average WER of 32.97% and 63.38% for low and very low intelligible speakers, respectively. This provided a reduction in WER in comparison to the latest results in the literature. As for continuous speech, the use of VM-MRFE augmentation technique provided a better reduction in WER compared to the use of speed and volume perturbation technique or the combination of both techniques. The average WER was 35.89%. The results also revealed that the appropriate augmentation technique depends on the nature of the utterance and whether it is isolated or continuous. Table V summarizes the reviewed papers in hybrid approach, the techniques used, and error rates.

TABLE V. SUMMARY OF HYBRID APPROACH PAPERS

Ref.	Year	Method	Techniques	Datasets	Error Rate
[43]	2020	Adaptation (TL) + Features (domain invariant features)	-CNN (baseline model) -DANN model - MTL model	UASpeech	Lowest avg. WER 25.09 % ** (DANN trained on labelled dysarthric & control speech)
[44]	2020	Data augmentation (speed perturbation & data combination	-E2E-ASR system (acoustic model based on Speech Transformer) - ASAT system	-LibriSpeech (normal speech)	Lowest PER 29.84% (with data augmentation - TS2-DA (TS2 + Teacher:S2))

		(normal & dysarthric)) + Adaptation (TL)		-TORGO (normal +dysarthric)	
[45]	2020	Data augmentation (speed perturbation) + Adaptation (TL)+ Features (MFCC, fMLLR)	-Light Gated Recurrent Units (LiGRU) acoustic model -Multi-task optimization techniques (Joint optimization + Monophone regularization) - AE-BN (feature extractor)	-LibriSpeech (normal) -TORGO (dysarthric)	Lowest avg. WER 52.37% (fMLLR+BN20 + mono)
[46]	2021	Data augmentation (speed perturbation) + Speaker adaptation (LHUC SAT) + Features	-Hybrid DNN acoustic model -VAEVE (to encode variability information) -LHUC (speaker adaptation technique)	UASpeech (training: normal + dysarthric, testing: dysarthric)	Lowest overall WER 25.7% (DNN + Data Aug. + LHUC SAT + VAEVE)
[47]	2021	Data augmentation (multi-condition training using corruption with noise data) + Features (MFCC, PPG)	CNN-PPG Speech Command Recognition (SCR) System	3 Mandarin speakers (19 speech commands)	Lowest avg. WER 6.51% (personalized SCR system - (CNN–PPG)) **
[48]	2021	Adaptation (TL) + Data augmentation (frequency masking & time masking)	-RNN-T architecture -SpecAugment (data augmentation technique)	Recordings of 432 speakers - English	Median WER 4.6% (personalized models)
[49]	2021	Data augmentation (Visual data augmentation & synthesis of dysarthric data) + Adaptation (TL)	-Spatial Convolutional Neural Network (S-CNN) - Deep convolutional text-to-speech (DC-TTS) generation system -Speech Vision ASR system	UASpeech	Absolute avg. WER 35.29% (Speech Vision: synthetic data included) **
[50]	2021	Data augmentation (synthesis of dysarthric speech) + Adaptation (TL)	-DC-TTS (generation system) -Speech Vision ASR system	-TORGO -UASpeech	Lowest avg. WER 35.69% (Speech Vision: synthetic data included) * **
[51]	2021	Data augmentation (spectra-temporal perturbation) + Adaptation (model-based speaker adaptation - auxiliary speaker embedding & model-based adaptation (LHUC, HUB & PAct) + Features (cross-domain generation of visual features)	-Manually designed DNN system (baseline system architecture) -Neural architecture search (NAS) auto-configured DNN system	-UASpeech (English dysarthric) -CUDYS dataset (Cantonese Dysarthric Speech)	Lowest avg. WER 25.21% for English (NAS DNN+ Data aug. +LHUC SAT+AV fusion) Lowest avg. CER 11.2% for Cantonese (system no. 6)
[39]	2021	Data augmentation (speed perturbation) + Adaptation (TL - neural domain adaptation - architectural and parametric adaptation of Bayesian TDNNs - LHUC speaker adaptation)	-LF-MMI TDNNs (baseline systems) -A Bayesian differentiable architectural search (DARTS) super-network -LHUC (speaker adaptation technique)	-LibriSpeech (normal English) -SpeechOcean (normal Cantonese) -DementiaBank (elderly speech) -CUDYS (Cantonese dysarthric)	English: Lowest WER 30.83% (Bayesian domain parametric & architectural adaptation+data augmentation+LHUC) Cantonese: Lowest CER 9.41% (Bayesian domain parametric & architectural adaptation+LHUC)
[7]	2021	Data augmentation (time warping, frequency masking & time masking) + Adaptation (TL)	CNN with pyramid CNN subsets	-LibriSpeech (normal for pre-training) -TORGO (dysarthric for fine-tuning)	Lowest avg. WER 15.88% (proposed model with fine-tuning)*
[52]	2022	Enhancement (mean squared error (MSE) log spectral amplitude extractor, MSE spectral amplitude extractor, wavelet denoising, probabilistic geometric approach, geometric approach, phase spectrum compensation) + Features (GFEERB, GFEMEL, GFEBARK, MGDFC, DCSTC & DOSTC)	Isolated digit recognition system (Modeling technique: Vector quantization (VQ) based clustering technique)	TORGO (number of isolated digits - 10)	6% intelligibility speaker: Lowest avg. WER 4% (integration of all features +enhancement techniques-isolated digit recognition) 95% intelligibility speaker: Lowest avg. WER 0% (integration of GFE features+ enhancement techniques-isolated digit recognition)

