EURASIP Journal on Audio, Speech, and Music Processing

REVIEW

Open Access

A survey of technologies for automatic Dysarthric speech recognition

Zhaopeng Qian[1*], Kejing Xiao[2] and Chongchong Yu[1]

Abstract

Speakers with dysarthria often struggle to accurately pronounce words and effectively communicate with others. Automatic speech recognition (ASR) is a powerful tool for extracting the content from speakers with dysarthria. However, the narrow concept of ASR typically only covers technologies that process acoustic modality signals. In this paper, we broaden the scope of this concept that the generalized concept of ASR for dysarthric speech. Our survey discussed the systems encompassed acoustic modality processing, articulatory movements processing and audio-visual modality fusion processing in the application of recognizing dysarthric speech. Contrary to previous surveys on dysarthric speech recognition, we have conducted a systematic review of the advancements in this field. In particular, we introduced state-of-the-art technologies to supplement the survey of recent research during the era of multi-modality fusion in dysarthric speech recognition. Our survey found that audio-visual fusion technologies perform better than traditional ASR technologies in the task of dysarthric speech recognition. However, training audio-visual fusion models requires more computing resources, and the available data corpus for dysarthric speech is limited. Despite these challenges, state-of-the-art technologies show promising potential for further improving the accuracy of dysarthric speech recognition in the future.

Keywords Dysarthric speech recognition, Automatic speech recognition, Audio-visual speech recognition, Multi-modality fusion technology

1 Introduction

Speech, as a carrier of linguistic expression, is generated through the coordinated movements of articulatory organs, which are regulated by neural activities in speech functional areas of the brain [1]. Speech plays an important role in people's daily communication and is an essential medium for them to carry out social activities [2, 3]. Dysarthria as a speech disorder only refers to neuromuscular disturbances concerning strength, speed, tone, steadiness or accuracy of the movements responsible for speech production. Dysarthria is not dyslexia and dysarthric speakers have no difficulties in writing, speech comprehension or cognition of words and grammatical structures. Due to cortical lesions, dysarthric speakers show a series of neuropathological characteristics and the degree of dysarthria is affected by the position and severity of neuropathies [1]. Dysarthric speakers seldom pronounce correctly. Especially for the speakers with severe dysarthria, communication with others is extremely difficult, which not only brings great inconvenience to the patients, but also increases their psychological burden [4]. Therefore, it is critical to study the ways for dysarthric speakers to rehabilitate, better communicate with others and return to the society.

Automatic speech recognition (ASR) can be very helpful for speakers with dysarthria [5]. Dysarthric speakers are easily exhausted, less able to express emotions and are prone to drooling and dysphagia. As a result, collecting

*Correspondence:
Zhaopeng Qian
qianzhaopeng@btbu.edu.cn
[1] School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing, China
[2] School of Information Engineering, Beijing Institute of Graphical Communication, Beijing, China

dysarthric speech is extremely difficult [6]. This difficulty leads to the scarcity of dysarthric speech data, which adds to the difficulty of ASR for dysarthric speech. In addition, as their pathogenesis differs, dysarthric speakers vary a lot in their pronunciation, which also results in a larger and more complex variation in the acoustic space of dysarthric speech compared with normal speech [7, 8]. In fact, human speech perception is inherently a bimodal process that uses both acoustic and visual information. Previously, many researches have demonstrated that incorporating visual modality can enhance the performance of noisy speech recognition task [9–12]. This has prompted the application of audio-visual speech recognition (AVSR) technology in addressing disordered speech [13–15].

Presently, reviews of dysarthria mainly include causes of pathology [16–18], computer-aided diagnosis [19–21], treatment and its assessment [22–36]. To our best knowledge, the previous reviews on ASR for dysarthric speech primarily focused on the challenges faced when applying ASR to the elderly with dysarthria [37] and explored both general and specific factors that affect the accuracy of ASR for dysarthric speech [38]. Moreover, the technologies of ASR (especially, the generalized ASR) for dysarthric speech have a great development. Therefore, the primary objective of our survey is to discuss the trends of generalized ASR technologies for dysarthric speech including research on dysarthric speech databases, technologies of ASR for dysarthric speech, technologies of AVSR for dysarthric speech. Our survey provides a more comprehensive and systematic review of the development of the technologies of generalized ASR (ASR and AVSR) technologies for dysarthric speech, highlighting the latest advancements and future directions in this field. The main contributions of this paper include the following three aspects:

1) We present some commonly used databases of dysarthric speech that are employed for training Automatic Speech Recognition (ASR) or Audio-Visual Speech Recognition (AVSR) systems;
2) We provide a comprehensive summary of both traditional and state-of-the-art technologies utilized in ASR for dysarthric speech, along with an analysis of the distinct characteristics of each ASR technology;
3) We introduce the latest audio-visual fusion technologies applied in the tasks of dysarthric speech recognition.

The rest of this paper will be as follows: Section 2 introduces how we retrieve and select the papers for review; Section 3 introduces the databases of dysarthric speech used to train ASR or AVSR, trends of ASR technologies for dysarthric speech, newest technologies of AVSR for dysarthric speech, respectively; Section 4 discusses the challenges and the future prospects dysarthric speech recognition; Section 5 gives a conclusion.

2 Methodology

Regarding the main objective, this review follows, where possible, well-established practices for conducting and reporting scoping reviews as suggested by the PRISMA statement [39]. Of the 27 items on the PRISMA checklist, we are able to follow 13 in the paper's title, introduction, methods, results, discussion and funding. Instead of focusing on different papers' details and specificities, we aim to review the technologies of automated recognition of dysarthric speech.

2.1 Retrieval

In the course of the scoping paper retrieval, the following databases were searched: "Web of Science Core Collection", "IEEE Xplore" and "Engineering Village". Time limitation is set as from 1900 to 2023. The key words including "Automatic Dysarthric Speech Recognition", "ASR for Dysarthric Speech", "Audio-Visual Dysarthric Speech Recognition", "AVSR for Dysarthric Speech" are used as the retrieval condition.

2.2 Selection

All authors jointly decide on the following selection criteria to reduce possible deviations during selection. Firstly, we exclude the papers that are not cited by other researchers as their contribution may be insufficient. Secondly, we exclude less related papers. At this screening stage, the all authors jointly decide whether one paper is relevant to our research. We also exclude duplicated papers due to eligibility concerns. Finally, we select 82 representative papers fitting the theme of our research. The papers fit in three categories of "dysarthric speech databases", "technologies of ASR for dysarthric" and "technologies of AVSR for dysarthric speech". The whole selection process can be found in Fig. 1.

In Fig. 1, symbols "a", "b", "c" and "d" respectively represent the total number of papers searched with key words "Automatic Dysarthric Speech Recognition", "ASR for Dysarthric Speech", "Audio-Visual Dysarthric Speech Recognition", "AVSR for Dysarthric Speech". During the three stages of "Screening", "Eligibility" and "Inclusion", symbol n represents the total number of papers searched according to the screening rules at each stage.

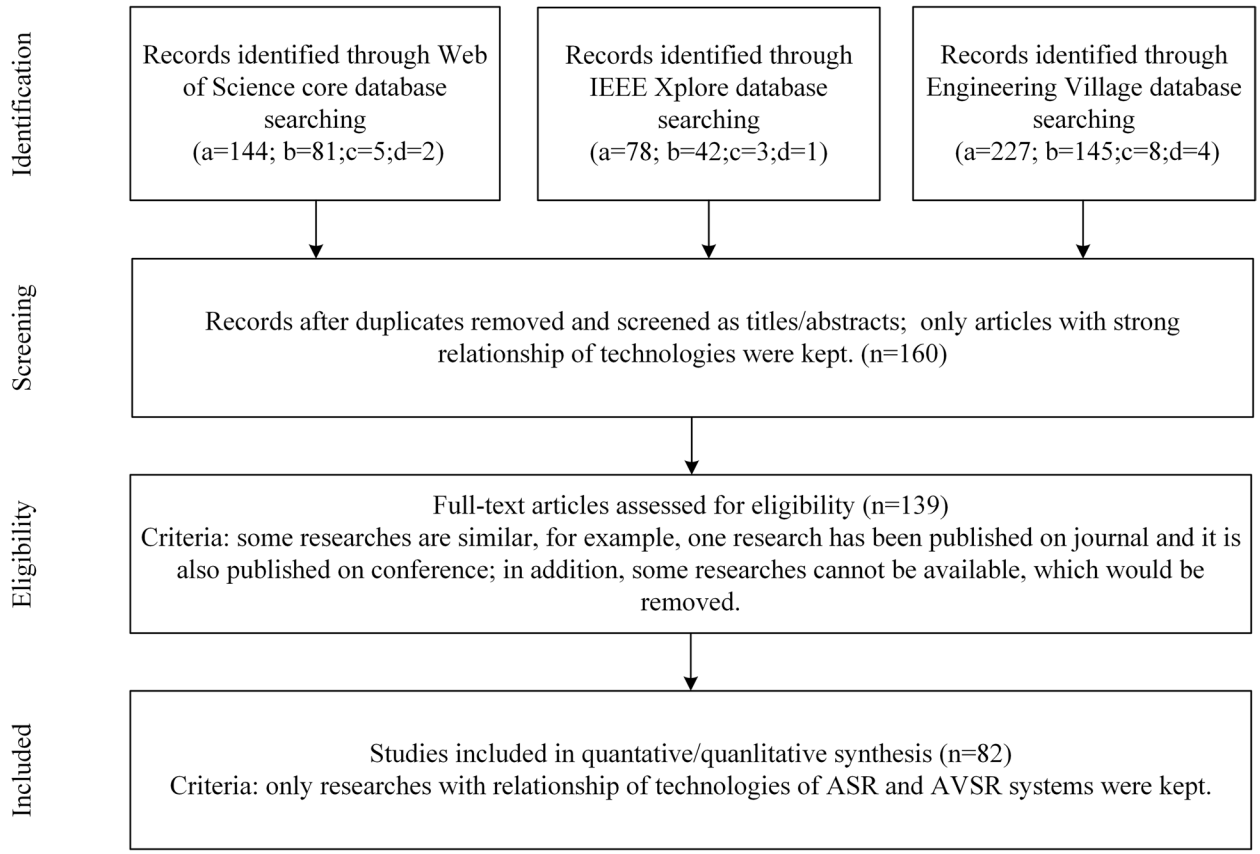

Fig. 1 PRISMA flow diagram of search methods

3 Results

From a technological perspective, factors that affect the performance of the ASR and AVSR system can be external and internal. External factors mainly concern about the characteristics of data, and internal factors mainly concern about the framework designed in ASR or AVSR system.

3.1 Commonly used Dysarthric speech databases

Databases of dysarthric speech are commonly used to train the models of ASR or AVSR. For example, the acoustic features extracted from audio files of dysarthric speech data corpus are used to train the acoustic model of ASR. The labelled text files are commonly used to train the language-lexical models of ASR or AVSR systems according to the phoneme dictionary. Additionally, the lip movements captured in video files are often fused with acoustic features to train the encoder model of an AVSR system. Table 1 lists some classic databases related to dysarthric speech that are discussed in this paper.

Whitaker is a database of dysarthric speech developed by Deller, et al. [40], which contains 19,275 isolated words spoken by 6 speakers with dysarthria resulting from cerebral palsy. Whitaker also contains the voices of healthy speakers as reference. The words in Whitaker database can be categorized into two groups: "TI-46" word list and "grandfather" word list. The "TI-46" word list contains 46 words, including 26 letters, 10 numbers and 10 control words of "start, stop, yes, no, go, help, erase, ruby, repeat, and enter". TI-46 is a standard vocabulary recommended by Texas Instruments Corporation [41] and has been widely used to test ASR algorithms. The "grandfather" word list contains 35 words, named after a paragraph that begins with "Let me tell you my grandfather..." and is generally used by the speech pathologists [42]. Each word in the TI-46 and grandfather word list was repeated for at least 30 times by 6 speakers with dysarthria. In most cases, 15 additional repetitions are also included, achieving a total of 45 repetitions. Normal speakers repeated 15 times for the database, serving as reference.

