Speech Communication 155 (2023) 103002

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Disordered speech recognition considering low resources and abnormal articulation ☆

Yuqin Lin [a], Jianwu Dang [a,b], Longbiao Wang [a,c,*], Sheng Li [d], Chenchen Ding [d]

[a] Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
[b] Pengcheng Laboratory, Shenzhen, China
[c] Huiyan Technology (Tianjin) Co., Ltd, Tianjin, China
[d] National Institute of Information and Communications Technology (NICT), Kyoto, Japan

ARTICLE INFO

Keywords:
Automatic speech recognition
Speech disorder
Dysarthria
Speech perception

ABSTRACT

The success of automatic speech recognition (ASR) benefits a great number of healthy people, but not people with disorders. The speech disordered may truly need support from technology, while they actually gain little. The difficulties of disordered ASR arise from the limited availability of data and the abnormal nature of speech, *e.g*, unclear, unstable, and incorrect pronunciations. To realize the ASR of disordered speech, this study addresses the problems of disordered speech in two respects, low resources, and articulatory abnormality. In order to solve the problem of low resources, this study proposes staged knowledge distillation (KD), which provides different references to the student models according to their mastery of knowledge, so as to avoid feature overfitting. To tackle the articulatory abnormalities in dysarthria, we propose an intended phonological perception method (IPPM) by applying the motor theory of speech perception to ASR, in which pieces of intended phonological features are estimated and provided to ASR. And further, we solve the challenges of disordered ASR by combining the staged KD and the IPPM. TORGO database and UASEECH corpus are two commonly used datasets of dysarthria which is the main cause of speech disorders. Experiments on the two datasets validated the effectiveness of the proposed methods. Compared with the baseline, the proposed method achieves 35.14%~38.12% relative phoneme error rate reductions (PERRs) for speakers with varying degrees of dysarthria on the TORGO database and relative 8.17%~13.00% PERRs on the UASPEECH corpus. The experiments demonstrated that addressing disordered speech from both low resources and speech abnormality is an effective way to solve the problems, and the proposed methods significantly improved the performance of ASR for disordered speech.

1. Introduction

The speech signal, the carrier of information transmission in human communication, is produced by the axrticulatory system and perceived by the auditory system. Therefore, research on speech signal processing has been focused on the mechanism of human speech production and auditory processing. Speech disorder arises from dysfunctions of motor programming at the higher level or motor execution at the lower level. Auditory comprehension of speech is difficult for listeners because of their lack of exposure to such speech signals of irregularity and distortion. Therefore, the task of ASR for speakers with speech disorder requires special procedures to predict speakers' intension by expanding available phonetic information to reconstruct language's phonology and then linguistic messages.

Automatic speech recognition (ASR) translates speech signals to corresponding text or commands. Studies on conventional speech recognition systems have focused on the acoustic model (Abdel-Hamid et al., 2012; Deliyski, 1993; Juang and Rabiner, 1991; Maas et al., 2013) related to the articulation process and the language model (Gandhe and Rastrow, 2020; Kuhn and De Mori, 1990; Steinbiss and Klakow, 2004) in relation to the text sequence. In recent years, with the help of high-performance deep learning approaches, the acoustic model and the language model are integrated into a single neural network called "end-to-end (E2E) ASR model", which contributes significantly

☆ This work was supported in part by the National Key R&D Program of China (Grant No. 2018YFB1305200), the National Natural Science Foundation of China (Grant No. 62176182), and the National Natural Science Foundation of China (Grant No. 62276185).

* Corresponding author.
E-mail addresses: linyuqin@tju.edu.cn (Y. Lin), jdang@jaist.ac.jp (J. Dang), longbiao_wang@tju.edu.cn (L. Wang), sheng.li@nict.go.jp (S. Li), chenchen.ding@nict.go.jp (C. Ding).

https://doi.org/10.1016/j.specom.2023.103002
Received 14 October 2022; Received in revised form 15 August 2023; Accepted 23 October 2023
Available online 29 October 2023

improvement s to ASR (Amodei et al., 2016; Chan et al., 2016; Dong et al., 2018; Qin et al., 2019; Shan et al., 2019; Zhang et al., 2020).

Unfortunately, the E2E ASR model is heavily dependent on the scale of data, so the recognition of some low-resource speech data is difficult, *e.g.*, minor language speech, dialect speech, and disordered speech. In previous studies, many methods have been used to improve the performance of low-resource ASR (Lin et al., 2020; Meng et al., 2019; Shor et al., 2019; Takashima et al., 2019). Among them, teacher-student learning (TS) achieved competitive results on low-resource data because it makes full use of a large amount of normal speech, transferring knowledge from normal speech to disordered speech since the feature overfitting. However, it omits the problem that blindly learning from the teacher model for normal speech is not conducive to knowledge transfer to the student model for disordered speech. The speech features can be decomposed into crucial features and auxiliary features, whether the former ones describe the common parts between speakers pronouncing the same phone, while the auxiliary ones describe the details of the speech such as the clarity and the style. The disordered speech are usually lost and/or replaced the auxiliary features. If we equally transferred both features from the teacher model to the student model, the student model would not perform well for the disordered speech recognition. This study enhanced the TS to solve the low-resource problem of disordered ASR, aiming to learn crucial features from the teacher model and avoid the model overfitting to auxiliary features.

Speech disorders affect a speaker's ability to produce natural sounds, as seen in stuttering, apraxia of speech, and dysarthria. The disorders reduce the intelligibility of speech to varying degrees. For example, dysarthria—a specific disorder, resulting from weakness or paralysis of speech muscles caused by damages to the motor system (Fritsch and Magimai-Doss, 2021; Narendra and Alku, 2018), may cause the speech unclear, unstable, and mixed with incorrect pronunciation. This paper introduces a dysarthric ASR task as a benchmark for measuring the effectiveness of the proposed methods for low resources and abnormal articulation.

To illustrate the problems of dysarthria, we refer a diagram modified from the Speech Chain (Denes et al., 1993) as shown in Fig. 1. The figure includes two loops; one is between the speaker and a listener, and another is between the speech production and perception in the speaker side. Speech production is an extremely complex process of motor coordination involved in the brain (Asaei et al., 2017). In the process, speaker's intention is translated into linguistic representation. Then, the phonological and phonetic encoding system extract the phonemes, intonation and duration of the language from the representation. Thus the speech planning center in the brain programs articulatory movements according to the linguistic representation. The sounds are produced by motor commands to drive the speech organs—the lungs, larynx, tongue, lips, etc. Any injury in the loop between speech production and perception may cause certain speech disorders. Dysarthria occurs when there is an injury in the motor planning or execution (Xian et al., 2017), as shown in the figure by a cross marker.

Speech perception is the process by which the spoken language is heard, interpreted, and understood. In this process, listeners perceive the speech sound by extracting acoustic cues and phonetic information to infer the articulatory movements (McGurk and MacDonald, 1976). The combination of these cues and this information is often reformed as abstract representations of phonemes and applied to speech recognition (Heba et al., 2019; Lippmann, 1996). Automatic speech attribute transcription (ASAT) is a system for extracting speech perception information. It detects articulatory attributes (e.g., the place and manner of articulation) from speech signals and has the potential to assist disease monitoring (Connaghan et al., 2019).