[53]	2022	Data augmentation (speed perturbation) + Features (acoustic (MFCC) & articulatory (lip))	-Acoustic model (CNN- multi-layer perceptrons (MLP) and LiGRU) -Monophone regularization (auxiliary task for optimization)	TORGO (dysarthric and typical speech)	Lowest avg. WER 43.2% (MFCC+lip with concat-2 fusion level)
[54]	2023	Data augmentation (VM-MRFE - speed & volume perturbation) + Adaptation (TL)	Transfer learning with DNN architecture (TDNN-F incorporating CNNs)	-TORGO -UASpeech -SSN-Tamil corpus by authors	Lowest avg. WER: UASpeech 35.86% * (VM-RFE +speed & volume perturbation - isolated words) TORGO 41.99% * (VM-MRFE - isolated words) SSN-Tamil 35.89% * (VM-MRFE - isolated words & sentences)

*Indicates that the average WER was not directly provided in the paper and therefore it was calculated using the formula: Avg = sum of error rates reported/number of error rates.

**indicates that the results were given in terms of recognition accuracy, and it was converted to error rate for comparison reans using the formula: Error rate = 100-accuracy.

III. DATASETS

One of the key requirements to develop and improve dysarthric speech recognition systems is the availability of dysarthric speech datasets. These datasets can be helpful in training the recognition system and in evaluating the system's performance. Table VI shows the available datasets and their languages.

TABLE VI. DYSARTHRIC DATASETS

Dataset / Ref.	Contents	No. of speakers	Dysarthria type/cause	Content type	Language
UASpeech [55]	765 isolated words per speaker (uncommon words, digits, computer commands, radio alphabet & common words)	15 dysarthric & 13 age-matched non-dysarthric	spastic dysarthria (Cerebral Palsy (CP))	Audiovisual recordings	American English
TORGO [56]	Single words or restricted sentences & unrestricted sentences (description of the content of some photos)	7 non-dysarthric & 8 dysarthric	spastic, athetoid, or ataxic (cerebral palsy) & amyotrophic lateral sclerosis (ALS)	Audiovisual recordings & electromagnetic articulography (aligned acoustic & articulatory recordings)	American English
Nemours [57]	814 short nonsense sentences & 74 sentences	11 dysarthric males	-	Audio recordings	American English
homeService [58]	command words	5 dysarthric speakers	severe dysarthric speakers	Audio recordings	British English
EasyCall [59]	dysarthric speech command dataset	24 non-dysarthric & 31 dysarthric	Parkinson's Disease (PD), ALS Huntington's Disease, peripheral neuropathy, myopathic, myasthenic lesions	Audio recordings	Italian
SSNCE [60]	365 utterances per speaker (single words & sentences including combination of common & uncommon phrases)	20 dysarthric & 10 non-dysarthric	Cerebral Palsy (CP)	Audio recordings	Tamil
PC-GITA [61]	21 isolated words per speaker	50 dysarthric & 50 non-dysarthric	PD with dysarthria	Audio recordings	Spanish
Dutch dysarthric speech database [62]	Isolated words & sentences	16 dysarthric speakers	PD, traumatic brain injuries (TBI) & cerebrovascular accident	Audio recordings	Dutch Netherlands
Korean dysarthric QoLT corpus [63]	isolated words & restricted sentences	10 non-dysarthric & 70 dysarthric	Cerebral Palsy (CP)	Audio recordings	Korean
IDEA [64]	211 isolated common words	45 dysarthric	ASL, Ataxia (ATX), Huntington's Chorea (HC), Multiple Sclerosis (MS),Myotonic Dystrophy,TBI, (MD),Neuropathy, PD, Stroke	Audio recordings	Italian