UASpeech database [43] was created by collecting dysarthric speech from 19 speakers with cerebral palsy. This database comprises 765 isolated words, including 455 unique words, of which 155 were repeated for three times. In addition, the corpus contains 300 rare words,

Table 1 Classical databases of dysarthric speech

Databases	Amount of Data	Number of Speakers	Isolated Word/Continuous Speech	Contain Video Files or Not
Whitaker	19,275(utterances)	6	Isolated Word	Not
UASpeech	765(words)	19	Isolated Word	Yes
TORGO	6177(utterances)	7	Mixed with Isolated Words and Continuous Speech	Not
Nemours	74(utterances)	11	Continuous Speech	Not
MOCHA-TIMIT	460(utterances)	2(+38 schedule recording)	Continuous Speech	Not
DEED	1680(utterances)	21(4 with dysarhtira and 17 healthy)	Continuous Speech	Yes

numbers, computer commands, radio alphabets, and common words to ensure maximum phoneme sequences diversity.

The TORGO database [44] was completed jointly by the University of Toronto Departments of Computer Science and Speech-Language Pathology and the Toronto Holland-Bloorview Kids Rehabilitation Hospital. Speech pathologists from the Bloorview Institute in Toronto recruited 7 dysarthric subjects aged between 16 and 50 years old, all of whom suffered from dysarthria caused by cerebral palsy (such as spasticity, athetosis and ataxia). An additional subject diagnosed with amyotrophic lateral sclerosis (ALS) was also included in the study. Speech language pathologists evaluated the speech motor function of each dysarthric speaker. The acoustic data in the database were collected directly using a headset microphone and a directional microphone, while the occlusal data were obtained through electromagnetic articulography (EMA), which measured tongue and other occlusal organs of the speakers during their pronunciations. The collected data underwent 3-dimensional (3D) reconstruction from binocular video sequences. The stimuli used in their study were sourced from various sources, including the TIMIT database, lists of identified telephone contacts, and assessments of speech intelligibility. For example, "non words" were used to control the baseline capabilities of dysarthric speakers, especially to gauge their artistic control in the presence of plosives and prosody. Speakers pronouncing /iy-p-ah/,/ah-p-iy/, and /p-ah-t-ah-k-ah/ respectively were asked to repeat for 5–10 times. These pronunciation sequences could help analyse the characteristics of pronunciation around blasting consonants [45]. The speakers were asked to keep pronouncing treble and bass vowels for over 5 seconds (i.e. pronouncing "e-e-e" for 5 seconds) This operation enabled researchers to explore how prosody could be applied to speech assistance technology, as many dysarthric speakers have limited control over their pitch [46]. "Short words" were critical for acoustic research [47] as voice activity detection was not necessary here. The stimuli included formant conversion between consonants and vowels, formant frequency of vowels, and sound energy of plosive consonants. The dysarthric speakers were asked to pronounce words like English numbers, yes/no, up/down/left/right/forward/back/select/menu, alphabet letters, the 50 words from the Frenchay dysarthria assessment [48], 360 words and 162 sentences from the "Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech" [49], and 10 most common words in the British National Corpus [50]. Restricted sentences were used to help the recording of complete and syntactically correct sentences for ASR. The content for recognition included the pre-selected sentences with rich phonemes, such as "The quick brown fox jumps over the lazy dog." and "She had your dark suit in great water all year.", the "grandfather" passage from the Nemours database [51], and the 460 TIMIT derived sentences used as prompts in the MOCHA database [52, 53]. Unrestricted sentences were used to supplement restricted sentences as they included sentences that are not fluent and have syntactic variations. All the participants spontaneously read sentences from the description of Webber Photo Cards: Story Starters [54].

The Nemours database [51] contains 814 short nonsense sentences out of which 74 were spoken by the 11 male speakers with varying severity of dysarthria. In addition, the database also contains two continuous paragraphs recorded by the 11 speakers. The recordings of all speakers were carried out in a special room with Sony PCM-2500 microphones. Speakers repeated the content following the instructor for an average time of 2.5 to 3 hours including breaks. The database was marked on word and phoneme levels. Words were labelled manually while phonemes were marked by Deep Neural Networks (DNN) and Hidden Markov Model (HMM) (DNN-HMM)-based ASR and manually corrected later. Due to the sparse phoneme distribution in the Nemours database, researchers find it challenging to use the database

to train ASR models and explore the potential impact of dysarthria [55, 56].

The dysarthric MOCHA-TIMIT database was created by Alan Wrench [52] in 1999. He selected the 460 short sentences from the TIMIT database [53] to include the major connected speech processes in English. The researcher used EMA (500 Hz sampling rate), laryngography (16 kHz sampling rate) and electropalatography (EPG, 200 Hz sampling rate) to collect the movement of the speakers' upper lip, lower lip, upper incisor, lower incisor, tongue tip, tongue blade (1 cm from tip of tongue) and tongue dorsum (1 cm from blade of tongue). The researcher planned to record the speech of 2 healthy speakers (one male and one female) and 38 speakers with dysarthria in May 2001. At present, detailed recording results are unclear and dysarthric data are unavailable online.

DEED is an audio-visual British English database that contains both dysarthric and normal speech. DEED has been ethically approved by the University of Sheffield, UK [57]. The whole name of DEED is the Dysarthric Expressed Emotional Database. In DEED data corpus, six basic emotions including "happiness", "sadness", "anger", "surprise", "fear" and "disgust" can support the researchers to explore the dysarthric emotion classification task. The DEED data corpus can be available online.[1] The dysarthric speech part is recorded by 4 speakers: one female speaker with dysarthria due to cerebral palsy and 3 speakers with dysarthria due to Parkinson's disease (2 female and 1 male). The text material of DEED is a subset of the material used in the SAVEE database [58]. Besides, the text material consists of 10 TIMIT sentences per emotion.

3.2 Trends of ASR Technologies for Dysarthria Speech

Our survey summarizes the technologies used to design ASR for dysarthric speech from 1900 to 2023. The development trend of technologies can be found in Fig. 2.

3.2.1 Early machine learning methods-based ASR for Dysarthric speech

To our best knowledge, the beginning research of ASR for dysarthric speech was initialized by Jayaram and Abdelhamied [59]. They delved into an artificial neural networks (ANN)-based acoustic model and successfully tested it on dysarthric speakers. The data used in their study [59] included 10 words recorded by a male dysarthric speaker, whose speech had only 20% intelligibility. The highest accuracy of ASR achieved was 78.25% [59]. Subsequently, Blaney and Wilson [60] tried to elucidate the cause of articulatory difference between dysarthric and healthy speakers, and explored the relationship between articulatory difference and the accuracy of ASR, for instance the relationship between the voice offset time (VOT) of voiced plosives, fricatives and vowels and ASR accuracy. Furthermore, speakers with moderate dysarthria exhibit greater variability in acoustic tests compared to those with mild dysarthria and healthy speakers. In an experiment for comparing different acoustic features in ASR based on the HMM, Polur and Miller [61] studied the fast Fourier transform, linear predictive, and Mel-Frequency Cepstrum Coefficients (MFCC) extracted from data provided training input to several whole-word hidden Markov model configurations. Their experimental results show that a 10-state ergodic model using 15 msec frames was better than other configurations and the MFCC performs better than the fast Fourier transform and linear predictive coding for training the ASR. According to the results of the above researches, the question arises: "how does the variation in acoustic features extracted from dysarthric speech affect the outcomes of acoustic models?" To answer this question, Fager [62] examined the duration variability of the isolated words and voice types of 10 speakers with dysarthria caused by traumatic brain injury (TBI) and healthy speakers. Their study explored the relationship between intelligibility and duration as well as between intelligibility and variability of dysarthric speech. The results revealed significant statistical differences between dysarthric and normal speech in terms of pronunciation duration and voice types. Consequently, some researchers have attempted to reduce the variability between dysarthric and normal speech to further enhance the accuracy of ASR for dysarthric speech. For example, Hawley, et al. [63] developed a limited vocabulary, speaker-dependent ASR that proved robust in dealing with speech variabilities. The accuracy for the training data set improved from 88.5% to 95.4% ($p < 0.001$). Even for speakers with severe dysarthria, the average word level recognition accuracy reached 86.9%.

In addition to the acoustic model, researchers have also made efforts to improve the accuracy of ASR for dysarthric speech. For instance, Greenm, et al. [64] developed an ASR based on context dependent HMM (CD-HMM) to recognize the audio commands from speakers with severe dysarthria. Their study found that these speakers with severe dysarthria had poor control over the target machine. Hain [65] proposed a method that gradually reduced the number of pronunciation variants for each word, similar to classification. This approach achieved good ASR performance using only a single pronunciation lexical model and was validated on both the Wall Street Journal and Switchboard datasets. Morales and

[1] https://sites.google.com/sheffield.ac.uk/deed

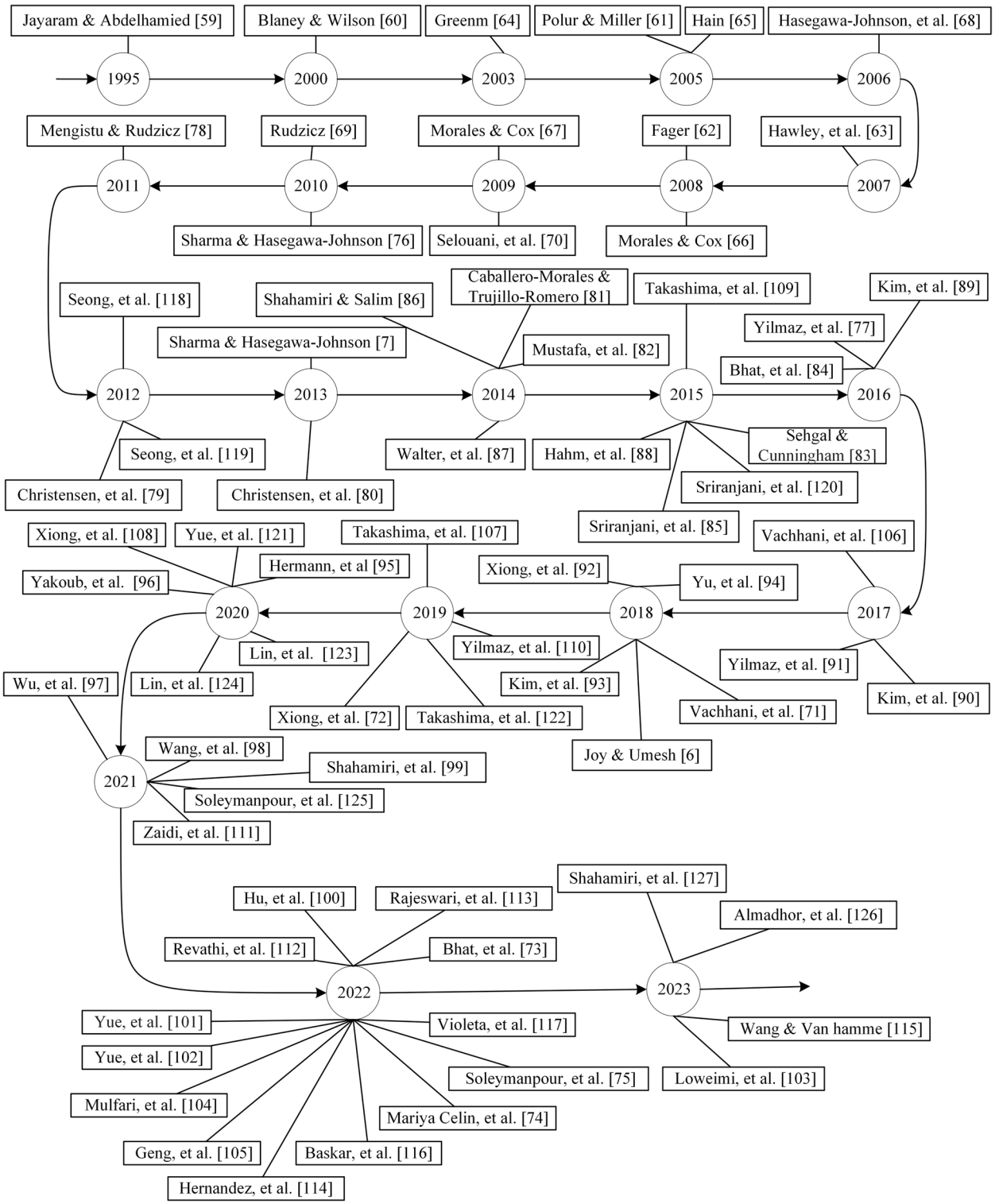

Fig. 2 The trends of ASR technologies for dysarthric speech

Cox [66, 67] improved the acoustic model using weighted finite-state transducers (WFST) at the levels of confusion matrix and word and language respectively. Their experimental results [66] showed that their approach outperformed maximum likelihood linear regression (MLLR) and Meta modelling.