The motor theory of speech perception is a thesis in psychology accounting the relationship between speech production and perception (Galantucci et al., 2006; Liberman and Mattingly, 1985; Liberman and Mattingly, 1989; Liberman and Whalen, 2000; McGurk and MacDonald, 1976). According to the motor theory of speech perception, speech

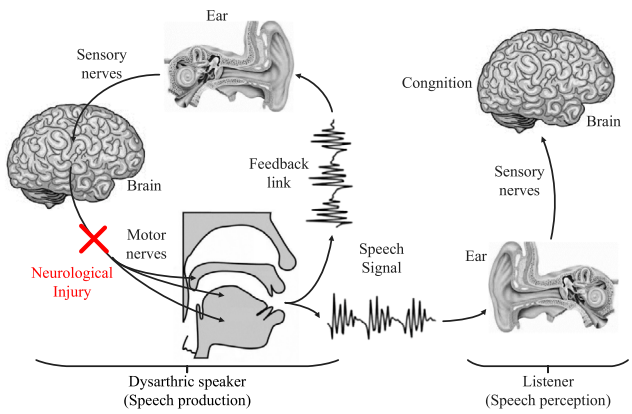

Fig. 1. Speech Chain of dysarthria (production-perception relationship modified from "The Speech Chain" (Denes et al., 1993)).

production and perception are closely related because they share the same speech representation—articulatory gestures. The articulatory gestures are dynamically features composed of a group of abstract and distinctive articulatory features. Normal pronunciation can be fully described by the articulatory features, but in speakers with speech disorders, some of the articulatory features may be lost or replaced. The motor theory suggests that when a listener encounters deformed speech, they may rely heavily on the speaker's articulatory gestures to comprehend the intended message. This phenomenon can be observed in everyday situations. For instance, when a speaker utters during eating, the speech can be perceived correctly, even if it is disturbed by the food in the mouth, because most of the listeners have such experiences. Dysarthria is a specific type of speech disorder that falls under the category of deformed speech. In perception of disordered speech, humans tend to infer the intended phonological features (IPFs) from partial articulatory features, ignoring the lost or replaced articulatory features. This process inspires us to consider perceptual information to solve abnormal problems when recognizing speech. *"How can machines estimate IPFs as humans?"* is the key to applying the motor theory of speech perception to machine speech recognition. The IPFs describe the intended pronunciation of a desired but mispronounced sound. Previous studies showed that the lost or replaced articulatory features tend to be consistent within a speaker, but still vary widely across speakers with speech disorders (Mengistu and Rudzicz, 2011a). Since articulatory movements are coherent, it is possible to use the majority of articulatory features (ignores lost/replaced) to estimate the IPFs of dysarthric speakers.

This study decomposes the challenge of disordered ASR into two sub-problems: a low-resource problem and an abnormality problem because such a speech sound is unclear and unstable with some incorrect pronunciation. In our previous studies (Lin et al., 2020; Lin et al., 2020), we conducted a preliminary study on building a well-performed ASR and ASAT system for dysarthric speech. To solve the low-resource problem of dysarthric ASR, the TS method was enhanced by staged training on dysarthric ASR, which is verified to be effective in avoiding overfitting. Besides, the method is successfully applied to an E2E dysarthric ASAT system. However, our previous studies did not analyze the influence of the stage boundary for dysarthric ASR in-depth and the experimental verification was limited. In addition, we did not consider the speech abnormality problem, which is the primary cause of degradation in dysarthric ASR.

Therefore, to demonstrate the effectiveness of solving the disordered ASR by addressing the two sub-problems, we extended our previous studies with detailed analyses, more experiments and a solution for speech abnormality problem. We first deal with the low-resource problem from the aspect of model training. A model-refactoring method

(Lin et al., 2020) is proposed to use the limited data more effectively. Based on it, a staged knowledge distillation (KD) is proposed to transfer knowledge from an informed teacher model to the dysarthric ASR model (Lin et al., 2020). Then, we further solve the abnormal problem from the perspective of human speech production and perception. We propose an intended phonological perception method (IPPM), in which the IPFs are estimated by reference to the crucial articulatory features (typical features) through an intended phonological perception loss function (IPP-loss). The articulatory features are extracted from an E2E ASAT system. The IPFs of dysarthric speech are used to correct the ambiguous decisions of the phonemes in ASR. Finally, the proposed methods are combined to solve the problems of disordered ASR. Experiments conducted on two popular datasets of dysarthria show that the proposed methods achieve a significant improvement over existing methods.

The remainder of this paper is organized as follows: Section 2 reviews the related work of disordered speech. Section 3 introduces our proposed method. Section 4 describes the experimental evaluations and analysis. Section 5 makes further discussions. Conclusions and plans for future works are presented in Section 6.

2. Related work

This section introduces related work in two areas: end-to-end ASR and dysarthric speech recognition.

2.1. End-to-end speech recognition

A conventional ASR system includes three independent components: an acoustic model, language model and lexicon (Jurafsky, 2000). As a consequence of the development of applications of neural networks, end-to-end approaches to ASR have attracted extensive attention. This approach directly learns the mapping between acoustic features and phonemes (or words) using a single neural network framework, without a language model or lexicon. Graves et al. (2006) proposed connectionist temporal classification (CTC), which laid the foundation for end-to-end speech recognition. Later, Graves and Jaitly (2014) proposed the RNN transducer, which uses a deep bidirectional long short-term memory (LSTM) network with CTC. Ueno et al. (2018) introduced an attention mechanism into ASR and obtained competitive results. The state-of-the-art approaches are based on self-attention, among which transformer-based ASR models are popular, as exemplified by Miao et al. (2020), Salazar et al. (2019), Shetty and Sagaya Mary N.J. (2020), Winata et al. (2020) and Yeh et al. (2019).

2.2. Disordered speech recognition

Related studies on disordered speech recognition focus on speech features and acoustic modeling. Some studies added disordered speakers features to the acoustic features. Deng et al. (2009) used the acoustic signal, the surface EMG (sEMG) signals, and their fusion by considering the articulation of the speakers. In studies on ASR for dysarthria, there were two categories of methods. One is exploring the improvement of ASR at the features level, and another category is exploring the way to fusion acoustic and articulatory features. Xiong et al. (2019) applied speech tempo transformations to reduce mismatches between normal and dysarthric speech. Illa et al. (2018) proposed new articulatory features to capture information from dysarthric speech. The other studies capture the variations of dysarthric speakers' pronunciation and learn the relationships between dysarthric speech features and phonemes or words (Bhat et al., 2018; Hasegawa-Johnson et al., 2006; Kim et al., Kim et al.; Kim et al., 2013). In fusion acoustic and articulatory features, a straightforward approach to fuse the acoustic features and articulatory features is feature concatenation (FC) (Yue et al., 2022). Another popular approach for adding distinguishing articulatory features is to use multitask learning (MTL) (Bayerl et al., 2022; Heba et al., 2019).

Table 1
English consonant list with the articulatory place attributes.

Articulatory place	Phonemes
Labial	p, b, m, f, v
Dental	θ, ð
Alveolar	t, d, n, s, z, l
Post-alveolar	ʧ, ʤ, ʃ, ʒ, r
Palatal	j
Velar	k, g, ŋ, w
Glottal	h

However, at the feature level, previous studies did not consider the case in which a few articulatory features are lost or replaced in dysarthric speech. One focus of this study is dealing with this issue. Besides, these approaches cannot capture perceptual differences between the predicted phonemes and the ground-truth phonemes because they are optimized only by the cross-entropy loss function. As a result, the phonological and articulatory features that the ASR system perceives are inconsistent. The improvement of the performance of disordered speech recognition is limited.