AllSpeak [65]	23 commands	8 non-dysarthric & 8 dysarthric	Amyotrophic Sclerosis Lateral (ALS)	Audio recordings	Italian
Whitaker [66]	19275 isolated-word utterances - 81 isolated words	6 dysarthric & 1 non-dysarthric	Cerebral Palsy (CP)	Audio recordings	American English
CCM [67]	words, sentences, & spontaneous speech	860 dysarthric & 80 non-dysarthric	PD, paralytic dysarthria, ALS, MS, ATX, Friedreich disease	Audio & some electro-glottographic recordings	French
The Aix Neurology-Hospital corpus (ANH)[a] [67]	vowels, sentences, & spontaneous speech	990 dysarthric & 160 non-dysarthric	PD & Parkinsonian syndromes	Audio & aerodynamic recordings	French
The TYPALOC Corpus [68]	Sentences & spontaneous speech (natural continuous speech)	28 dysarthric & 12 non-dysarthric	Extrapyramidal system with PD, Pyramidal system with ALS & Cerebellar system with Cerebellar ataxia (CA).	Audio recordings	French
The MSDM Database [69]	Syllables, characters, words, sentences, & spontaneous speech	25 dysarthric & 25 non-dysarthric	subacute stroke patients	Audio-visual recordings	Mandarin Chinese
CUDYS [70]	61 single words, 23 short sentences, passage, conversation & articulatory tasks	11 dysarthric & 5 non-dysarthric	cerebellar degeneration (spino-cerebellar ataxia (SCA))	Audio & video recordings	Cantonese
Copas [71]	utterances & words	182 dysarthric & 122 non-dysarthric	-	Audio recordings	Dutch Flemish
Italian dataset using CapisciAMe app [72]	isolated words (commands)	156 dysarthric speakers	neuromotor disabilities	Audio recordings	Italian

[a.] Also Known as AHN (Aix Hospital Neurology) corpus.

IV. Conclusion and Future Work

This survey discussed the latest efforts of dysarthric speech recognition and the different approaches and techniques that can be used to increase the recognition accuracy and support dysarthric speakers. It can be noticed that in most cases combining multiple approaches or systems yielded better results and lower error rates. Moreover, the availability of dysarthric speech datasets can boost the experiments carried out to support a certain language. As it can be observed, most of the research utilized English dysarthric datasets either for training and testing if the system is developed for English dysarthric speakers or to pre-train the model and then fine-tune it using the targeted language of ASR system. This is because of the availability of these datasets and the quantity of these recordings. Future work may consider developing datasets for low resource languages or languages that have none.

References

[1] F. Rudzicz, "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech," in *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, Tempe, Arizona, USA, 2007.

[2] M. J. Kim, J. Yoo and H. Kim, "Dysarthric Speech Recognition Using Dysarthria-Severity-Dependent and Speaker-Adaptive Models," in *Interspeech 2013*, Lyon, France, 2013.

[3] Z. Qian and K. Xiao, "A Survey of Automatic Speech Recognition for Dysarthric Speech," *Electronics*, vol. 12, no. 20, pp. 1-23, 2023.

[4] S. Yadav, D. M. Yadav and K. R. Desai, "A comprehensive survey of automatic dysarthric speech recognition," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 12, no. 3, pp. 242-250, 2023.

[5] 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," *Frontiers in Computer Science*, vol. 4, no. 770210, pp. 1-8, 2022.

[6] B. Vachhani, C. Bhat and S. K. Kopparapu, "Data Augmentation using Healthy Speech for Dysarthric Speech Recognition," in *Interspeech 2018*, Hyderabad, India, 2018.