Several researchers have compared different machine learning methods in designing ASR for dysarthric speech. For example, Hasegawa-Johnson, et al. [68] compared HMM-based and support vector machines (SVM)-based ASR to evaluate the performance of different approaches when dealing with dysarthric speech. This study collected data from three dysarthric speakers (two males and one female). Their experimental results showed that HMM-based ASR effectively recognized the speech of all dysarthric speakers, but had poor recognition accuracy for consonants that were damaged. SVM-based ASR had low accuracy for the stuttering speaker and high accuracy for the other two dysarthric speakers. Therefore, this research demonstrated that HMM-based ASR was robust for dysarthric speech with large variations in word length, while SVM-based ASR was better suited for processing dysarthric speech with missing consonants (with an average recognition accuracy of 69.1%). However, the intelligibility of the dysarthric speech used in their study was unknown. Additionally, Rudzicz [69] incorporated articulatory knowledge of the vocal tract to mark segmented and non-segmented sequences of non-typical speech. Their research [69] combined the above models with the discriminative learning models such as ANN, SVM and conditional random fields (CRF). Selouani, et al. [70] developed an auxiliary system that combines ASR and Text-to-Speech (TTS) to enhance the intelligibility of dysarthric speakers' speech. The re-synthesized speech demonstrated high intelligibility.

3.2.2 Technologies for Dealing with scarcity of Dysarthric speech data

The scarcity of dysarthric speech data has had a significant impact on the performance of ASR systems. To address this issue, researchers have explored various methods to improve the accuracy of ASR for dysarthric speech. For example, Vachhani, et al. [71] used temporal and speed modifications to healthy speech to simulate dysarthric speech, aiming to expand the training data corpus and improve the performance of ASR for dysarthric speech. Their experimental results showed a 4.24% and 2% absolute improvement using tempo-based and speed-based data augmentation, respectively, compared to the baseline using healthy speech alone for training. Xiong, et al. [72] proposed a nonlinear approach to modify speech rhythm, reducing the mismatch between typical and atypical speech by either modifying dysarthric speech into typical speech or modifying typical speech into dysarthric speech for data augmentation. The latter approach of their study was found to be more effective, improving absolute accuracy by nearly 7% when tested on the UASpeech database.

Bhat, et al. [73] used deep auto-encoder to modify and perturb healthy speech, thereby augmenting dysarthric speech data. They tested their data augmentation approach using an End-to-End ASR system and achieved an average of word error rate (WER) of 20.6% on the UASpeech, which represents an absolute improvement of 16% over a baseline without data augmentation. Mariya Celin, et al. [74] designed a speaker-dependent ASR system based on transfer learning, trained on UASpeech and SSN-Tamil databases. In their research, they employed virtual microphone array synthesis and multiresolution feature extraction (VM-MRFE) to augment the data. Their experimental results showed that their VM-MRFE-based data augmentation reduced WER for isolated word recognition by up to 29.98% and WER for continuous speech recognition by 24.95%, outperforming conventional speed and volume perturbation-based data augmentation methods.

Furthermore, employing TTS to generate simulated dysarthric speech is another effective data augmentation technique that can enhance the performance of ASR systems for dysarthric speech. Soleymanpour, et al. [75] developed a multi-speaker End-to-End TTS system to synthesize the dysarthric speech, incorporating dysarthria severity level and pause insertion mechanisms alongside other control parameters such as pitch, energy, and duration. Their experimental results demonstrated that their ASR system based on a DNN-HMM model trained on additional synthetic dysarthric speech achieved a WER improvement of 12.2% compared to a model not trained on synthetic data. Additionally, the inclusion of severity level and pause insertion controls resulted in a 6.5% reduction in WER for dysarthric speech recognition.

3.2.3 ASR models of speaker-dependent, speaker-adaptive and speaker-independent for Dysarthric speech

Researchers have also focused on speaker-dependent, speaker-adaptive, and speaker-independent problems in the development of ASR systems for dysarthric speech. Sharma and Hasegawa-Johnson [76] trained a speaker dependent ASR system for normal speakers using the TIMIT database and validated it with UASpeech data from seven dysarthric speakers. Their experimental results showed that the average accuracy of the speaker-adaptive ASR system was 36.8%, while the average accuracy of the speaker-dependent ASR system was 30.84%.

Yilmaz, et al. [77] trained a speaker-independent acoustic model based on DNN-HMM using speech from different Dutch languages and tested it with Flemish speech. Their findings demonstrated that training the ASR system with speech from various Dutch languages improved its performance for Flemish data.

3.2.4 Specific strategies of improving ASR for Dysarthric speech

Several strategies have proven useful in further improving the performance of ASR systems for dysarthric speech. For instance, Mengistu and Rudzicz [78] proposed using acoustic and lexical adaptation to improve the ASR for dysarthric speech. Their experimental results showed that acoustic adaptation reduced the world error rate by 36.99%, leading to a significant improvement in accuracy. Additionally, the pronunciation lexicon adaptation (PLA) model further decreased the WER by an average of 8.29% when six speakers with moderate to severe dysarthria were asked to pronounce a large vocabulary pool of over 1500 words. The speaker-dependent system with five-fold cross-validation demonstrated that PLA-based ASR also reduced the average WER by 7.11%. Christensen, et al. [79] proposed using MLLR and maximum a posteriori (MAP) adaptation strategies to improve ASR for dysarthric speech. Their approach improved recognition performance for all speakers with an average increase in accuracy of 3.2% and 3.5% respectively against the speaker dependent and independent baseline systems. Their approach improved recognition performance for all speakers, resulting in an average increase in accuracy of 3.2% and 3.5% against the speaker-dependent and independent baseline systems, respectively. They later proposed an alternative domain adaptive approach, combining in-domain and out-domain data to train a deep belief network (DBN) to enhance ASR for dysarthric speech [80]. The key point in works [80] was that during acoustic feature extraction, out-domain data training was used to generate a DBN. Augmented Multi-party Interaction (AMI) meeting corpus and TED talk corpus were then applied to optimize the previously trained model. Finally, the optimized model was verified on the UASpeech database, achieving an average recognition accuracy of 62.5%, which was 15% higher than the baseline method before optimization. Sharma & Hasegawa-Johnson [7] proposed an interpolation-based approach that obtained prior articulatory knowledge from healthy speakers and applied this knowledge to dysarthric speech through adaptation. Their study was validated on the UASpeech database. The experimental results showed that compared with the baseline approach of MAP adaptation, the interpolation-based approach achieved an absolute improvement of 8% and a relative improvement of 40% in recognition accuracy. Caballero-Morales and Trujillo-Romero [81] integrated multiple pronunciation patterns to improve the performance of ASR for dysarthric speech. This integration was achieved by weighing the response of the ASR system when different language model restrictions were set. The response weight parameters were estimated using a genetic algorithm, which also optimized the structure of the HMM-based implementation process (Meta-models). Their research was tested on the Nemours speech database and the experimental results showed that the integrated approach had higher accuracy than the standard Meta-model and speaker adaptation approach. Mustafa, et al. [82] used well-known adaptive technologies like MLLR and constrained-MLLR to improve ASR for dysarthric speech. The model trained using dysarthric speech and normal speech was applied as the source model. The experimental results showed that training normal and dysarthria speech together could effectively improve the accuracy of ASR systems for dysarthric speech. Constrained-MLLR had better performance than MLLR in dealing with mildly and moderately dysarthric speech. In addition, phoneme confusion was the main factor causing errors in the ASR of severely dysarthric speech. Sehgal and Cunningham [83] discussed the applicability of various speaker-independent systems, as well as the effectiveness of speaker adaptive training in implicitly eliminating the differences in pronunciation among dysarthric speakers. Their research relied on hybrid MLLR-MAP for both speaker-independent and speaker-adaptive training systems, which were tested on the UASpeech database. Their experimental results showed that compared with the baseline approach, the research achieved an increase of 11.05% in absolute accuracy and 20.42% in relative accuracy. Furthermore, the speaker adaptive training system was more suitable for dealing with severely dysarthric speech and had better performance than speaker-independent systems. Bhat, et al. [84] proposed combining multi-taper spectrum estimation and multiple acoustic features (such as jitter or shimmer) with MLLR (fMLLR) and speaker-adaptive methods to improve ASR for dysarthric speech. Sriranjani, et al. [85] proposed using fMLLR to process and combine pooled data and dysarthric speech to normalize the effect of inter-speaker variability. The results showed that combining features achieved a relative improvement of 18.09% and 50.00% over the baseline system for the Nemours database and UASpeech (digit set) database, respectively.

3.2.5 Deep learning technologies of ASR for Dysarthric speech

Deep learning methods have demonstrated superior performance in Automatic Speech Recognition (ASR)

applications compared to traditional machine learning techniques. In recent years, researchers have focused on how to use deep learning methods to further improve the accuracy of ASR systems for dysarthric speech. Shahamiri and Salim [86] proposed a dysarthric multi-networks speech recognizer (DM-NSR), which employs a multi-views, multi-learners strategy known as multi-nets ANN. The approach effectively accommodates the variability inherent in dysarthric speech. Experiment results on the UASpeech database revealed that DM-NSR achieved a 24.67% improvement in accuracy compared to single-network dysarthric speech recognizers. Walter, et al. [87] investigated unsupervised learning models for automatic dysarthric speech recognition. Their approach involved using vector quantization (VQ) to obtain Gaussian posterior-grams at the frame level, followed by training acoustic unit descriptors (AUD) and phone-like units' Hidden Markov Models (HMMs) in an unsupervised manner. Hahm, et al. [88] explored three across-speakers normalization methods in the acoustic and articulatory spaces of speakers with dysarthria: Procrustes Matching (a physiological method in the articulatory space), Vocal Tract Length Normalization (VTLN, a data-driven method in the acoustic space), and MLLR. These methods were employed to address the significant variation in phonation among individuals with dysarthria. Their study [88] was based on the ALS database and demonstrated that training the triple phoneme DNN-HMM (Triph-DNN-HMM) using acoustic and articulatory data and normalizing methods yielded the best performance. The phoneme error rate of this optimal combination was 30.7%, which is 15.3% lower than that of the baseline method, "triple phoneme Gaussian Mixed Model-Hidden Markov Model (Triph-GMM-HMM) trained using acoustic data."

Kim, et al. [89] developed an ASR system for dysarthric speech using Kullback-Leibler (KL)-HMM approach. In their research, the emission probability of each state was modelled based on the posterior probability distribution of phonemes estimated by DNN. Their approach was trained on a corpus recorded by 30 speakers (12 speakers with mild dysarthria, 8 speakers with moderate dysarthria and 10 healthy speakers) and the speakers were asked to pronounce several hundred words. Their experimental results showed that the DNN-HMM approach based on KL divergence outperformed traditional GMM-HMM and DNN approaches. Subsequently, the researchers [90] proposed using KL-HMM to capture the variations of dysarthric speech. In their framework [90], the state emission probability was predicted by the posterior probability value of phoneme. Additionally, the researchers introduced a speaker adaptation method based on "L2-norm" regularization (also known as ridge regression) to further reflect the specific speech patterns of individual speakers, thereby reducing confusion. Their approach improved the distinguishability of state classification distributions in KL-HMM while retaining the specific speaker information. Their research [90] was conducted on a self-made database comprising 12 speakers with mild dysarthria, 8 speakers with moderate dysarthria, and 10 normal individuals. Their experimental results showed that combining DNN with KL-HMM yielded better performance than traditional speaker-adaptive DNN-based approaches in dysarthric speech recognition tasks.