3. Disordered ASR considering low-resource and abnormal articulation

In this section, we first introduce the automatic speech attribute transcription (ASAT) system. Then, we describe the proposed two solutions to the low resources and abnormality challenges in disordered ASR, respectively. Finally, we introduce the final solutions for disordered ASR.

3.1. Automatic speech attribute transcription

Automatic speech attribute transcription (ASAT) is a system for extracting speech perception information. It detects articulatory attributes from speech signals and has the potential to assist disease monitoring (Connaghan et al., 2019; Lippmann, 1996). Articulatory attributes describe the process of human speech production according to the deformation or movement of the lips, tongue, and other speech organs. A consonant can be determined by the attributes of three dimensions: the place and manner of articulation, and distinction between the voiced or voiceless. According to the articulatory place, consonants are divided into the labial, dental, alveolar, post-alveolar, palatal, velar, and glottal. According to the manner of articulation, consonants are divided into the plosives, affricates, nasals, fricatives, approximants, and laterals. Vowels can be determined by three positional attributes: tongue height, tongue backness, and lip rounding. Tongue height describes the tongue's vertical positions, and tongue backness describes its horizontal positions.

In the revised motor theory, speech perception is related to articulatory gestures but not absolute (or typical) articulatory. In this study, we trained an articulatory place detection system following Lin et al. (2020) to provide hidden features of articulatory movement for ASR. It is because the manner features of articulation are easier to obtain from the speech features than the articulatory place features. Therefore, the accurate extraction of the articulatory place features requires additional attribute annotations in the ASAT. The articulatory features obtained from the hidden layer output of the ASAT contain the majority of the articulatory features (not lost/replaced), mixed with the information of the correct place and manner of articulation, irrespective of intended phonological features (IPFs). They have the potential to describe the underlying articulatory gestures since the points on the articulatory gesture can be mapped into a distribution around the typical articulatory place of the phoneme. This paper classifies consonants by the articulatory place using the mapping rules in Lin et al. (2020), which are shown in Table 1. The place attribute of vowels is not taken into account in this paper because vowels have more tolerance for recognition than consonants.

3.2. The solution on low-resource: Staged knowledge distillation

The end-to-end based dysarthric ASR suffered from limited speech data. Knowledge distillation (KD)/Teacher-student learning (TS) is a popular transfer learning method, which has been shown to be effective for adaptation (Li et al., 2017; Meng et al., 2018; Mošner et al., Meng et al.). It makes the student model (dysarthric ASR) learn from a large teacher model (general ASR) and the transcriptions. The traditional KD approaches are not suitable for dysarthric speech due to the low resources of dysarthric speech and its heavy deviation from normal speech. They learn both crucial features and auxiliary features of normal speech equally from the teacher model, which may lead the student model to overfit in the auxiliary features and harms the performance of transferring. This study proposed a staged KD, aiming to learn crucial features from the teacher model and avoid learning auxiliary features since they are usually lost and/or replaced in disordered speech, and overfitting them harms the transfer performance.

Given input speech features $\mathbf{x} = \{x_1, \dots x_L\}$ with length of L, and the corresponding transcriptions $\mathbf{y} = \{y_1, \dots, y_N\}$ with length of N, the teacher network is trained by optimizing the loss between the transcriptions $\mathbf{y}$ and the output softmax of the teacher $\mathbf{o}^t = \{o_1^t, \dots, o_N^t\} \in \mathbb{R}^{N\times D}$. D is the number of target classes. In conditional TS, the student network is trained to learn from selected labels, that is made up of the transcriptions (hard labels) $\mathbf{y}$, and the outputs softmax of the teacher network (soft labels) $\mathbf{o}^t$. The loss function is defined between the selected labels and the outputs softmax of the student network (predicted label) $\mathbf{o}^s = \{o_1^s, \dots, o_N^s\}$ $o_i^s \in \mathbb{R}^{N\times D}$. However, this approach is ineffective when the resources of data are limited as in dysarthric speech. To make a full use of limited data resources, a staged training strategy is adopted. The latest study in Takashima et al. (2020) shows staged training (first adapted to multiple dysarthric speakers, and then to the target speaker) is effective for dysarthric speech. Different from that, the student model in the proposed staged KD is first adapted to the mixture of multiple dysarthric and healthy speakers, and then the adapted model is further adapted for the target dysarthric speakers.

The staged KD has two learning stages. The selected labels are defined differently at different stages. In the first stage, the selected labels are made up of the hard labels $\mathbf{y}$ and soft labels $\mathbf{o}^t$. In the second stage, the selected labels are made up of the hard labels $\mathbf{y}$ and predicted labels $\mathbf{o}^s$. The specific definition is as follows:

$$\tilde{y}_i(o_i) = \begin{cases} o_i^{t/s}, & \underset{j\in\{1,2,\dots,D\}}{\arg\max}\ o_{i,j} = \underset{k\in\{1,2,\dots,D\}}{\arg\max}\ y_{i,k} \\ y_i, & \text{otherwise} \end{cases} \tag{1}$$

where $o_i^{t/s}$ represents the ith output softmax of the teacher network or student network. That is to say, the student at first learns knowledge from both the teacher and ground truth and then focuses on the more difficult aspects when it has learned most of the knowledge.

A boundary value λ is introduced to divide the two stages. The training process enters the second stage when the prediction accuracy of the student network is higher than λ. In brief, the student network is trained to optimize the following loss function:

$$\mathcal{L}_{staged\_KD} = \begin{cases} -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{D} \tilde{y}_{i,j}(o_i^t)\log o_i^s, & \text{acc} \le \lambda \\ -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{D} \tilde{y}_{i,j}(o_i^s)\log o_i^s, & \text{otherwise,} \end{cases} \tag{2}$$

$$\text{acc}_i = \begin{cases} 1, & \underset{j\in\{1,2,\dots,D\}}{\arg\max}\ o_{i,j}^s = \underset{k\in\{1,2,\dots,D\}}{\arg\max}\ y_{i,k} \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

$$\text{acc} = \frac{1}{N}\sum_{i=1}^{N} \text{acc}_i, \tag{4}$$

where $\mathcal{L}_{staged\_KD}$ is the loss function for the proposed staged KD. acc_i indicates whether the model accurately predicts the ith output, and acc is the overall accuracy of prediction. λ is the tunable tradeoff parameter.

3.3. The solution on abnormal articulation: Intended phonological perception method

According to the motor theory of speech perception (Liberman and Mattingly, 1985), listeners would perceive pronunciation heavily reference with articulatory gestures when the sound deteriorates. Indeed, dysarthric speech is deteriorated by missing or replacing a few articulatory features in articulatory gestures. In the perception of normal speech, humans can easily identify phonological features from articulatory gestures, while when it comes to disordered speech, humans may infer the intended phonological features (IPFs) by relying on the majority of typical articulatory features, provided that there is consistency between the perceptual features and articulatory features. The coordination of movements suggests the feasibility of inferring the IPFs from the predominant accurate articulatory features.