[7] L. Wu, D. Zong, S. Sun and J. Zhao, "A SEQUENTIAL CONTRASTIVE LEARNING FRAMEWORK FOR ROBUST DYSARTHRIC SPEECH RECOGNITION," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, 2021.

[8] Y. Matsuzaka, R. Takashima, C. Sasaki and T. Takiguchi, "Data Augmentation for Dysarthric Speech Recognition Based on Text-to-Speech Synthesis," in *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*, Osaka, Japan, 2022.

[9] Y. Takashima, T. Takiguchi and Y. Ariki, "End-to-end dysarthric speech recognition using multiple databases," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[10] A. Misbullah, H.-H. Lin, C.-Y. Chang, H.-W. Yeh and K.-C. Weng, "Improving Acoustic Models for Dysarthric Speech Recognition using Time Delay Neural Networks," in *2020 International Conference on Electrical Engineering and Informatics (ICELTICs)*, Aceh, Indonesia, 2020.

[11] F. Xiong, J. Barker and H. Christensen, "PHONETIC ANALYSIS OF DYSARTHRIC SPEECH TEMPO AND APPLICATIONS TO ROBUST PERSONALISED DYSARTHRIC SPEECH RECOGNITION," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[12] T. A. M. Celin, T. Nagarajan and P. Vijayalakshmi, "Data Augmentation Using Virtual Microphone Array Synthesis and Multi-Resolution Feature Extraction for Isolated Word Dysarthric Speech Recognition," *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING*, vol. 14, no. 2, pp. 346-354, 2020.

[13] E. Hermann and M. M. Doss, "DYSARTHRIC SPEECH RECOGNITION WITH LATTICE-FREE MMI," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[14] E. Hermann and M. M. Doss, "Handling acoustic variation in dysarthric speech recognition systems through model combination," in *INTERSPEECH 2021*, Brno, Czechia, 2021.

[15] J. Harvill, D. Issa, M. Hasegawa-Johnson and C. Yoo, "SYNTHESIS OF NEW WORDS FOR IMPROVED DYSARTHRIC SPEECH RECOGNITION ON AN EXPANDED VOCABULARY," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, 2021.

[16] M. Soleymanpour, M. T. Johnson, R. Soleymanpour and J. Berry, "SYNTHESIZING DYSARTHRIC SPEECH USING MULTI-SPEAKER TTS FOR DYSARTHRIC SPEECH RECOGNITION," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022.

[17] Z. Yue, E. Loweimi and Z. Cvetkovic, "RAW SOURCE AND FILTER MODELLING FOR DYSARTHRIC SPEECH RECOGNITION," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022.

[18] Z. Yue, E. Loweimi, H. Christensen, J. Barker and Z. Cvetkovic, "Dysarthric Speech Recognition From Raw Waveform with Parametric CNNs," in *Interspeech 2022*, Incheon, Korea, 2022.

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

[20] C. Bhat, B. Das, B. Vachhani and S. Kumar Kopparapu, "Dysarthric Speech Recognition Using Time-delay Neural Network Based Denoising Autoencoder," in *Interspeech 2018*, Hyderabad, 2018.

[21] D. Wang, J. Yu, X. Wu, S. Liu, L. Sun, X. Liu and H. Meng, "End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[22] R. Rajeswari, T. Devi and S. Shalini, "Dysarthric Speech Recognition Using Variational Mode Decomposition and Convolutional Neural Networks," *Wireless Personal Communications*, vol. 122, no. 1, p. 293–307, 2022.

[23] C. Ding, S. Sun and J. Zhao, "Multi-Task Transformer with Input Feature Reconstruction for Dysarthric Speech Recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, 2021.

[24] L. Prananta, "Improving Automatic Speech Recognition For Dysarthric Speech," Delft University of Technology, Delft, Netherlands, 2021.

[25] L. Prananta, B. M. Halpern, S. Feng and O. Scharenborg, "The Effectiveness of Time Stretching for Enhancing Dysarthric Speech for Improved Dysarthric Speech Recognition," in *Interspeech 2022*, Incheon, Korea, 2022.