Yilmaz, et al. [91] proposed a multi-stage DNN training scheme, aiming to achieve high performance of ASR for dysarthric speech with a small amount of in-domain training data. Their experimental results demonstrated that this multi-stage DNN approach significantly outperformed a single-stage baseline system trained with a large amount of normal speech or a small amount of in-domain data, achieving higher accuracy in recognizing Dutch dysarthric speech.

Xiong, et al. [92] proposed a method that employs long short-term memory (LSTM) to simulate the inverse mapping from acoustic to articulatory space, aiming to enhance the accuracy of ASR for dysarthric speech. Their proposed approach supplemented information for DNN, taking advantage of acoustic and articulatory information. Kim, et al. [93] developed a convolutional LSTM (CLSTM)-Recurrent Neural Networks (RNN) (CLSTM-RNN)-based ASR for dysarthric speech, which was validated on a self-made database comprising 9 dysarthric speakers. Their experimental results showed that the CLSTM-RNN achieved better performance than convolutional neural networks (CNN) and LSTM-RNN. Joy and Umesh [6] explored a variety of methods to improve the performance of ASR for dysarthric speech. They adjusted the parameters of different acoustic models, used speaker-normalized Cepstrum features, trained a speaker-specific acoustic model using a complex DNN-HMM model with dropout and sequence-discrimination strategies, and incorporated specific information from dysarthric speech to enhance recognition accuracy for severely and severely-moderately dysarthric speakers. Their research was tested on the TORGO database and achieved ideal recognition accuracy. Yu, et al. [94] proposed a series of deep-neural-network-framework acoustic models based on time delayed neural networks (TDNN), LSTM-RNN, and their advanced variants to develop an ASR system for dysarthric speech. They also utilized learning hidden unit contribution (LHUC) to adapt to the acoustic variations of dysarthric speech and improved feature extraction by constructing

a semi-supervised complementary auto-encoder. Test results on the UASpeech dataset showed that this integrated approach achieved an overall word recognition accuracy of 69.4% on a test set containing 16 speakers.

Hermann and Doss [95] proposed to use the lattice-free maximum mutual information (LF-MMI) in the advanced sequence discriminative model. Their research [95] aimed to further improve the accuracy of ASR for dysarthric speech. Experimental results on the TORGO database showed that the performance of ASR for dysarthric speech using LF-MMI was effectively improved. Yakoub, et al. [96] proposed a deep learning architecture including CNN, empirical mode decomposition and Hurst (EMDH)-based mode selection to improve the accuracy of ASR for dysarthric speech. The k-fold cross-validation test conducted on the Nemours database showed that this architecture outperformed both the GMM-HMM and CNN baseline approaches without EMDH enhancement, achieving overall accuracy improvements of 20.72% and 9.95%, respectively.

Wu, et al. [97] proposed a contrastive learning framework to capture the acoustic variations of dysarthric speech, aiming to obtain robust recognition results of dysarthric speech. Their study also explored data augmentation strategies to alleviate the scarcity of speech data. Wang, et al. [98] proposed using meta-learning to re-initialize the basic model to tackle the mismatch between statistical distributions of normal and dysarthric speech. They extended model-agnostic meta learning (MAML) and Reptile algorithms to update the basic model, repeatedly simulating adaptation to different speakers with dysarthria. Experimental results on the UASpeech dataset showed that this meta-learning approach reduced the relative WER by 54.2% and 7.6% compared to a DNN-HMM-based ASR without fine-tuning and an ASR with fine-tuning, respectively. Shahamiri [99] developed a specific ASR system called Speech Vision, which learned to recognize the shape of words pronounced by dysarthric speakers. Their visual acoustic modeling approach helped eliminate phoneme-related challenges, and visual data augmentation was used to address data scarcity. Experimental results on the UASpeech database demonstrated a 67% improvement in accuracy.

Hu, et al. [100] applied neural architecture search (NAS) to automatically learn the two hyper-parameters of factored time delay neural networks (TDNN-Fs), namely the left and right splicing context offsets and the dimensionality of the bottleneck linear projection at each hidden layer. They utilized differentiable neural architecture search (DARTS) to integrate architecture learning with lattice-free maximum mutual information (LF-MMI) training, Gumbel-Softmax and Pipelined DARTS to reduce confusion over candidate architectures and improve generalization of architecture selection, and penalized DARTS to incorporate resource constraints to balance the trade-off between system performance and complexity. Their experimental results based on the UASpeech database demonstrated that the NAS approach for TDNN-Fs achieved significant improvement in ASR for dysarthric speech.

Yue et al. [101] proposed a multi-stream model as the acoustic model of ASR for dysarthric speech, which consists of convolutional, recurrent, and fully connected layers neural networks. This framework allows for pre-processing various information streams and fusing them at an optimal level of abstraction. Their experimental results based on the TORGO and UASpeech databases showed that the WERs can achieve 35.3% and 30.3%, respectively. In their study, they also compared the results of multi-stream model-based ASR trained by different acoustic features of dysarthric speech, such as MFCC, filter bank (FBank), raw waveform, and i-vector. Their research demonstrated that such a multi-stream processing leverages information encoded in the vocal tract and excitation components and leads to normalizing nuisance factors such as speaker attributes and speaking style. This operation can lead to better handling of dysarthric speech that exhibits large inter- and intra-speaker variabilities and results in a notable performance gain. Subsequently, researchers [102] studied how to effectively further improve the performance of data augmentation and multi-stream acoustic modelling through combining non-parametric and parametric CNNs fed by hand-crafted and raw waveform features. Their experimental results based on the TORGO database showed that parametric CNNs outperform non-parametric CNNs, with an average WER reaching up to 35.9% tested on dysarthric speech. Loweimi, et al. [103] used the raw real and imaginary parts of the Fourier transform of speech signals to investigate the multi-stream acoustic modelling approach. In their framework, the real and imaginary parts are treated as two streams of information, pre-processed via separate convolutional networks, and they combined at an optimal level of abstraction, followed by further post-processing via recurrent and fully connected layers of neural networks. Their experimental results based on TORGO show that the WER of dysarthric speech achieved to 31.7%. The multi-stream modelling approach provides a novel direction to improve the performance of ASR for dysarthric speech. This operation can reduce the loss of information caused by using single feature extracted by speech.

Mulfari, et al. [104] exploited a CNN architecture to predict the presence of a reduced number of speech commands within an atypical speech. In their

research, they focused on isolated word recognition. Their ASR model was trained on a 21 K speech data corpus. Geng, et al. [105] proposed speech spectrum decomposition based on singular value decomposition (SVD) to facilitate speaker adaptation of hybrid DNN/TDNN and end-to-end Conformer speech recognition systems based on auxiliary feature. Their experimental results based on UASpeech and TORGO showed that their proposed spectro-temporal deep feature adapted systems outperformed the i-vector and x-vector adaptation by up to 2.63% absolute reduction in WER. The best average of WER of their method can achieve 25.05% based on 16 dysarthric speakers from UASpeech.

3.2.6 Transfer learning technologies of ASR for Dysarthric speech

Knowledge transferred from the acoustic space of normal speech to the acoustic space of dysarthric speech can be very helpful to improve the performance of ASR for dysarthric speech by pre-training methodology. Vachhani, et al. [106] proposed to use a deep auto-encoder to enhance the MFCC representation performance of ASR for dysarthric speech. In the research [106], normal speech was used to train the auto-encoder, and then the trained auto-encoder was used in transfer learning to improve the representation of acoustic features. Test results on the UASpeech showed that the accuracy of ASR for dysarthric speech improved by 16%. Takashima, et al. [107] proposed to use transfer learning to obtain knowledge from normal and dysarthric speech, and then used the target dysarthric speech to fine-tune the pre-trained model. This approach was tested on Japanese dysarthric speech. Experimental results showed that this transfer learning approach significantly improved the performance of ASR for dysarthric speech. Xiong, et al. [108] proposed an improved transfer learning framework, which was suitable for increasing the robustness of the dysarthric speech recognition. The proposed approach was utterance-based and selected source domain data based on the entropy of posterior probability. The ensuing statistical analysis obeyed a Gaussian distribution. Compared with convolutional neural networks time delay deep neural networks (CNN-TDNN) trained by source domain data (as the transfer learning baseline), the proposed approach performed better on the UASpeech database. The proposed approach could accurately select potentially useful source domain data and improved absolute accuracy by nearly 20% against the transfer learning baseline for recognition of moderately and severely dysarthric speech.

3.2.7 Representation learning technologies of ASR for Dysarthric speech

Representation of acoustic features can significantly impact the performance of ASR for dysarthric speech. Takashima, et al. [109] used the pre-trained convolutional bottleneck network (CBN) to extract acoustic features from dysarthric speech and trained an ASR system using these features. Their study was based on patients with hand foot cerebral palsy, which is challenging due to their limited ability to pronounce words. To address overfitting issues caused by limited data, they incorporated convolution limited Boltzmann machines during pre-training. Word recognition experiments demonstrated that this approach outperformed networks without pre-training. Yilmaz, et al. [110] explored the use of gammatone features in ASR for dysarthric speech and compared them to traditional Mel-filters. Gammatone features were found to better capture resolution variation in the spectrum, making them more representative of vocal tract kinematics. Using gammatone features improved the robustness of ASR by explaining the variability observed in the acoustic space. Zaidi, et al. [111] explored how to combine DNN, CNN and LSTM to improve the accuracy of ASR for dysarthric speech. They compared MFCC, Mel-Frequency Spectrum Coefficient (MFSC), and Perceptual Linear Prediction (PLP) in their study. Results on the Nemours database showed that CNN-based ASR achieved an accuracy of 82% when PLP parameters were used, which was 11% and 32% higher than LSTM-based and GMM-HMM-based ASR, respectively. Revathi, et al. [112] proposed a combination of gammatone energy with filters calibrated in different non-linear frequency scales (GFE), stockwell features, modified group delay cepstrum (MGDFC), speech enhancement, and VQ-based classification. After fusing all acoustic feature parameters at the decision level of speech enhancement, the WER of ASR for dysarthric speech (with an intelligibility of 6%) was reduced to 4%. Additionally, the WER of ASR for dysarthric speech (with an intelligibility of 95%) was reduced to 0%. However, their research was based on a corpus where only digital pronunciation was available, making it difficult to evaluate its applicability. Rajeswari, et al. [113] treated dysarthric speech as distorted or noisy voice and enhanced it using variational mode decomposition (VMD) and wavelet thresholding before recognizing it using CNN as characters. Experimental results on the UASpeech database showed that the average accuracy of ASR for dysarthric speech was 91.8% without enhancement and improved to 95.95% with VMD enhancement. However, their results lack statistical significance due to the absence of standard deviations or confidence intervals provided in their study.