In terms of this notion, we propose the intended phonological perception method (IPPM), which consists of two subsystems that aim to learn both articulatory and auditory information from speech sounds. These subsystems comprise a number of layers and contain various levels of articulatory and auditory information in different layers. Our objective was to identify the layers in the two subsystems which have the highest relation between the articulation and auditory, and optimize the ASR learning process through their interactions. The IPPM estimating IPFs through minimizing the difference in the perceptual phonological features and articulatory features. The phonological perceptual features are mapped to articulatory features in an optimized measure, thereby allowing for the inference of IPFs. The two features are extracted from the decoder of the ASR and ASAT, respectively. The difference is defined by an IPP loss. The IPFs are incorporated into ASR to correct the ambiguous decisions of the phonemes by optimizing the joint loss functions. Therefore, IPFs refer to the phonetic characteristics that speakers aim to produce based on the category of phonemes and considering the majority of articulatory features. Fig. 2 shows the schematic diagram of the IPPM.

Given input speech features $\mathbf{x} = \{x_1, \dots x_L\}$, with length L, corresponding phoneme sequences $\mathbf{y}^{ASR} = \{y_1^{ASR}, \dots y_N^{ASR}\}$, with length N, and corresponding articulatory place sequences $\mathbf{y}^{ASAT} = \{y_1^{ASAT}, \dots y_N^{ASAT}\}$, with length N, the predicted phoneme sequences $\tilde{\mathbf{y}}^{ASR}$ and predicted articulatory place sequences $\tilde{\mathbf{y}}^{ASAT}$ can be obtained from an ASR model and ASAT model, respectively. The phonological features T_i^{phono} and articulatory features T_i^{arti} are extracted from the ith layer decoder output of a training ASR system and a trained ASAT system, respectively. Where T_i^{phono} and T_i^{arti} has the same shape of $B_i \times N_i \times D_i$, and B_i, and Di denote the number of sentences per batch and the dimension of features, respectively. Features are from the decoder because the layers in the decoder contain abstract information about the mapping between speech features and the corresponding phoneme or articulatory place, which the encoder does not.

To estimate the IPFs, the IPP-loss gives the mean square error (MSE) between T_i^{arti} and T_i^{phono}. The use of MSE rather than other loss functions such as cross-entropy loss, is the fact that the outputs of the decoder layers are likelihood probabilities. The IPP-loss function can be written as

$$\mathcal{L}_{IPP} = (B_i N_i D_i)^{-1} \| T_i^{\text{phono}} - T_i^{\text{arti}} \|_{\text{F}}^2. \tag{5}$$

We compute the IPP-loss between decoders because the decoder of the Transformer has the potential to extract features related to the phonemes and articulatory information. Encoders tend to extract shallow generic speech representations because it does not obtain accurate phoneme or articulation information.

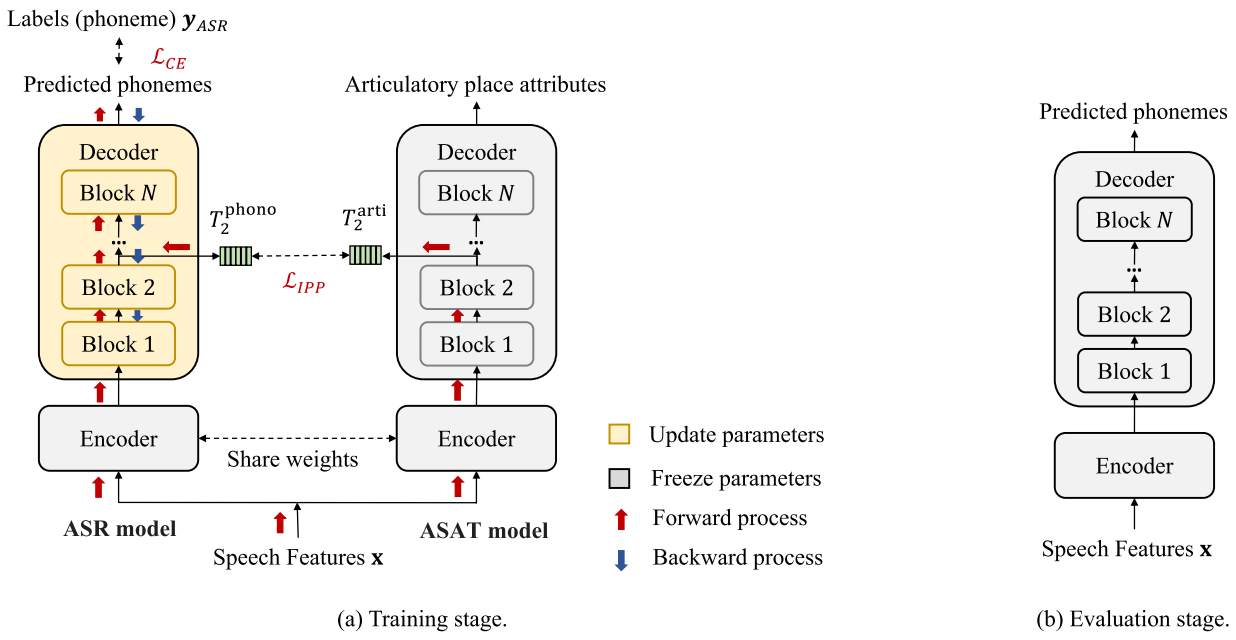

Fig. 2. Schematic diagram of the IPPM with the case where $i = 2$.

To incorporate the IPFs into ASR, the model is trained to jointly optimize two loss functions: cross-entropy loss $\mathcal{L}_{CE}$ and IPP-loss $\mathcal{L}_{IPP}$. The total loss function is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \beta\mathcal{L}_{IPP}, \tag{6}$$

where β is a tunable parameter, representing the degree of reliance of ASR on IPFs. Once the training process is completed, the lower i layers of the ASR decoder acquire the ability to estimate IPFs directly from the speech input. All the experiments in this study were conducted with $\beta = 1$. We emphasize that we paid less attention for tuning the β, since the major purpose of this study is to find out an effective way to estimate IPFs of dysarthric speech and take them into account in ASR.

3.4. Final system: Combining the staged KD and the IPPM

This study solves the challenge of disordered ASR by dealing with two sub-problem: low resources and abnormal articulation. For the low resources problem, we propose the staged KD to more effectively use the limited resources. For the abnormal articulation problem, we propose the IPPM, which estimates IPFs and uses the features to correct ambiguous phonemes. Based on them, we develop the final system by combining the two solutions. The performance of disordered ASR is further improved. Specifically, we joint optimizing the loss function in the staged KD and the IPPM. The final loss function can be expressed as:

$$\mathcal{L}_{final} = \mathcal{L}_{staged\_KD} + \beta\mathcal{L}_{IPP}, \tag{7}$$

where β is a tunable parameter, which is set to 1 same as in Section 3.3.

4. Experiments and results

4.1. Datasets

A series of experiments were conducted on two open-source corpus of speakers with dysarthria to evaluate the effectiveness of the proposed method: the TORGO database[1] and the UASpeech.[2] In addition, another 500-h recording of normal speech from the Librispeech corpus (Panayotov et al., 2015) was used for pretraining in the E2E framework because of the limited amount of dysarthric speech data.