[26] B.-F. Zaidi, M. Boudraa, S.-A. Selouani, D. Addou and M. S. Yakoub, "Automatic Recognition System for Dysarthric Speech Based on MFCC's, PNCC's, JITTER and SHIMMER Coefficients," in *Advances in Computer Vision. CVC 2019. Advances in Intelligent Systems and Computing*, vol. 944, Cham, Switzerland, Springer, 2019, pp. 500-510.

[27] J. B. Mathew, J. Jacob, K. Sajeev, J. Joy and R. Rajan, "Significance of Feature Selection for Acoustic Modeling in Dysarthric Speech Recognition," in *2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, India, 2018.

[28] M. Kim, B. Cao, K. An and J. Wang, "Dysarthric Speech Recognition Using Convolutional LSTM Neural Network," in *Interspeech 2018*, Hyderabad, 2018.

[29] S. Hu, S. Liu, H. F. Chang, M. Geng, J. Chen, L. W. Chung, T. K. Hei, J. Yu, K. H. Wong, X. Liu and H. Meng, "The CUHK Dysarthric Speech Recognition Systems for English and Cantonese," in *INTERSPEECH 2019: Show & Tell Contribution*, Graz, Austria, 2019.

[30] B. F. Zaidi, S. A. Selouani, M. Boudraa and M. S. Yakoub, "Deep neural network architectures for dysarthric speech analysis and recognition," *Neural Computing and Applications*, vol. 33, no. 15, p. 9089–9108, 2021.

[31] S. Chandrakala, "Machine Learning Based Assistive Speech Technology for People with Neurological Disorders," in *Recent Advances in Intelligent Assistive Technologies: Paradigms and Applications*, vol. 170, H. Costin, B. Schuller and A. M. Florea, Eds., Cham, Switzerland, Springer Cham, 2020, pp. 143-163.

[32] A. Hernandez, P. A. Perez-Toro, E. Noth, J. R. Orozco-Arroyave, A. Maier and S. H. Yang, "Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition," in *Interspeech 2022*, Incheon, Korea, 2022.

[33] B. A. Al-Qatab, M. B. Mustafa, S. S. Salim and A. A. Sani, "Determining the adaptation data saturation of ASR systems for dysarthric speakers," *International Journal of Speech Technology*, vol. 24, no. 1, p. 183–192, 2021.

[34] Y. Sawa, R. Takashima and T. Takiguchi, "Adaptation of a Pronunciation Dictionary for Dysarthric Speech Recognition," in *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*, Osaka, Japan, 2022.

[35] F. Xiong, J. Barker, Z. Yue and H. Christensen, "SOURCE DOMAIN DATA SELECTION FOR IMPROVED TRANSFER LEARNING TARGETING DYSARTHRIC SPEECH RECOGNITION," in *2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[36] Y. Takashima, R. Takashima, T. Takiguchi and Y. Ariki, "Dysarthric Speech Recognition Based on Deep Metric Learning," in *INTERSPEECH 2020*, Shanghai, China, 2020.

[37] R. Takashima, T. Takiguchi and Y. Ariki, "TWO-STEP ACOUSTIC MODEL ADAPTATION FOR DYSARTHRIC SPEECH RECOGNITION," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[38] Y. Takashima, R. Takashima, T. Takiguchi and Y. Ariki, "Knowledge Transferability Between the Speech Data of Persons With Dysarthria Speaking Different Languages for Dysarthric Speech Recognition," *IEEE Access*, vol. 7, pp. 164320-164326, 2019.

[39] J. Deng, F. R. Gutierrez, S. Hu, M. Geng, X. Xie, Z. Ye, S. Liu, J. Yu, X. Liu and H. Meng, "Bayesian Parametric and Architectural Domain Adaptation of LF-MMI Trained TDNNs for Elderly and Dysarthric Speech Recognition," in *INTERSPEECH 2021*, Brno, Czechia, 2021.

[40] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Y. LAM, X. Wu, K. H. Wong, X. Liu and H. Meng, "Development of the CUHK Dysarthric Speech Recognition System for the UASpeech Corpus," in *Interspeech 2018*, Hyderabad, 2018.

[41] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, A. Hassidim and Y. Matias, "Personalizing ASR for dysarthric and accented speech with limited data," in *Interspeech 2019*, Graz, Austria, 2019.

[42] D. Wang, J. Yu, X. Wu, L. Sun, X. Liu and H. Meng, "Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Hong Kong, 2021.