Hernandez, et al. [114] pre-trained an acoustic model with features extracted from Wav2Vec, Hubert, and the cross-lingual XLSR (a cross-lingual data corpus). Their findings suggest that speech representations pre-trained on large unlabelled data can enhance ASR performance for dysarthric speech. Wang & Van hamme [115] compared various mono- or cross-lingual pre-training methodologies and quantitatively examined the benefits of pre-training for Dutch dysarthric speech recognition. Baskar, et al. [116] explored integrating wav2vec with fMLLR features or x-vectors during fine-tuning. They proposed an adaptation network for fine-tuning wav-2vec using these features, achieving a 57.72% WER for high severity in UASpeech. Violeta, et al. [117] investigated the self-supervised learning frameworks (wav2vec 2.0 and WavLM, a large scale self-supervised pre-trained model proposed by Azure Speech Group of Microsoft) using different setups and compared the performances of ASR systems for pathological speech (including electro-laryngeal speech and dysarthric speech) with different supervised pre-training setups. Their experimental results based on UASpeech show that the best WER of extremely severe dysarthric speech can achieve 51.8%, however, the result is not better than the only-used acoustic feature Mel-scale FBank. This is because the discrepancy between the normal speech and dysarthric speech is too large. To further improve the performance of ASR using un-supervised, self-supervised or semi-supervised strategy, we should find an effective way to minify this discrepancy between the normal speech and dysarthric speech. In fact, a large amount of speech data corpus used to pre-train the representation model by un-supervised learning method can effectively make the neural networks learn the prior knowledge from the training data. This operation combining transfer-learning methods would make the framework perform better in ASR for dysarthric speech than the model without pre-trained by speech data corpus.

3.2.8 Language and lexical model of ASR for Dysarthric speech

Language-lexical models are crucial components of ASR, and their improvement can further enhance the accuracy of ASR for dysarthric speech. Seong, et al. [118] proposed a multiple pronunciation lexical modelling based on phoneme confusion matrix to improve the performance of ASR for dysarthric speech. The system first created a confusion matrix based on phoneme recognition results, then extracted pronunciation variation rules from the analysis of this matrix. These rules were applied to develop a speaker-dependent multiple pronunciation lexicon, which reduced relative WER by 5.06% compared to a group-dependent multiple pronunciation lexicon. Subsequently, the researchers [119] also used interpolation to integrate the lexicon and WFST of context dependent confusion matrix, aiming at correcting the wrongly recognized phonemes. Their approach reduced world error rate by 5.93% and 13.68% compared to baseline and error correction approaches with context-independent confusion matrices, respectively. Sriranjani, et al. [120] proposed to use the state specific vector (SSV) of the acoustic model trained by phoneme cluster adaptation (Phone-CAT) to identify the pronunciation errors of each speaker with dysarthria. SSV is a low-dimensional vector estimated for each binding state, with each element representing the weight of a specific mono-phoneme. Their method improved the relative accuracy of all speakers by 9% on the Nemours database compared to a standard lexical model-based ASR system.

Yue, et al. [121] used the classic TORGO database to investigate the impact of language model (LM) on ASR systems. By training the LM with different vocabularies, they analysed the confusion results of speakers with varying degrees of dysarthria. Their findings revealed that the optimal complexity of the LM is highly dependent on the speaker.

3.2.9 End-to-end ASR for Dysarthric speech

End-to-End ASR is a potent system for processing speech, but it requires a substantial amount of data for training. Unfortunately, obtaining dysarthric speech is challenging due to the speaker's inability to pronounce fluently. Researchers are exploring ways to enhance the accuracy of dysarthric speech recognition using End-to-End ASR with limited speech data. For example, Takashima et al., [122] proposed an End-to-End ASR framework that integrates an acoustic and language model. The acoustic model component of their framework is shared among speakers with dysarthria, while the language model portion is assigned to each language regardless of dysarthria. Their experimental results from a self-created Japanese data corpus demonstrated that End-to-End ASR can effectively recognize dysarthric speech even with minimal speech data.

Lin, et al., [123] suggested restructuring the acoustic model parameters into two layers, with only one layer being retrained. This approach aims to effectively utilize limited data for training End-to-End ASR systems for dysarthric speech. Additionally, they [124] proposed a staged knowledge distillation method to design an End-to-End ASR system and an automatic speech attribute transcription system for speakers with dysarthria resulting from either cerebral palsy or amyotrophic lateral sclerosis. Their experimental results on the TORGO database demonstrated that their proposed method achieved a

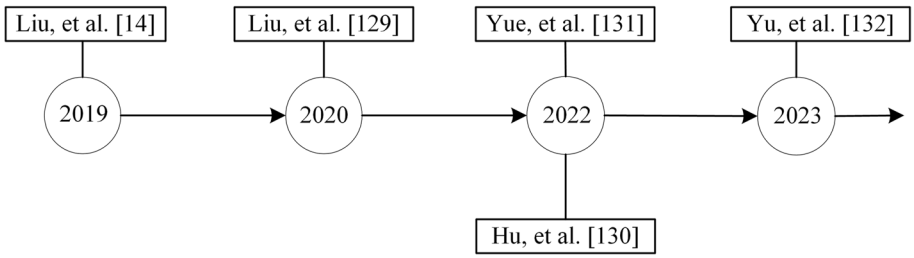

Fig. 3 The trends of AVSR technologies for dysarthric speech

38.28% relative phone error rate compared to the baseline method.

Different from the above ways in dealing with dysarthric speech, Soleymanpour, et al., [125] proposed a specialized data augmentation approach to enhance the performance of End-to-End ASR based on sub-word models. Their proposed methods contain two parts: "prosodic transformation" and "time-feature masking". Their experimental results of TORGO database showed that their approach reduced the character error rate (CER) by 11.3% and WER by 11.4%.

Almadhor, et al. [126] proposed a spatio-temporal ASR system based on Spatial CNN and multi-head attention Transformer to visually extract the acoustic features from dysarthric speech. Their experimental results on the UASpeech database showed that the best word recognition accuracy (WRA, they used WRA to evaluate their proposed method) achieved to 90.75%, 61.52%, 69.98% and 36.91% of low level dysarthric speech, mild level dysarthric speech, high level dysarthric speech and very high level dysarthric speech, respectively. However, overfitting problem influences the generalization of the system. Shahamiri, et al. [127] proposed to use Transformer framework-based End-to-End ASR for dysarthric speech. In their research, they designed a two-phase transfer-learning pipeline to leverage healthy speech. They investigated neural freezing configurations and used data augmentation for audio samples. After training 45 speaker-adaptive dysarthric ASR, their experimental results based on UASpeech showed that the WRAs of their best approach surpassed those of the benchmarks Fig. 3.

3.3 Trends of AVSR Technologies for Dysarthric Speech

The speech perception of humans exhibits a bi-modal processing characteristic [128]. Visual information is not affected by acoustic signal damage, providing compensatory information for ASR systems. Researchers have explored utilizing visual information to enhance the performance of ASR for dysarthric speech. Liu, et al. [14] proposed to use the Bayesian gated neural network (BGNN) to design the AVSR for dysarthric speech. In their research, the Bayesian gated control of contributions from visual features allows a robust fusion of audio and video modality. Their experimental results based on UASpeech showed that the WER of BGNN-based AVSR outperformed the DNN-based ASR by 4.5% and AVSR by 4.7%, respectively. However, severe voice quality degradation and large mismatch against normal speech affect AVSR's performance for dysarthric speech. Subsequently, the researchers [129] proposed a cross-domain visual feature generation approach to address the above problems. Their experimental results on the UASpeech corpus demonstrated that the AVSR based on cross-domain visual feature generation outperformed baseline ASR and AVSR without this approach. Hu, et al. [130] proposed a cross-domain acoustic-to-articulator inversion approach. Their model was pre-trained using parallel acoustic-articulatory data from the 15-hour TORGO corpus. After pre-training, the model was adapted to the 102.7-hour UASpeech corpus to generate articulatory features. The cross-domain acoustic-to-articulator inversion approach was designed using mixture density networks, with a cross-domain feature adaptation network reducing the mismatch between TORGO and UASpeech data. Their experiments showed that their best performing system incorporating video modality, cross-domain articulatory features, data augmentation, and learning hidden unit contributions speaker adaptation achieved an average WER of 24.82% on the 16 dysarthric speakers from the UASpeech corpus.

Yue, et al. [131] proposed a multi-stream framework that combines convolutional, recurrent, and fully connected layers to fuse articulatory and acoustic features extracted from dysarthric speech. While the fusion of articulatory and acoustic features does not constitute a true AVSR for dysarthric speech, the articulatory information collected through EMA can serve as visual movement information, providing compensatory information for the acoustic aspects of dysarthric speech. Therefore, we also include this research in our discussion.

Relying solely on lip movement as visual information to fuse acoustic features cannot fully cover the movements of articulatory. However, EMA has some drawbacks such as high cost, difficult acquisition, etc. To address the

Table 2 Comparison of ASR and AVSR technologies for dysarthric speech

Type of Model	ASR or AVSR	Training Time	Required Video?	Parameter Size of Model	Required GPU?
Machine Learning	ASR	Hours Level	No	≤Kilobyte Level	No Need
Deep Learning	ASR	Days Level	No	>Mbyte Level	Yes
Deep Learning	AVSR	Days (or even Weeks Level)	Yes	> > Mbyte Level (Gigabyte Level)	Yes

above problems, Yu, et al. [132] proposed a multi-stage fusion framework to further improve the performance of AVSR for dysarthric speech. In their research, their framework includes two-stage fusion operation. The first stage is the visual fusion. During this stage, Yu, et al. [132] obtained the facial speech functional area from the speakers frame by frame and fused these areas into the visual code. During the second stage, they fused the visual code and acoustic features using Hubert framework. After pre-training the AVSR model by LRS2 data corpus mixed with UASpeech, their fine-tuned AVSR can perform excellent. Their experimental results based on UASpeech show that the best WER was reduced by 13.5% on moderate dysarthric speech. In addition, for the mild dysarthric speech, the best result that the WER arrived at 6.05%. Even for the extremely severe dysarthric speech, the WER achieved at 63.98%, which reduced by 2.72% and 4.02% compared with the WERs of wav2vec and HuBERT, respectively.

3.4 Summary of ASR and AVSR Technologies for Dysarthric Speech

According to the above analysis of ASR and AVSR technologies for dysarthric speech, we can find that technologies before deep learning and after deep learning have a large difference. Table 2 provides the comparison of computing complexity, training time cost, amount of training data and type of computing resources, etc.

As shown in Table 2, the computing complexity of AVSR is significantly higher than that of ASR. Furthermore, deep learning-based ASR has a higher computing complexity compared to machine learning-based ASR. Training AVSR requires more audio and video data, which can be challenging due to the difficulty in collecting such data from speakers with dysarthria. The scarcity of training data is one of the main challenges in dealing with dysarthric speech. Additionally, both trained models of ASR and AVSR are prone to overfitting.

4 Discussion

The scoping review aims to provide a comprehensive overview of the development of ASR and AVSR technologies for dysarthric speech. Unlike previous surveys, our systematic review examines the trends in ASR and AVSR technologies specifically for dysarthric speech. To the best of our knowledge, this is the first survey that focuses on AVSR for dysarthric speech.

Over the past few decades, ASR technologies for dysarthric speech have evolved into various subdomains, such as "machine learning-based ASR", "technologies for data augmentation", "technologies for dealing with the speaker-adaptation of ASR", "specific strategies of improving ASR", "deep learning-based ASR", "transfer-learning-based ASR", "representation-learning-based ASR", "language-lexical model of ASR" and "end-to-end ASR". In fact, the edges of these above subdomains are blurred. Dividing them into the above subdomains is purely to facilitate our discussion. One major challenge we face is how to further improve the accuracy of ASR with limited resources for dysarthric speech, as collecting data from speakers with dysarthria can be extremely difficult.

The emergence of AVSR for dysarthric speech presents a promising new direction. Multi-modality fusion speech recognition allows us to leverage information from articulatory movements to compensate for the loss of acoustic information in dysarthric speech. However, there are still some obstacles to overcome in this novel field. For instance, EMA devices are costly, and using them for data collection may disturb the pronunciation process of dysarthric speakers. Additionally, obtaining synchronous audio and video signals from dysarthric speakers can be challenging. In the future, improving the performance of fusion frameworks and expanding the audio-visual data available for dysarthric speech will be crucial in advancing this field.