TORGO The TORGO database (Rudzicz et al., 2012) consists of data from 8 dysarthric speakers with varying degrees of cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), and 7 healthy control speakers. There are three levels of speech disorders: severe, moderate, and mild. Speech files in the dataset are recorded by a microphone array and a head-worn microphone with a 16 kHz sampling rate. The recordings last 20.4 h (8.7 h for dysarthric and 11.7 h for typical control speech) after applying standard three-way speed perturbation with factors of 0.9, 1.0, and 1.1 (Ko et al., 2015).

UASPEECH The UASPEECH corpus consists of data from 15 dysarthric speakers with CP and 13 healthy control speakers. Speakers were classifiered in four severity levels, namely severe, severe to moderate, moderate and mild (Kim et al., 2008). Speech files in the dataset are recorded by 7 microphones with a 16 kHz sampling rate. The original recording contains long silent segments at the beginning and end of the audio. After cleaning up the data as in Xiong et al. (2018), the recordings last 78.5 h, where 47.8 h for dysarthric speech and 30.7 h for typical control speech.

4.2. Experimental setup

All experiments were conducted using the Kaldi speech recognition toolkit (Povey et al., 2011) and the open-source Transformer-based machine translation model in tensor2tensor.[3] Due to the limited training data, all models for ASR or ASAT are pre-trained with 500-h labeled speech in the Librispeech corpus. The E2E baseline is the model fine-tuned with the target datasets. The details of the general settings for all models in the experiments are as follows.

Banks of log Mel-filterbank energy features were used as input speech features. They were computed with a sliding window of 25 ms wide, shifted by 10 ms each time step. The log Mel-filterbank energy features were stacked by 40 dimensional static filterbank features, and the delta- and double-delta filterbank features. After computing the

[1] http://ifp-08.ifp.uiuc.edu/protected/UASPEECH/.
[2] https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html.
[3] https://github.com/tensorflow/tensor2tensor.

Table 2
Comparison of dysarthric speech recognition performance (PER%) of different methods on the TORGO database and the UASPEECH corpus. 'S/M' indicates speakers wore severe or moderate degree of dysarthria.

Framework	Method	Articulatory features	TORGO			UASPEECH		
			S/M	Mild	Avg.	S/M	Mild	Avg.
Conventional	GMM-HMM	w/o	87.01	77.69	84.45	99.21	67.70	80.90
	DNN-HMM	w/o	80.22	72.96	77.84	90.26	55.40	73.52
E2E	E2E baseline	w/o	43.57	17.89	33.08	56.94	45.04	52.65
	Model-ref. (Lin et al., 2020)	w/o	33.11	12.47	24.69	68.47	49.24	61.53
	Conditional TS (Meng et al., 2019)	w/o	29.67	9.83	21.57	57.16	44.06	52.43
	Staged KD (Lin et al., 2020) (ours)	w/o	**28.41**	**9.05**	**20.51**	**56.58**	**43.60**	**51.90**
	FC (Yue et al., 2022)	w/	30.51	10.55	22.36	56.56	42.86	51.62
	MTL (Bayerl et al., 2022)	w/	30.86	9.54	22.15	57.23	44.48	52.63
	IPPM (ours)	w/	**28.87**	**9.37**	**20.91**	**52.98**	**39.06**	**47.96**
Final system	Staged KD + IPPM	w/	**28.26**	**9.17**	**20.47**	**52.29**	**37.93**	**47.11**

normalized mean and variance for each speaker, the three left frames were spliced with the current frame. In this case, the input speech features had a dimension of 480.

The Transformer models for ASR or ASAT had an encoder with six layers and a decoder with six layers. The number of heads in the multihead attention layers was set to 8. All sublayers in the model, and the input/output embedding layers, had a dimension of 512. In ASR, the vocabulary set contained 39 phonemes according to the CMU pronouncing dictionary (Weide, 1998). In ASAT, the vocabulary set contained the articulatory places, mapped from the 39 phonemes.

Early-stopping monitors training steps by the performance of the model on the held-out validation set. However, it is not an efficient for dysarthric ASR. It is because the speech sounds vary greatly for each speaker even if the type and degree of the speech diseases are the same, and applying early stopping only on the data of several speakers benefits those speakers, but their speech may be quite different from others. In dysarthric ASR, a common approach is to set a fixed number of iteration training steps (Xiong et al., 2018; Xiong et al., 2020; Huang et al., 2022). In this study, we set 60 epochs for the TORGO database and 20 epochs for the UASPEECH corpus according to the model convergence. The training model was saved every 200 steps, and the parameters of the last 20 saved models were averaged to avoid overfitting. The division of training set and test set for the TORGO database and the UASPEECH corpus are the same as in Lin et al. (2020) and in Xiong et al. (2018), respectively.

When training, we set the max length of the minibatch as 16 000 frames on the TORGO database and 10 000 frames on the UASPEECH corpus. All the models were optimized by the Adam optimizer with a warm-up learning rate (Vaswani et al., 2017). The maximum learning rate was set to 1 on the TORGO database and 0.1 on the UASPEECH corpus. When evaluating, the beam search algorithm, with a beam size of 20, was used for decoding the ASR system. Any ASAT component is not required. All the models in our experiments have the same schematic diagram of the evaluation stage.

4.3. Results

Table 2 shows the phoneme error rate (PER%) of ASR for speakers with different degrees of dysarthria achieved by different methods on the TORGO database and the UASPEECH corpus. 'S/M' indicates speakers wore severe or moderate degree of dysarthria. The conventional-based ASR methods are being compared. Besides, we applied the feature concatenate (FC) (Yue et al., 2022) and multi-task learning (MTL) (Bayerl et al., 2022) on the Transformer for comparison on E2E framework. All of these methods, including IPPM, involve the issue of deciding which layer to use for feature fusion. We conducted experiments and presented the best results for each method in Table 2.

From Table 2, E2E models performed significantly better than the conventional methods. The proposed staged knowledge distillation (KD) and intended phonological perception method (IPPM) showed competitive performances compared to other methods on the both datasets. The staged KD is more effective than conditional TS. On the TORGO database, compared with the baseline, the IPPM achieved 33.74% relative phoneme error rate reduction (PERR) on speakers with severe or moderate degree dysarthria and achieved 47.62% relative PERR on speakers with mild degree dysarthria. On the UASPEECH corpus, compared with the baseline, the IPPM achieved 6.95% relative PERR on speakers with severe or moderate degree dysarthria and achieved 13.27% relative PERR on speakers with mild degree dysarthria. The performance improvement of IPPM for dysarthric ASR is more obvious when the data is extremely limited. The model refactoring method (Lin et al., 2020) performs well on the TORGO database while not on the UASPPECH corpus because the parameter sharing between layers degrades the capability of the model. The FC method and MTL method obtains competitive performance while they need more trainable parameters; it easily causes overfitting or insufficient training.

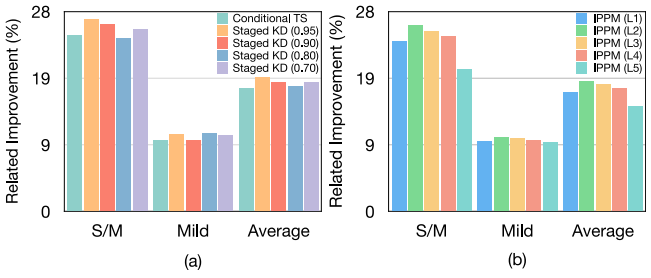

Fig. 3. Visualizing the relative improvement of the proposed methods on the TORGO dataset compared with the baseline. (a) shows the results of the staged KD with different λ values. (b) shows the results of the IPPM that computes IPP-loss on different layers of the decoder.