[43] D. Woszczyk, S. Petridis and D. Millard, "Domain adversarial neural networks for dysarthric speech recognition," in *Interspeech 2020*, Shanghai, China, 2020.

[44] Y. Lin, L. Wang, S. Li, J. Dang and C. Ding, "Staged knowledge distillation for end-to-end dysarthric speech recognition and speech attribute transcription," in *Interspeech 2020*, Shanghai, China, 2020.

[45] Z. Yue, H. Christensen and J. Barker, "Autoencoder Bottleneck Features with Multi-Task Optimisation for Improved Continuous Dysarthric Speech Recognition," in *INTERSPEECH 2020*, Shanghai, China, 2020.

[46] X. Xie, R. Ruzi, X. Liu and L. Wang, "Variational Auto-Encoder Based Variability Encoding for Dysarthric Speech Recognition," in *INTERSPEECH 2021*, Brno, Czechia, 2021.

[47] Y.-Y. Lin, W.-Z. Zheng, W. C. Chu, J.-Y. Han, Y.-H. Hung, G.-M. Ho, C.-Y. Chang and Y.-H. Lai, "A Speech Command Control-Based Recognition System for Dysarthric Patients Based on Deep Learning Technology," *Applied Sciences* , vol. 11, no. 6, p. 2477, 2021.

[48] J. R. Green, R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson and K. Tomanek, "Automatic Speech Recognition of Disordered Speech: Personalized models outperforming human listeners on short phrases," in *INTERSPEECH 2021*, Brno, Czechia, 2021.

[49] 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.

[50] A. Hu, D. Phadnis and S. R. Shahamiri, "Generating synthetic dysarthric speech to overcome dysarthria acoustic data scarcity," *Journal of Ambient Intelligence and Humanized Computing*, 2021.

[51] S. Liu, M. Geng, S. Hu, X. Xie, M. Cui, J. Yu, X. Liu and H. Meng, "Recent Progress in the CUHK Dysarthric Speech Recognition System," *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 29, pp. 2267-2281, 2021.

[52] A. Revathi, R. Nagakrishnan and N. Sasikaladevi, "Comparative analysis of Dysarthric speech recognition: multiple features and robust templates," *Multimedia Tools and Applications*, vol. 81, no. 22, p. 31245–31259, 2022.

[53] Z. Yue, E. Loweimi, Z. Cvetkovic, H. Christensen and J. Barker, "Multi-Modal Acoustic-Articulatory Feature Fusion For Dysarthric Speech Recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022.

[54] T. A. Mariya Celin, P. Vijayalakshmi and T. Nagarajan, "Data Augmentation Techniques for Transfer Learning-Based Continuous Dysarthric Speech Recognition," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 601-622, 2023.

[55] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin and S. Frame, "Dysarthric speech database for universal access research," in *INTERSPEECH 2008*, Brisbane, Australia, 2008.

[56] F. Rudzicz, A. K. Namasivayam and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523-541, 2012.

[57] 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)*, Philadelphia, PA, USA, 1996.

[58] M. Nicolao, H. Christensen, S. Cunningham, P. Green and T. Hain, "A Framework for Collecting Realistic Recordings of Dysarthric Speech - the homeService Corpus," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 2016.

[59] R. Turrisi, A. Braccia, M. Emanuele, S. Giulietti, M. Pugliatti, M. Sensi, L. Fadiga and L. Badino, "EasyCall corpus: a dysarthric speech dataset," in *INTERSPEECH 2021*, Brno, Czechia, 2021.

[60] T. A. Mariya Celin, T. Nagarajan and P. Vijayalakshmi, "Dysarthric speech corpus in Tamil for rehabilitation research," in *2016 IEEE Region 10 Conference (TENCON)*, Singapore, 2016.

[61] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. González-Rátiva and E. Nöth, "New Spanish Speech Corpus Database for the Analysis of People Suffering from Parkinson's Disease," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014.

[62] E. Yılmaz, M. Ganzeboom, L. Beijer, C. Cucchiarini and H. Strik, "A Dutch Dysarthric Speech Database for Individualized Speech Therapy Research," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 2016.

[63] D.-L. Choi, B.-W. Kim, Y.-W. Kim, Y.-J. Lee, Y. Um and M. Chung, "Dysarthric Speech Database for Development of QoLT Software Technology," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012.