5 Conclusion

This scoping survey analysed 82 papers selected from 160 papers in the field of ASR and AVSR for dysarthric speech. The large variations among dysarthric speakers make it challenging for ASR to reduce speaker dependence due to poor generalization applicability of the acoustic model. This issue poses a significant challenge for the commercialization and popularization of ASR systems. Furthermore, dysarthric speakers exhibit substantial differences in their pronunciation, and available speech data is scarce. Data scarcity makes it difficult to

meet the required amount needed to train models using big data. Therefore, data scarcity is also an obstacle that hinders further improvement of models' performance in ASR for dysarthric speech. Despite the limited research on AVSR for dysarthric speech, this technology still holds promise. The scoping survey of ASR and AVSR for dysarthric speech can serve as a valuable technical reference for researchers in this field.

Acknowledgements
Not applicable.

Authors' contributions
Conceptualization, Zhaopeng Qian and Chongchong Yu.; Methodology, Zhaopeng Qian and Kejing Xiao.; formal analysis, Zhaopeng Qian and Chongchong Yu; writing—original draft preparation, Zhaopeng Qian.; writing—review and editing, Kejing Xiao.; funding acquisition, Zhaopeng Qian and Kejing Xiao. All authors have read and agreed to the published version of the manuscript.

Funding
This research was supported by the Humanity and Social Science Youth Foundation of Ministry of Education of China (No. 21YJCZH117), Research Foundation for Youth Scholars of Beijing Institute of Graphic Communication-Key Technologies for False News Detection Based on Deep Learning Research (No. 27170123034) and Humanities and Social Sciences Research Planning Fund of the Ministry of Education of China (No. 21A10011003).

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Not applicable.

Received: 29 July 2023 Accepted: 8 November 2023
Published online: 11 November 2023

References

1. L. Rampello, L. Rampello, F. Patti, M. Zappia, When the word doesn't come out: A synthetic overview of dysarthria. J. Neurol. Sci. **369**, 354–360 (2016). https://doi.org/10.1016/j.jns.2016.08.048
2. J.P. Rauschecker, S.K. Scott, Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. Nat. Neurosci. **12**(6), 718–724 (2009). https://doi.org/10.1038/nn.2331
3. M.D. Hauser, N. Chomsky, W.T. Fitch, The faculty of language: What is it, who has it, and how did it evolve? Science. **298**(5598), 1569–1579 (2002). https://doi.org/10.1126/science.298.5598.1569
4. S. Sapir, A.E. Aronson, The relationship between psychopathology and speech and language disorders in neurologic patients. J. Speech Hear. Disord. **55**(3), 503–509 (1990). https://doi.org/10.1044/jshd.5503.503
5. E. Sanders, M.B. Ruiter, L. Beijer, H. Strik, in *7th International Conference on Spoken Language Processing, ICSLP2002 – INTERSPEECH*. Automatic recognition of Dutch Dysarthric speech: A pilot study (Denver, Colorado, USA, 2002), pp. 661–664. https://doi.org/10.21437/ICSLP.2002-217
6. N.M. Joy, S. Umesh, Improving acoustic models in TORGO Dysarthric speech database. IEEE Trans. Neural Syst. Rehabilitation. Eng. **26**(99), 637–645 (2018). https://doi.org/10.1109/TNSRE.2018.2802914
7. H.V. Sharma, M. Hasegawa-Johnson, Acoustic model adaptation using in-domain background models for dysarthric speech recognition. Comput. Speech Lang. **27**(6), 1147–1162 (2013). https://doi.org/10.1016/j.csl.2012.10.002
8. M. Tu, A. Wisler, V. Berisha, J.M. Liss, The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance. J. Acoust. Soc. Am. **140**(5), EL416-EL422 (2016). https://doi.org/10.1121/1.4967208
9. J. Huang, B. Kingsbury, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Audio-visual deep learning for noise robust speech recognition (IEEE, Vancouver, BC, Canada, 2013), pp. 7596–7599. https://doi.org/10.1109/ICASSP.2013.6639140
10. Y. Mroueh, E. Marcheret, V. Goel, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep multimodal learning for audio-visual speech recognition (IEEE, South Brisbane, QLD, Australia, 2015), pp. 2130–2134. https://doi.org/10.1109/ICASSP.2015.7178347
11. S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, M. Pantic, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. End-to-end audiovisual speech recognition (IEEE, Calgary, AB, Canada, 2018), pp. 6548–6552. https://doi.org/10.1109/ICASSP.2018.8461326
12. S. Zhang, M. Lei, B. Ma, L. Xie, in *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Robust audio-visual speech recognition using bimodal DFSMN with multi-condition training and dropout regularization (IEEE, Brighton, UK, 2019), pp. 6570–6574. https://doi.org/10.1109/ICASSP.2019.8682566
13. C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, I. Li, in *2010 IEEE International Workshop on Multimedia Signal Processing*. Multimodal speech recognition of a person with articulation disorders using AAM and MAF (IEEE, Saint-Malo, France, 2010), pp. 517–520. https://doi.org/10.1109/MMSP.2010.5662075
14. S. Liu, S. Hu, Y. Wang, J. Yu, R. Su, X. Liu, H. Meng, in *Interspeech 2019*. Exploiting visual features using Bayesian gated neural networks for disordered speech recognition (Graz, Austria, 2019), pp. 4120–4124. https://doi.org/10.21437/Interspeech.2019-1536
15. S. Hu, S. Liu, H.F. Chang, M. Geng, J. Chen, L.W. Chung, T.K. Hei, J. Yu, K.H. Wong, X. Liu, H. Meng, *The CUHK Dysarthric speech recognition systems for English and Cantonese*, vol 15–19 (Interspeech, Graz, Austria, 2019), pp. 3669–3670
16. V. Di Stefano, M.V. De Angelis, C. Montemitro, M. Russo, C. Carrarini, M. di Giannantonio, F. Brighina, M. Onofrj, D.J. Werring, R. Simister, Clinical presentation of strokes confined to the insula: A systematic review of literature. Neurol. Sci. **42**, 1697–1704 (2021). https://doi.org/10.1007/s10072-021-05109-1
17. G. Noffs, T. Perera, S.C. Kolbe, C.J. Shanahan, F.M.C. Boonstra, A. Evans, H. Butzkueven, A. van der Walt, A.P. Vogel, What speech can tell us: A systematic review of dysarthria characteristics in multiple sclerosis. Autoimmun. Rev. **17**(12), 1202–1209 (2018). https://doi.org/10.1016/j.autrev.2018.06.010
18. S. Sapir, Multiple factors are involved in the dysarthria associated with Parkinson's disease: A review with implications for clinical practice and research. J. Speech. Lang. Hear. Res. **57**(4), 1330–1343 (2014). https://doi.org/10.1044/2014_JSLHR-S-13-0039
19. J. Rusz, T. Tykalova, L.O. Ramig, E. Tripoliti, Guidelines for speech recording and acoustic analyses in Dysarthrias of movement disorders. Mov. Disord. **35**(4), 803–814 (2020). https://doi.org/10.1002/mds.28465
20. L.K. Butler, S. Kiran, H. Tager-Flusberg, Functional near-infrared spectroscopy in the study of speech and language impairment across the life span: A systematic review. Am. J. Speech. Lang. Pathol. **29**(3), 1674–1701 (2020). https://doi.org/10.1044/2020_AJSLP-19-00050
21. B.E. Murdoch, Physiological investigation of dysarthria: Recent advances. Int. J. Speech. Lang. Pathol. **13**(1), 28–35 (2011). https://doi.org/10.3109/17549507.2010.487919
22. F. Yuan, X. Guo, X. Wei, F. Xie, J. Zheng, Y. Huang, Z. Huang, Z. Chang, H. Li, Y. Guo, J. Chen, J. Guo, B. Tang, B. Deng, Q. Wang, Lee Silverman voice treatment for dysarthria in patients with Parkinson's disease: A systematic review and meta-analysis. Eur. J. Neurol. **27**(10), 1957–1970 (2020). https://doi.org/10.1111/ene.14399
23. R. Chiaramonte, M. Vecchio, Dysarthria and stroke. The effectiveness of speech rehabilitation. A systematic review and meta-analysis of the