Fig. 3(a) visualized the relative improvement of the staged KD with λ set as 0.95, 0.90, 0.80, 0.70. The staged KD makes use of the limited data more effectively than the conditional TS. The results show that $\lambda = 0.95$ is the best boundary in the staged KD. Fig. 3(b) visualized the relative improvement on IPPM that computes IPP-loss on different layers of the decoder, compared with the baseline. The IPPMs that compute IPP-loss on each layer all effective for dysarthric ASR. The improvement trend for speakers with varying degrees of dysarthria is almost constant. It achieves the best results when the perceptive features are extracted from the second layers of the ASAT decoders through the IPP-loss.

Table 3
Proportion (%) of the three categories of the mispronunciations on the TORGO database and the UASPEECH corpus, comparing the methods with/without IPPM.

Datasets	Method	Mispronunciation type		
		VV	CC	CrossVC
TORGO	Model-ref.	4.02	5.96	1.86
	+IPPM	3.4	5.09	1.67
	RC (%)	15.42	14.60	**10.22**
UASPEECH	E2E-baseline	5.86	18.05	5.35
	+IPPM	5.6	17.26	4.66
	RC (%)	4.44	4.38	**12.90**

Table 4
Proportion (%) of the four types of the errors in defective consonants on the TORGO database and the UASPEECH corpus, comparing the methods with/without IPPM.

Datasets	Method	Error type			
		PE	ME	VE	MixE
TORGO	Model-ref.	6.80	12.97	3.87	5.09
	+IPPM	5.93	12.27	3.42	4.33
	RC (%)	**12.79**	5.40	11.63	**14.93**
UASPEECH	E2E-baseline	19.79	26.30	10.64	18.22
	+IPPM	19.01	24.86	10.54	17.07
	RC (%)	3.94	**5.48**	0.93	**6.31**

4.4. Analysis

This section mainly analysis the function of the IPPM on phoneme decision correction. We roughly divide mispronounced phonemes into three categories: mispronunciations between vowels (VV), mispronunciations between consonants (CC), and crossing mispronunciations between vowels and consonants (Cross VC). Table 3 shows the proportion of the three categories of mispronunciations on the TORGO database and the UASPEECH corpus. We compared the methods without or with IPPM and calculated the relative correction (RC%). From the table, the IPPM is effective on all categories of mispronunciations. The relative correction for the three mispronunciations varied according to the datasets. It is worth noting that IPPM showed a significant effect on feature correction for the Cross VC category in both datasets. It suggests that restoring IPFs can effectively improve the machine's recognition of vowel and consonant sounds. Therefore, it can be inferred that the IPPM is easier to estimate vowel-consonant-distinctive IPFs from the typical features.

Consonants in dysarthric speech are more difficult than vowels. This is evidenced by the high proportion of the CC category in both datasets, shown in Table 3. For further analysis, we classify the defective consonants into four types according to the place and manner of articulation, and voiced or voiceless. Details are as follows:

(1) Place error (PE): intended consonants have defective articulatory places in pronunciation (*e.g.*, confuse /p/ with /t/ or /s/).
(2) Manner error (ME): intended consonants have wrong manner of pronunciation (*e.g.*, confuse /f/ with /p/ or /t/).
(3) Voicing error (VE): intended consonants have incorrect voice type of voiced/voiceless (*e.g.*, confuse /s/ with /z/, or vise versa).
(4) Mixture error (MixE): intended consonants have a mixed error that consists of more than one type of errors described above (*e.g.*, confuse /p/ with /s/ or /r/).

Table 4 shows the proportion and relative correction of the four types of the errors, comparing the models with/without the IPPM. One can seen that the estimation is able to recover the IPFs from the defective articulation of the consonants. Especially, the IPPM demonstrated a powerful capability in recovering the IPFs from the deteriorate articulation with wrong manners as well as the one with mixed articulatory errors when the data is extremely limited. As the amount of data increases, the role of IPPM in recovering IPFs decreases, and one possible reason is the powerful data-based learning ability of E2E ASR. Still, IPPM can correct mispronunciation at the hidden feature level. The statistical results verify the important role of estimated IPFs in phoneme correction in speech recognition. In addition, it suggests the necessity of recovering the IPFs in dysarthric speech recognition. It is worth noting that the IPFs are not so effective in recovering the voicing error type. A possible reason is that the voicing features are lost in the ASAT system. This may prompt ideas of improving the method by considering voicing attributes.

5. Discussions

In the above sections, we introduce two proposed methods for solving disordered ASR. To deal with the low-resource problem, the staged knowledge distillation transfers the knowledge from the teacher model in the stage to avoid the model overfitting to the auxiliary features in normal speech. The performance shown in Section 4.3 validated the necessity. The intended phonological perception method (IPPM) addressed the challenge of the abnormality of speech by applying the motor theory of speech perception to disordered ASR. This section further reveals the mechanism of IPPM, mainly discussing the function of the IPP-loss in the IPPM. The inspiration gained from these discussions may prompt ideas for improved methods and future research.

The Speech Chain formed by speech production and perception plays an important role in human communication. According to the motor theory of speech perception (Liberman and Mattingly, 1985), speech production and perception are sharing the same articulatory gestures. The articulatory gestures are made up of a group of distinctive articulatory features. In the perception of normal speech, articulatory features are regarded as decisive factor for phoneme decision. Disordered speech such as dysarthric, lost or replaced partial articulatory features, and deteriorated integrity of the articulatory gestures. However, human can estimate the intended phonological features (IPFs) by means of the articulatory features no matter that some of them are lost. According to the potential links between speech production and perception, and the continuity of articulatory movement, this study aims at finding an effective method to estimate the IPFs in dysarthric speech for reconstructing language phonology and then provide the IPFs for ASR to guide the ambiguous phonemes decision. In principle, the IPP-loss is an implementation based on the motor theory of speech perception (Galantucci et al., 2006).

In terms of implementation structure, the IPP-Loss verifies the characterization function of different levels of neural network. Computer vision studies demonstrated that the bottom layers of the deep neural network are good at describing the overt features, while the higher layers are good at the abstract features (Yosinski et al., 2014; Johnson et al., 2016). For example, lower layers tend to represent the simple geometric parts of visual images and higher layers preserve overall spatial structure but lost color, texture, and exact shape (Yosinski et al., 2015). According to the analogue of audio and visual cognitions (Zhou et al., 2018), we speculate that the function of the lower layers mainly reflects simple features of the articulation, while the higher layers correspond to the abstract articulatory gestures. Therefore, we deduce the function of IPP-loss by referring to the process of image processing. The IPP-loss estimated the IPFs in articulatory gestures by minimizing the perceptive differences between phonological features and articulatory features. The effectiveness of IPP-loss on the second layer of decoders demonstrates that the second layer plays important role in obtaining the IPFs of abnormal speech and shares common gestures between speech production and perception. According to the architecture of the Transformer (Dong et al., 2018), the decoder of the ASAT system accepts articulatory place attributes in the first layer, and the encoded speech features between each layer. It can be inferred

that the first layer of the ASAT decoder extracted discrete features of articulatory place attributes, and the second layer of the ASAT decoder extracted more continuous and abstract features. These features seems to the articulatory features sharing between speech production and perception, making sense to recover the IPFs in ASR.