[64] M. Marini, M. Viganò, M. Corbo, M. Zettin, G. Simoncini, B. Fattori, C. D'Anna, M. Donati and L. Fanucci, "Idea: An Italian Dysarthric Speech Database," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, 2021.

[65] C. Di Nardi, R. Turrisi, A. Inuggi, N. Riva, I. Mauri and L. Badino, "An automatic speech recognition Android app for ALS patients," *Book series Studi AISV*, vol. 4, pp. 217--225, 2018.

[66] J. R. Deller Jr, M. S. Liu, L. J. Ferrier and P. Robichaud, "The Whitaker database of dysarthric (cerebral palsy) speech," *The Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3516-3518, 1993.

[67] C. Fougeron, L. Crevier-Buchman, C. Fredouille, A. Ghio, C. Meunier, C. Chevrie-Muller, N. Audibert, J.-F. Bonastre, A. Colazo Simon, C. Delooze, D. Duez, C. Gendrot, T. Legou, N. Levèque, C. Pillot-Loiseau, S. Pinto, G. Pouchoulin, D. Robert, J. Vaissiere, F. Viallet and C. Vincent, "The DesPho-APaDy Project: Developing an acoustic-phonetic characterization of dysarthric speech in French," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

[68] C. Meunier, C. Fougeron, C. Fredouille, B. Bigi, L. Crevier-Buchman, E. Delais-Roussarie, L. Georgeton, A. Ghio, I. Laaridh, T. Legou, C. Pillot-Loiseau and G. Pouchoulin, "The TYPALOC Corpus: A Collection of Various Dysarthric Speech Recordings in Read and Spontaneous Styles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 2016.

[69] J. Liu, X. Du, S. Lu, Y.-M. Zhang, H. An-ming, M. L. Ng, R. Su, L. Wang and N. Yan, "Audio-video database from subacute stroke patients for dysarthric speech intelligence assessment and preliminary analysis," *Biomedical Signal Processing and Control*, vol. 79, p. 104161, 2023.

[70] K. H. Wong, Y. T. Yeung, E. H. Y. Chan, P. C. M. Wong, G.-A. Levow and H. Meng, "Development of a Cantonese Dysarthric Speech Corpus," in *INTERSPEECH 2015*, Dresden, Germany, 2015.

[71] "Corpus Pathological and Normal Speech (COPAS)," University of Antwerp, University of Ghent, 2011. [Online]. Available: the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-n3.

[72] D. Mulfari, G. Campobello, G. Gugliandolo, A. Celesti, M. Villari and N. Donato, "Comparison of Noise Reduction Techniques for Dysarthric Speech Recognition," in *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Messina, Italy, 2022.

أحدث التطورات في أنظمة التعرف على كلام ذوي صعوبات النطق: الطرق ومجموعات البيانات

تهاني الراجحي[2,1]، مراد يخلف[1]، أحمد السند[1]

[1] قسم نظم المعلومات، جامعة الملك سعود، الرياض، المملكة العربية السعودية

[2] قسم علم المعلومات، جامعة الإمام عبدالرحمن بن فيصل، الدمام، المملكة العربية السعودية

taalrajhi@iau.edu.sa, ykhlef@ksu.edu.sa, aasanad@ksu.edu.sa

المستلخص: عسر النطق هو اضطراب لفظي عصبي حركي ينتج عن إعاقة جسدية ويحد من وضوح الكلام. يمكن لذوي صعوبات النطق الاستفادة من أنظمة التعرف على الكلام لمساعدتهم على التواصل مع الآخرين بشكل أفضل. تستعرض هذه الورقة أحدث الدراسات على أنظمة التعرف على كلام ذوي صعوبات النطق والتي أجريت خلال السنوات الخمس الماضية وتحديداً من عام 2018 وحتى 2023. صنفت هذه الأعمال بناءً على النهج المتبع لتحسين نظام التعرف على كلام ذوي صعوبات النطق. تتضمن هذه النُهُج زيادة البيانات وتحسين الكلام العسر والعمل على خصائص الكلام والصوت وتكييف النظم واستخدام هجين من عدة طرق.

الكلمات المفتاحيةـــ عسر النطق، التعرف التلقائي على الكلام، التعرف التلقائي على كلام ذوي صعوبات النطق، اعتلال الكلام