studies. Eur. J. Phys. Rehabil. Med. **57**(1), 24–43 (2020). https://doi.org/10.23736/s1973-9087.20.06242-5
24. C. Whillans, M. Lawrie, E.A. Cardell, C. Kelly, R. Wenke, A systematic review of group intervention for acquired dysarthria in adults. Disabil. Rehabil. **44**(13), 3002–3018 (2020). https://doi.org/10.1080/09638288.2020.1859629
25. Z. Wu, K. Hu, Y. Guo, Y. Tu, H. Zhang, Y. Wang, Acupuncture combined with speech rehabilitation training for post-stroke spasmodic dysphonia: A multicenter randomized controlled trial. World J. Acupuncture-Moxibustion. **24**(4), 12–16 (2014). https://doi.org/10.1016/S1003-5257(15)60021-6
26. N. Munoz-Vigueras, E. Prados-Roman, M.C. Valenza, M. Granados-Santiago, I. Cabrera-Martos, J. Rodriguez-Torres, I. Torres-Sanchez, Speech and language therapy treatment on hypokinetic dysarthria in Parkinson disease: Systematic review and meta-analysis. Clin. Rehabil. **35**(5), 639–655 (2020). https://doi.org/10.1177/0269215520976267
27. R. Chiaramonte, P. Pavone, M. Vecchio, Speech rehabilitation in dysarthria after stroke: A systematic review of the studies. Eur. J. Phys. Rehabil. Med. **56**(5), 547–562 (2020). https://doi.org/10.23736/s1973-9087.20.06185-7
28. Y.J. Park, J.M. Lee, Effect of acupuncture intervention and manipulation types on Poststroke dysarthria: A systematic review and Meta-analysis. Evid. Based Complement. Alternat. Med. **2020**, 4981945 (2020). https://doi.org/10.1155/2020/4981945
29. A Fletcher, M McAuliffe, Examining variation in treatment outcomes among speakers with dysarthria. Seminars in speech and language. Thieme Medical Publishers, 38(3), 191–199 (2017). https://doi.org/10.1055/s-0037-1602838
30. L. Pennington, N.K. Parker, H. Kelly, N. Miller, Speech therapy for children with dysarthria acquired before three years of age. Cochrane Database Syst. Rev. **7**, CD006937 (2016). https://doi.org/10.1002/14651858.CD006937.pub3
31. R. Kaipa, A.M. Peterson, A systematic review of treatment intensity in speech disorders. Int. J. Speech. Lang. Pathol. **18**(6), 507–520 (2016). https://doi.org/10.3109/17549507.2015.1126640
32. S.A. Borrie, M.J. McAuliffe, J.M. Lissb, Perceptual learning of Dysarthric speech: A review of experimental studies. J. Speech. Lang. Hear. Res. **55**(1), 290–305 (2012). https://doi.org/10.1044/1092-4388(2011/10-0349)
33. C. Mitchell, A. Bowen, S. Tyson, Z. Butterfint, P. Conroy, Interventions for dysarthria due to stroke and other adult-acquired, non-progressive brain injury. Cochrane Database Syst. Rev. **1**, CD002088 (2017). https://doi.org/10.1002/14651858.CD002088.pub3
34. M. Trail, C. Fox, L.O. Ramig, S. Sapir, J. Howard, E.C. Lai, Speech treatment for Parkinson's disease. NeuroRehabilitation. **20**(3), 205–221 (2005). https://doi.org/10.3233/NRE-2005-20307
35. S. Pinto, C. Ozsancak, E. Tripoliti, S. Thobois, P. Limousin-Dowsey, P. Auzou, Treatments for dysarthria in Parkinson's disease. Lancet Neurol. **3**(9), 547–556 (2004). https://doi.org/10.1016/S1474-4422(04)00854-3
36. K.M. Yorkston, K.A. Spencer, J.R. Duffy, Behavioral management of respiratory/phonatory dysfunction from dysarthria: A systematic review of the evidence. J. Med. Speech-Lang. Pathol. **11**(2), xiii-xxxviii (2003)
37. V. Young, A. Mihailidis, Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. Assist. Technol. **22**(2), 99–112 (2010). https://doi.org/10.1080/10400435.2010.483646
38. M.B. Mustafa, F. Rosdi, S.S. Salim, M.U. Mughal, Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. Expert Syst. Appl. **42**(8), 3924–3932 (2015). https://doi.org/10.1016/j.eswa.2015.01.033
39. D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, The PRISMA Group, Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. **6**(7), e1000097 (2009)
40. D. JR Jr., M.S. Liu, L.J. Ferrier, P. Robichaud, The Whitaker database of dysarthric (cerebral palsy) speech. J. Acoustical Soc. Am. **93**(6), 3516–3518 (1993). https://doi.org/10.1121/1.405684
41. G.R. Dodding, T.B. Schalk, Speech recognition: Turning theory to practice. IEEE Spectr. **18**(9), 26–32 (1981). https://doi.org/10.1109/MSPEC.1981.6369809
42. W. Johnson, F. Darley, D. Spriestersbach, *Diagnostic Methods in Speech Pathology* (Harper & Row, New York, 1963)
43. H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, S. Frame, in *Ninth Annual Conference of the International Speech Communication Association (Interspeech 2008)*. Dysarthric speech database for universal access research (Brisbane, Australia, 2008), pp. 1741–1744
44. F. Rudzicz, A.K. Namasivayam, T. Wolff, The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang. Resour. Eval. **46**(4), 523–541 (2012). https://doi.org/10.1007/s10579-011-9145-0
45. J.W. Bennett, P.H.H.M. van Lieshout, C.M. Steele, Tongue control for speech and swallowing in healthy younger and older subjects. Int. J. Orofacial Myology. **33**, 5–18 (2007)
46. R. Patel, Prosodic control in severe dysarthria: Preserved ability to mark the question-statement contrast. J. Speech. Language. Hear. Res. **45**(5), 858–870 (2002). https://doi.org/10.1044/1092-4388(2002/069)
47. N. Roy, H.A. Leeper, M Blomgren RM Cameron, a description of phonetic, acoustic, and physiological changes associated with improved intelligibility in a speaker with spastic dysarthria. Am. J. Speech. Lang. Pathol. **10**(3), 274–290 (2001). https://doi.org/10.1044/1058-0360(2001/025)
48. P. Enderby, Frenchay dysarthria assessment. Br. J. Disord. Commun. **15**(3), 165–173 (1980). https://doi.org/10.3109/13682828009112541
49. K.M. Yorkston, D.R. Beukelman, C. Traynor, *Assessment of Intelligibility of Dysarthric Speech* (Pro-ed, Austin, TX, 1984)
50. J.H. Clear, in *In: The digital word: Text-based computing in the humanities*. The British national corpus (MIT Press, Cambridge, MA, 1993), pp. 163–187
51. X. Menendez-Pidal, J.B. Polikoff, S.M. Peters, J.E. Leonzio, H.T. Bunnell, in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96.* The Nemours database of dysarthric speech (IEEE, Philadelphia, PA, USA, 1996), pp. 1962–1965. https://doi.org/10.1109/ICSLP.1996.608020
52. A Wrench, The MOCHA-TIMIT articulatory database. 1999. url:https://data.cstr.ed.ac.uk/mocha/
53. V. Zue, S. Seneff, J. Glass, Speech database development at MIT: TIMIT and beyond. Speech. Comm. **9**(4), 351–356 (1990). https://doi.org/10.1016/0167-6393(90)90010-7
54. S.G. Webber, *Webber Photo Cards: Story Starters* (2005)
55. F. Rudzicz, in *Assets 07: 9th international ACM SIGACCESS conference on Computers and Accessibility*. Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech (New York, NY, United States, 2007), pp. 255–256. https://doi.org/10.1145/1296843.1296899
56. F. Rudzicz, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Applying discretized articulatory knowledge to dysarthric speech (IEEE, Taipei, Taiwan, China, 2009), pp. 4501–4504. https://doi.org/10.1109/ICASSP.2009.4960630
57. L. Alhinti, S. Cunningham, H. Christensen, The Dysarthric expressed emotional database (DEED): An audio-visual database in British English. PLoS One. **18**(8), e0287971 (2023). https://doi.org/10.1371/journal.pone.0287971
58. P. Jackson, S. Haq, *Surrey Audio-Visual Expressed Emotion (Savee) Database* (University of Surrey, Guildford, UK, 2014)
59. G. Jayaram, K. Abdelhamied, Experiments in dysarthric speech recognition using artificial neural networks. J. Rehabil. Res. Dev. **32**, 162–162 (1995)
60. B. Blaney, J. Wilson, Acoustic variability in dysarthria and computer speech recognition. Clin. Linguist. Phon. **14**(4), 307–327 (2000). https://doi.org/10.1080/02699200050024001
61. P.D. Polur, G.E. Miller, Experiments with fast Fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model. IEEE Trans. Neural Syst. Rehabilitation Eng. **13**(4), 558–561 (2005). https://doi.org/10.1109/TNSRE.2005.856074
62. S.K. Fager, *Duration and Variability in Dysarthric Speakers with Traumatic Brain Injury (Dissertation)* (The University of Nebraska-Lincoln, 2008)
63. M.S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, O. Peter, R. Palmer, A speech-controlled environmental control system for people with severe dysarthria. Med. Eng. Phys. **29**(5), 586–593 (2007). https://doi.org/10.1016/j.medengphy.2006.06.009

64. P.D. Greenm, J. Carmichael, A. Hatzis, P. Enderby, M.S. Hawley, M. Parker, in *8th European Conference on Speech Communication and Technology, (EUROSPEECH 2003 - INTERSPEECH 2003), ISCA 2003*. Automatic speech recognition with sparse training data for dysarthric speakers (Geneva, Switzerland, 2003), pp. 1189–1192. https://doi.org/10.21437/Eurospeech.2003-384
65. T. Hain, Implicit modelling of pronunciation variation in automatic speech recognition. Speech. Comm. **46**(2), 171–188 (2005). https://doi.org/10.1016/j.specom.2005.03.008
66. S.O.C. Morales, S.J. Cox, in *9TH Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*. Application of weighted finite-state transducers to improve recognition accuracy for dysarthric speech (Brisbane, Australia, 2008), pp. 1761–1764. https://doi.org/10.21437/Interspeech.2008-485
67. S.O.C. Morales, S.J. Cox, Modelling errors in automatic speech recognition for Dysarthric speakers. Eurasip J. Adv. Signal Process. **1**, 1–14 (2009). https://doi.org/10.1155/2009/308340
68. M. Hasegawa-Johnson, J. Gunderson, A. Penman, T. Huang, in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*. Hmm-based and Svm-based recognition of the speech of talkers with spastic dysarthria (Toulouse, France, 2006), p. III.1060-III.1063. https://doi.org/10.1109/ICASSP.2006.1660840
69. F Rudzicz, Articulatory knowledge in the recognition of dysarthric speech. IEEE Trans. Audio Speech Lang. Process., 19(4), 947–960 (2010). https://doi.org/https://doi.org/10.1109/TASL.2010.2072499
70. S.A. Selouani, M.S. Yakoub, D. O'Shaughnessy, Alternative speech communication system for persons with severe speech disorders. Eurasip J. Adv. Signal Process. **1-12** (2009). https://doi.org/10.1155/2009/540409
71. B. Vachhani, C. Bhat, S.K. Kopparapu, in *Interspeech 2018*. Data augmentation using healthy speech for Dysarthric speech recognition (Hyderabad, 2018), pp. 471–475. https://doi.org/10.21437/Interspeech.2018-1751
72. F. Xiong, J. Barker, H. Christensen, in *44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition (IEEE, Brighton, England, 2019), pp. 5836–5840. https://doi.org/10.1109/ICASSP.2019.8683091
73. C. Bhat, A. Panda, H. Strik, *Improved ASR Performance for Dysarthric Speech Using Two-stage DataAugmentation* (Interspeech 2022, Incheon, Korea, 2022), pp. 46–50. https://doi.org/10.21437/Interspeech.2022-10335
74. T.A.M. Celin, P. Vijayalakshmi, T. Nagarajan, Data augmentation techniques for transfer learning-based continuous Dysarthric speech recognition. Circuits, Syst. Signal Process. **42**, 601–622 (2022). https://doi.org/10.1007/s00034-022-02156-7
75. M. Soleymanpour, M.T. Johnson, R. Soleymanpour, J. Berry, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Synthesizing Dysarthric speech using multi-speaker Tts for Dysarthric speech recognition (IEEE, Singapore, Singapore, 2022), pp. 7382–7386. https://doi.org/10.1109/ICASSP43922.2022.9746585
76. H.V. Sharma, M. Hasegawa-Johnson, in *Proceedings of the NAACL HLT 2010 workshop on speech and language processing for assistive technologies*. State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition (Los Angeles, CA, USA, 2010), pp. 72–79
77. E. Yilmaz, M.S. Ganzeboom, C. Cucchiarini, H. Strik, in *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*. Combining non-pathological data of different language varieties to improve DNN-HMM performance on pathological speech (San Francisco, USA, 2016), pp. 218–222. https://doi.org/10.21437/Interspeech.2016-109
78. K. Mengistu, F. Rudzicz, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Adapting acoustic and lexical models to dysarthric speech (Prague, Czech Republic, IEEE, 2011), pp. 4924–4927. https://doi.org/10.1109/ICASSP.2011.5947460
79. H. Christensen, S. Cunningham, C. Fox, P. Green, T. Hain, in *Interspeech'12: 13th Annual Conference of the International Speech Communication Association*. A comparative study of adaptive, automatic recognition of disordered speech (Portland, OR, USA, 2012), pp. 1776–1779. https://doi.org/10.21437/Interspeech.2012-484
80. H. Christensen, M.B. Aniol, P. Bell, P. Green, T. Hain, S. King, P. Swietojanski, in *14TH Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*. Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech (Lyon, France, 2013), pp. 3642–3645. https://doi.org/10.21437/Interspeech.2013-324
81. S.O. Caballero-Morales, F. Trujillo-Romero, Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition. Expert Syst. Appl. **41**(3), 841–852 (2014). https://doi.org/10.1016/j.eswa.2013.08.014
82. M.B. Mustafa, S.S. Salim, N. Mohamed, B. Al-Qatab, C.E. Siong, Severity-based adaptation with limited data for ASR to aid dysarthric speakers. PLoS One. **9**(1), e86285 (2014). https://doi.org/10.1371/journal.pone.0086285
83. S. Sehgal, S. Cunningham, in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2015)*. Model adaptation and adaptive training for the recognition of dysarthric speech (Dresden, Germany, 2015), pp. 65–71. https://doi.org/10.18653/v1/W15-5112
84. C. Bhat, B. Vachhani, S. Kopparapu, in *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*. Recognition of Dysarthric speech using voice parameters for speaker adaptation and multi-taper spectral estimation (San Francisco, USA, 2016), pp. 228–232. https://doi.org/10.21437/Interspeech.2016-1085
85. R. Sriranjani, M.R. Reddy, S. Umesh, in *2015 Twenty First National Conference on Communications (NCC)*. Improved acoustic modeling for automatic dysarthric speech recognition (IEEE, Mumbai, India, 2015), pp. 1–6. https://doi.org/10.1109/NCC.2015.7084856
86. S.R. Shahamiri, S.S.B. Salim, Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. Adv. Eng. Inform. **28**(1), 102–110 (2014). https://doi.org/10.1016/j.aei.2014.01.001
87. O. Walter, V. Despotovic, R. Haeb-Umbach, J.F. Gemnzeke, B. Ons, H. Van Hamme, in *15TH Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*. An evaluation of unsupervised acoustic model training for a dysarthric speech interface (Singapore, Singapore, 2014), pp. 1013–1017
88. S. Hahm, D. Heitzman, J. Wang, in *6th Workshop on Speech and Language Processing for Assistive Technologies*. Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization, vol 11 (Dresden, Germany, 2015), pp. 47–54
89. M. Kim, J. Wang, H. Kim, in *17th Annual Rence of the International Speech Communication Association (INTERSPEECH 2016)*. Dysarthric speech recognition using Kullback-Leibler divergence-based hidden Markov model (San Francisco, USA, 2016), pp. 2671–2675. https://doi.org/10.21437/Interspeech.2016-776
90. M. Kim, Y. Kim, J. Yoo, J. Wang, H. Kim, Regularized speaker adaptation of KL-HMM for dysarthric speech recognition. IEEE. Trans. Neural. Syst. Rehabil. Eng. **25**(9), 1581–1591 (2017). https://doi.org/10.1109/TNSRE.2017.2681691
91. E. Yilmaz, M.S. Ganzeboom, C. Cucchiarini, H. Strik, in *18TH Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*. Multi-stage DNN training for automatic recognition of Dysarthric speech (Stockholm, Sweden, 2017), pp. 2685–2689. https://doi.org/10.21437/Interspeech.2017-303
92. F. Xiong, J. Barker, H. Christensen, in *Speech Communication. 13th ITG-Symposium*. Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition, vol 16 (VDE, Oldenburg, Germany, 2018), pp. 1–5
93. M. Kim, B. Cao, K. An, J. Wang, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*. Dysarthric speech recognition using convolutional LSTM neural network (Hyderabad, 2018), pp. 2948–2952. https://doi.org/10.1109/10.21437/speech.2018-2250
94. J.W. Yu, X.R. Xie, S.S. Liu, S.K. Hu, M.E.K. Lam, X.X. Wu, K.H. Wong, X.Y. Liu, H. Meng, in *19TH Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*. Development of the CUHK Dysarthric speech recognition system for the UA speech Corpus (Hyderabad, India, 2018), pp. 2938–2942. https://doi.org/10.21437/Interspeech.2018-1541
95. E. Hermann, M.M. Doss, in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Dysarthric speech recognition with lattice-free MMI (Barcelona, Spain, 2020), pp. 6109–6113. https://doi.org/10.1109/ICASSP40776.2020.9053549