In summary, the IPP-loss function a bridge between speech production and perception in the articulatory gesture level. It realizes the unification of theory and technology.

Previous studies (Mengistu and Rudzicz, 2011b) shown that recognition errors of dysarthric speech primarily arise from imprecise articulation and improper breathing, particularly in multiword utterances. In this study, our focus primarily rested on addressing imprecise articulation, with insufficient consideration given to addressing breathing irregularities. The errors resulting from improper breathing were not excluded in this study, where they will be decreased after using a garbage model. Consequently, we will incorporate this aspect to effectively address these errors in future work. In addition, many of Davide Mulfari's works (Mulfari et al., 2023, 2022a,b), proposing methods to build high-accuracy voice servers for dysarthric speakers, are highly informative. For example, they used self-supervised systems of controlled vocabulary to improve ASR performance. However, in our work, our test sets provided no additional information apart from the audio recordings. Words in the test set may not appear in the training set. Therefore, the controlled vocabulary could not be directly applied in this work. In the future, we will explore the incorporation of a language model to govern the selection of output words.

6. Conclusion

This study focused on ASR for disordered speakers considering low resources and abnormal articulation. To solve the low-resource problem, we proposed staged knowledge distillation, which staged transfers knowledge from general ASR to dysarthric ASR according to the accuracy of the model, avoiding the model overfitting to the auxiliary features in the normal speech. To solve the speech abnormality problem, we proposed the intended phonological perception method by retrieving the intended phonological features (IPFs) in ASR for correcting ambiguous phoneme decisions. Experiments on the TORGO database and UASPEECH dataset confirmed the effectiveness of the proposed methods in solving two problems. Furthermore, the experiments demonstrated the necessity of recovering the IPFs in disordered speech on correcting ambiguous phoneme decisions. Finally, this paper suggests to enhance the method by considering voicing attributes in the future research.

CRediT authorship contribution statement

Yuqin Lin: Conceptualization of this study, Methodology, Software, Writing – original draft, Formal analysis, Investigation. **Jianwu Dang:** Supervision, Data analysis, Validation, Writing – review & editing, Resources. **Longbiao Wang:** Supervision, Methodology, Investigation, Writing – review & editing. **Sheng Li:** Investigation, Validation, Writing – review & editing. **Chenchen Ding:** Formal analysis, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Penn, G., 2012. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proc. ICASSP. pp. 4277–4280.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al., 2016. Deep speech 2: End-to-end speech recognition in English and Mandarin. In: Proc. Int. Conf. Mach. Learn.. pp. 173–182.

Asaei, A., Cernak, M., Bourlard, H., 2017. Perceptual information loss due to impaired speech production. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (12), 2433–2443.

Bayerl, S.P., Wagner, D., Nöth, E., Riedhammer, K., 2022. Detecting dysfluencies in stuttering therapy using wav2vec 2.0. arXiv preprint arXiv:2204.03417.

Bhat, C., Das, B., Vachhani, B., Kopparapu, S.K., 2018. Dysarthric speech recognition using time-delay neural network based denoising autoencoder. In: Proc. INTERSPEECH. pp. 451–455.

Chan, W., Jaitly, N., Le, Q., Vinyals, O., 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: Proc. ICASSP. pp. 4960–4964.

Connaghan, K.P., Green, J.R., Paganoni, S., Chan, J., Weber, H., Collins, E., Richburg, B., Eshghi, M., Onnela, J.-P., Berry, J.D., 2019. Use of beiwe smartphone app to identify and track speech decline in amyotrophic lateral sclerosis (ALS). In: Proc. INTERSPEECH. pp. 4504–4508.

Deliyski, D.D., 1993. Acoustic model and evaluation of pathological voice production. In: Proc. Conf. Speech Communicat., Technol..

Denes, P.B., Denes, P., Pinson, E., 1993. The Speech Chain. Macmillan.

Deng, Y., Patel, R., Heaton, J.T., Colby, G., Gilmore, L.D., Cabrera, J., Roy, S.H., Luca, C.J.D., Meltzner, G.S., 2009. Disordered speech recognition using acoustic and sEMG signals. In: Proc. Annu. Conf. Int. Speech Communicat. Associat..

Dong, L., Xu, S., Xu, B., 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In: Proc. ICASSP. pp. 5884–5888.

Fritsch, J., Magimai-Doss, M., 2021. Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features. IEEE Signal Process. Lett. 28, 224–228.

Galantucci, B., Fowler, C.A., Turvey, M.T., 2006. The motor theory of speech perception reviewed. Psychon. Bull. Rev. 13 (3), 361–377.

Gandhe, A., Rastrow, A., 2020. Audio-attention discriminative language model for ASR rescoring. In: Proc. ICASSP. pp. 7944–7948.

Graves, A., Fernández, S., Gomez, F., Schmidhuber, J., 2006. Connectionist temporal classification: Labeling unsegmented sequence data with recurrent neural networks. In: Proc. Int. Conf. Mach. Learn.. pp. 369–376.

Graves, A., Jaitly, N., 2014. Towards end-to-end speech recognition with recurrent neural networks. In: Proc. Int. Conf. Mach. Learn.. pp. 1764–1772.

Hasegawa-Johnson, M., Gunderson, J., Perlman, A., Huang, T., 2006. HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria. In: Proc. ICASSP, Vol. 3. pp. III–III.

Heba, A., Pellegrini, T., Lorré, J.-P., André-Obrecht, R., 2019. Char+ CV-CTC: combining graphemes and consonant/vowel units for CTC-based ASR using multitask learning. In: Proc. INTERSPEECH. pp. 1611–1615.

Huang, W.-C., Halpern, B.M., Violeta, L.P., Scharenborg, O., Toda, T., 2022. Towards identity preserving normal to dysarthric voice conversion. In: Proc. ICASSP. IEEE, pp. 6672–6676.

Illa, A., Patel, D., Yamini, B., Shivashankar, N., Veeramani, P.-K., Polavarapui, K., Nashi, S., Nalini, A., Ghosh, P.K., et al., 2018. Comparison of speech tasks for automatic classification of patients with amyotrophic lateral sclerosis and healthy subjects. In: Proc. ICASSP. pp. 6014–6018.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: Proc. Europ. Conf. Comput. Vis.. Springer, pp. 694–711.

Juang, B.H., Rabiner, L.R., 1991. Hidden Markov models for speech recognition. Technometrics 33 (3), 251–272.

Jurafsky, D., 2000. Speech & Language Processing. Pearson Education India.

Kim, M.J., Cao, B., An, K., Wang, J., 2018. Dysarthric speech recognition using convolutional LSTM neural network. In: Proc. INTERSPEECH. pp. 2948–2952.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T.S., Watkin, K., Frame, S., 2008. Dysarthric speech database for universal access research. In: Proc. INTERSPEECH.

Kim, M.J., Yoo, J., Kim, H., 2013. Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models. In: Proc. INTERSPEECH. pp. 3622–3626.

Ko, T., Peddinti, V., Povey, D., Khudanpur, S., 2015. Audio augmentation for speech recognition. In: Proc. INTERSPEECH. pp. 3586–3589.

Kuhn, R., De Mori, R., 1990. A cache-based natural language model for speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. 12 (6), 570–583.

Li, J., Seltzer, M.L., Wang, X., Zhao, R., Gong, Y., 2017. Large-scale domain adaptation via teacher-student learning. arXiv preprint arXiv:1708.05466.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21 (1), 1–36.

Liberman, A.M., Mattingly, I.G., 1989. A specialization for speech perception. Science 243 (4890), 489–494.

Liberman, A.M., Whalen, D.H., 2000. On the relation of speech to language. Trends Cogn. Sci. 4 (5), 187–196.
Lin, Y., Wang, L., Dang, J., Li, S., Ding, C., 2020. End-to-end articulatory modeling for dysarthric articulatory attribute detection. In: Proc. ICASSP. pp. 7349–7353.
Lin, Y., Wang, L., Li, S., Dang, J., Ding, C., 2020. Staged knowledge distillation for dysarthric speech recognition and speech attribute transcription. In: Proc. INTERSPEECH. pp. 4791—4795.
Lippmann, R., 1996. Speech perception by humans and machines. In: Proc. Europ. Audit. Bas. Speech Percept. Workshop. pp. 309–316.
Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: Proc. Int. Conf. Mach. Learn., Vol. 30. p. 3.
McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264 (5588), 746–748.
Meng, Z., Li, J., Gong, Y., Juang, B.-H., 2018. Adversarial teacher-student learning for unsupervised domain adaptation. In: Proc. IEEE-ICASSP. pp. 5949–5953.
Meng, Z., Li, J., Zhao, Y., Gong, Y., 2019. Conditional teacher-student learning. In: Proc. IEEE-ICASSP. pp. 6445–6449.
Mengistu, K.T., Rudzicz, F., 2011a. Adapting acoustic and lexical models to dysarthric speech. In: Proc. ICASSP. pp. 4924–4927.
Mengistu, K.T., Rudzicz, F., 2011b. Comparing humans and automatic speech recognition systems in recognizing dysarthric speech. In: Canadian Conf. Artific. Intellig.. Springer, pp. 291–300.
Miao, H., Cheng, G., Gao, C., Zhang, P., Yan, Y., 2020. Transformer-based online CTC/attention end-to-end speech recognition architecture. In: Proc. ICASSP. pp. 6084–6088.
Mošner, L., Wu, M., Raju, A., Parthasarathi, S.H.K., Kumatani, K., Sundaram, S., Maas, R., Hoffmeister, B., 2019. Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. In: Proc. IEEE-ICASSP. pp. 6475–6479.
Mulfari, D., Carnevale, L., Galletta, A., Villari, M., 2023. Edge computing solutions supporting voice recognition services for speakers with dysarthria. In: International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW). IEEE, pp. 231–236.
Mulfari, D., Celesti, A., Villari, M., 2022a. Exploring AI-based speaker dependent methods in dysarthric speech recognition. In: International Symposium on Cluster, Cloud and Internet Computing (CCGrid). IEEE, pp. 958–964.
Mulfari, D., La Placa, D., Rovito, C., Celesti, A., Villari, M., 2022b. Deep learning applications in telerehabilitation speech therapy scenarios. Comput. Biol. Med. 148, 105864.
Narendra, N., Alku, P., 2018. Dysarthric speech classification using glottal features computed from non-words, words and sentences. In: Proc. INTERSPEECH. pp. 3403–3407.
Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books. In: Proc. ICASSP. pp. 5206–5210.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit. In: Proc. IEEE Autom. Speech Recognit. Understanding Workshop.
Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I., Raffel, C., 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: Proc. Int. Conf. Mach. Learn.. pp. 5231–5240.
Rudzicz, F., Namasivayam, A.K., Wolff, T., 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang. Resour. Eval. 46 (4), 523–541.
Salazar, J., Kirchhoff, K., Huang, Z., 2019. Self-attention networks for connectionist temporal classification in speech recognition. In: Proc. ICASSP. pp. 7115–7119.
Shan, C., Weng, C., Wang, G., Su, D., Luo, M., Yu, D., Xie, L., 2019. Investigating end-to-end speech recognition for Mandarin-English code-switching. In: Proc. ICASSP. pp. 6056–6060.
Shetty, V.M., Sagaya Mary N.J., M., 2020. Improving the performance of transformer based low resource speech recognition for Indian languages. In: Proc. ICASSP. pp. 8279–8283.
Shor, J., Emanuel, D., Lang, O., Tuval, O., Brenner, M., Cattiau, J., Vieira, F., McNally, M., Charbonneau, T., Nollstadt, M., et al., 2019. Personalizing ASR for dysarthric and accented speech with limited data. arXiv preprint arXiv:1907.13511.
Steinbiss, V., Klakow, D., 2004. Language model based on the speech recognition history. US Patent 6, 823, 307.
Takashima, Y., Takiguchi, T., Ariki, Y., 2019. End-to-end dysarthric speech recognition using multiple databases. In: Proc. ICASSP. pp. 6395–6399.
Takashima, R., Takiguchi, T., Ariki, Y., 2020. Two-step acoustic model adaptation for dysarthric speech recognition. In: Proc. IEEE-ICASSP. pp. 6104–6108.
Ueno, S., Inaguma, H., Mimura, M., Kawahara, T., 2018. Acoustic-to-word attention-based model complemented with character-level CTC-based model. In: Proc. ICASSP. pp. 5804–5808.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Proc. Advances Neural Inf. Process. Syst.. pp. 5998–6008.
Weide, R.L., 1998. The CMU pronouncing dictionary. URL: http://www.speech.cs.cmu.edu/cgibin/cmudict.
Winata, G.I., Cahyawijaya, S., Lin, Z., Liu, Z., Fung, P., 2020. Lightweight and efficient end-to-end speech recognition using low-rank transformer. In: Proc. ICASSP. pp. 6144–6148.
Xian, Y., Lampert, C.H., Bernt, S., Zeynep, A., 2017. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. IEEE Trans. Pattern Anal. Mach. Intell. PP (99), 1.
Xiong, F., Barker, J., Christensen, H., 2018. Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. In: Speech Communicat; 13th ITG-Symposium. VDE, pp. 1–5.
Xiong, F., Barker, J., Christensen, H., 2019. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In: Proc. ICASSP. pp. 5836–5840.
Xiong, F., Barker, J., Yue, Z., Christensen, H., 2020. Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In: Proc. ICASSP. IEEE, pp. 7424–7428.
Yeh, C.-F., Mahadeokar, J., Kalgaonkar, K., Wang, Y., Le, D., Jain, M., Schubert, K., Fuegen, C., Seltzer, M.L., 2019. Transformer-transducer: End-to-end speech recognition with self-attention. arXiv preprint arXiv:1910.12977.
Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks. In: Proc. Advances Neural Inf. Process. Syst.. pp. 3320–3328.
Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H., 2015. Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579.
Yue, Z., Loweimi, E., Cvetkovic, Z., Christensen, H., Barker, J., 2022. Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition. In: Proc. ICASSP. pp. 7372–7376.
Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., Kumar, S., 2020. Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss. In: Proc. ICASSP. pp. 7829–7833.
Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L., 2018. Visual to sound: Generating natural sound for videos in the wild. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. pp. 3550–3558.