96. M.S. Yakoub, S.A. Selouani, B.F. Zaidi, A. Bouchair, Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. Eurasip J. Audio Speech Music Process. **1**, 1–7 (2020). https://doi.org/10.1186/s13636-019-0169-5
97. L.D. Wu, D.M. Zong, S.L. Sun, J. Zhao, in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A sequential contrastive learning framework for robust Dysarthric speech recognition (IEEE, Toronto, ON, Canada, 2021), pp. 7303–7307. https://doi.org/10.1109/ICASSP39728.2021.9415017
98. D. Wang, J. Yu, X. Wu, L.F. Sun, X.Y. Liu, H.E. Meng, in *12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization (IEEE, Hong Kong, China, 2021), pp. 1–5. https://doi.org/10.1109/ISCSLP49672.2021.9362068
99. S.R. Shahamiri, Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. IEEE Trans. Neural Syst. Rehabilitation Eng. **29**, 852–861 (2021). https://doi.org/10.1109/TNSRE.2021.3076778
100. S.K. Hu, X.R. Xie, M.Y. Cui, J.J. Deng, S.S. Liu, J.W. Yu, M.Z. Geng, X.Y. Liu, H.E. Meng, Neural architecture search for LF-MMI trained time delay neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **30**, 1093–1107 (2022). https://doi.org/10.1109/TASLP.2022.3153253
101. Z. Yue, E. Loweimi, H. Christensen, J. Barker, Z. Cvetkovic, Acoustic modelling from raw source and filter components for dysarthric speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **30**, 2968–2980 (2022). https://doi.org/10.1109/TASLP.2022.3205766
102. Z. Yue, E. Loweimi, H. Christensen, J. Barker, Z. Cvetkovic, in *INTERSPEECH 2022 ISCA-INST SPEECH COMMUNICATION ASSOC*. Dysarthric Speech Recognition From Raw Waveform with Parametric CNNs (Incheon, Korea, 2022), pp. 31–35. https://doi.org/10.21437/Interspeech.2022-163
103. E. Loweimi, Z. Yue, P. Bell, S. Renals, Z. Cvetković, Multi-stream acoustic modelling using raw real and imaginary parts of the Fourier transform. IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 876–890 (2023). https://doi.org/10.1109/TASLP.2023.3237167
104. D. Mulfari, A. Celesti, M. Villari, in *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. Exploring AI-based Speaker Dependent Methods in Dysarthric Speech Recognition (IEEE, Taormina, Italy, 2022), pp. 958–964. https://doi.org/10.1109/CCGrid54584.2022.00117
105. M. Geng, X. Xie, Z. Ye, T. Wang, G. Li, S. Hu, X. Liu, H. Meng, Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **30**, 2597–2611 (2022). https://doi.org/10.1109/TASLP.2022.3195113
106. B. Vachhani, C. Bhat, B. Das, S.K. Kopparapu, in *18th Annual Conference of the International-Speech-Communication-Association (INTERSPEECH 2017)*. Deep autoencoder based speech features for improved Dysarthric speech recognition (Stockholm, Sweden, 2017), pp. 1854–1858. https://doi.org/10.21437/Interspeech.2017-1318
107. Y. Takashima, R. Takashima, T. Takiguchi, Y. Ariki, Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition. IEEE Access. **7**, 164320–164326 (2019). https://doi.org/10.1109/ACCESS.2019.2951856
108. F.F. Xiong, J. Barker, Z.J. Yue, H. Christensen, in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Source domain data selection for improved transfer learning targeting dysarthric speech recognition (IEEE, Barcelona, Spain, 2020), pp. 7424–7428. https://doi.org/10.1109/ICASSP40776.2020.9054694
109. Y. Takashima, T. Nakashika, T. Takiguchi, Y. Arikii, in *23rd European Signal Processing Conference (EUSIPCO)*. Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition (Nice, France, 2015), pp. 1411–1415. https://doi.org/10.1109/EUSIPCO.2015.7362616
110. E. Yilmaz, V. Mitra, G. Sivaraman, H. Franco, Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech. Comput. Speech Lang. **58**, 319–334 (2019). https://doi.org/10.1016/j.csl.2019.05.002
111. B.F. Zaidi, S.A. Selouani, M. Boudraa, M.S. Yakoub, Deep neural network architectures for dysarthric speech analysis and recognition. Neural Comput. Applic. **33**(15), 9089–9108 (2021). https://doi.org/10.1007/s00521-020-05672-2
112. A. Revathi, R. Nagakrishnan, N. Sasikaladevi, Comparative analysis of Dysarthric speech recognition: Multiple features and robust templates. Multimed. Tools Appl. **81**(22), 31245–31259 (2022). https://doi.org/10.1007/s11042-022-12937-6
113. R. Rajeswari, T. Devi, S. Shalini, Dysarthric speech recognition using Variational mode decomposition and convolutional neural networks. Wirel. Pers. Commun. **122**(1), 293–307 (2022). https://doi.org/10.1007/s11277-021-08899-x
114. A Hernandez, PA Perez-Toro, E Noth, JR Orozco-Arroyave, A Maier, SH Yang, Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition. Interspeech 2022, Incheon, Korea. 2022, 51–55. https://doi.org/10.21437/Interspeech.2022-10674
115. P. Wang, H Van hamme, benefits of pre-trained mono-and cross-lingual speech representations for spoken language understanding of Dutch dysarthric speech. Eurasip J. Audio Speech Music Process. **2023**(1), 1–25 (2023). https://doi.org/10.1186/s13636-023-00280-z
116. M.K. Baskar, T. Herzigy, D. Nguyen, M. Diez, T. Polzehl, L. Burget, J. Cernocky, *Speaker adaptation for Wav2vec2 based dysarthric ASR* (Interspeech 2022, Incheon, Korea, 2022), pp. 3403–3407. https://doi.org/10.21437/Interspeech.2022-10896
117. L.P. Violeta, W.C. Huang, T. Toda, *Investigating self-supervised pretraining frameworks for pathological speech recognition* (Incheon, Korea, 2022), pp. 41–45. https://doi.org/10.21437/Interspeech.2022-10043
118. W.K. Seong, J.H. Park, H.K. Kim, Multiple pronunciation lexical modeling based on phoneme confusion matrix for dysarthric speech recognition. Adv. Sci. Technol. Lett. **14**, 57–60 (2012)
119. WK Seong, JH Park, HK Kim, Dysarthric speech recognition error correction using weighted finite state transducers based on context-dependent pronunciation variation. Computers Helping People with Special Needs. ICCHP'12: Proceedings of the 13th international conference on Computers Helping People with Special Needs, Linz, Austria, 11–13 July 2012, Part II, 475–482. https://doi.org/10.1109/10.1007/978-3-642-31534-3_70
120. R. Sriranjani, S. Umesh, M.R. Reddy, in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2015)*. Pronunciation adaptation for disordered speech recognition using state-specific vectors of phone-cluster adaptive training, vol 11 (Dresden, Germany, 2015), pp. 72–78
121. Z. Yue, F. Xiong, H. Christensen, J. Barker, in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition (IEEE, Barcelona, Spain, 2020), pp. 6094–6098. https://doi.org/10.1109/ICASSP40776.2020.9054343
122. Y. Takashima, T. Takiguchi, Y. Ariki, in *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. End-to-end dysarthric speech recognition using multiple databases (IEEE, Brighton, UK, 2019), pp. 6395–6399. https://doi.org/10.1109/ICASSP.2019.8683803
123. Y. Lin, L. Wang, J. Dang, S. Li, C. Ding, in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. End-to-end articulatory modeling for dysarthric articulatory attribute detection (IEEE, Barcelona, Spain, 2020), pp. 7349–7353. https://doi.org/10.1109/ICASSP40776.2020.9054233
124. Y. Lin, L. Wang, S. Li, J. Dang, C. Ding, in *Interspeech*. Staged knowledge distillation for end-to-end Dysarthric speech recognition and speech attribute transcription (Shanghai, China, 2020), pp. 4791–4795. https://doi.org/10.21437/Interspeech.2020-1755
125. M. Soleymanpour, M.T. Johnson, J. Berry, in *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. Dysarthric speech augmentation using prosodic transformation and masking for subword end-to-end ASR (IEEE, Bucharest, Romania, 2021), pp. 42–46. https://doi.org/10.1109/SpeD53181.2021.9587372
126. A. Almadhor, R. Irfan, J. Gao, N. Salleem, H.T. Rauf, S. Kadry, E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition. Expert Syst. Appl. **222**, 119797 (2023). https://doi.org/10.1016/j.eswa.2023.119797
127. S.R. Shahamiri, V. Lal, D. Shah, Dysarthric speech transformer: A sequence-to-sequence Dysarthric speech recognition system. IEEE. Trans. Neural. Syst. Rehabil. Eng. **31**, 3407–3416 (2023). https://doi.org/10.1109/TNSRE.2023.3307020

128. H. McGurk, J. MacDonald, Hearing lips and seeing voices. Nature. **264**(5588), 746–748 (1976)
129. S. Liu, X. Xie, J. Yu, S. Hu, M. Geng, R. Su, S. Zhang, X. Liu, H. Meng, in *Interspeech 2020*. Exploiting cross-domain visual feature generation for disordered speech recognition (Shanghai, China, 2020), pp. 711–715. https://doi.org/10.21437/Interspeech.2020-2282
130. S. Hu, S. Liu, X.R. Xie, M.Z. Geng, T.Z. Wang, S.K. Hu, M.Y. Cui, X. Liu, H. Meng, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Exploiting cross domain acoustic-to-articulatory inverted features for disordered speech recognition (IEEE, Singapore, Singapore, 2022), pp. 6747–6751. https://doi.org/10.1109/ICASSP43922.2022.9746989
131. Z Yue, E Loweimi, Z Cvetkovic, H Christensen, J Barker, Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition. ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Singapore, Singapore, 2022. 7372–7376. https://doi.org/https://doi.org/10.1109/ICASSP43922.2022.9746855
132. C. Yu, X. Su, Z. Qian, Multi-stage audio-visual fusion for Dysarthric speech recognition with pre-trained models. IEEE. Trans. Neural. Syst. Rehabil. Eng. **31**, 1912–1921 (2023). https://doi.org/10.1109/TNSRE.2023.3262001

Publisher's Note