CFDRN: A Cognition-Inspired Feature Decomposition and Recombination Network for Dysarthric Speech Recognition

Yuqin Lin, Longbiao Wang, *Member, IEEE*, Yanbing Yang, and Jianwu Dang, *Member, IEEE*

***Abstract*—As an essential technology in human–computer interactions, automatic speech recognition (ASR) ensures a convenient life for healthy people; however, people with speech disorders, who truly need support from such a technology, have experienced difficulties in the use of ASR. Disordered ASR is challenging because of the large variabilities in disordered speech. Humans tend to separately process different spectro-temporal features of speech in the left and right hemispheres of their brain, showing significantly better ability in speech perception than machines, especially in disordered speech perception. Inspired by human speech processing, this article proposes a cognition-inspired feature decomposition and recombination network (CFDRN) for dysarthric ASR. In the CFDRN, slow- and rapid-varying temporal processors are designed to decompose features into stable and changeable features, respectively. A gated fusion module was developed to selectively recombine the decomposed features. Moreover, this study utilised an adaptation approach based on unsupervised pre-training techniques to alleviate data scarcity issues in dysarthric ASR. The CFDRNs were added to the layers of the pre-trained model, and the entire model is adapted from normal speech to disordered speech. The effectiveness of the proposed method was validated on the widely used TORGO and UASpeech dysarthria datasets under three popular unsupervised pre-training techniques, wav2vec 2.0, HuBERT, and data2vec. When compared to the baseline methods, the proposed CFDRN with the three pre-training techniques achieved 13.73%∼16.23% and 4.50%∼13.20% word error rate reductions on the TORGO and UASpeech datasets, respectively. Furthermore, this study clarified several major factors affecting dysarthric ASR performance.**

***Index Terms*—Adaptation, automatic speech recognition, dysarthria.**

Manuscript received 31 March 2023; revised 30 July 2023; accepted 21 September 2023. Date of publication 26 September 2023; date of current version 20 October 2023. This work was supported by the National Natural Science Foundation of China under Grant 62176182. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hema A Murthy. *(Corresponding authors: Longbiao Wang; Jianwu Dang.)*

Yuqin Lin and Yanbing Yang are with the Tianjin Key Laboratory of Cognitive Computingand Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China (e-mail: linyuqin@tju.edu.cn; yangyanbing@tju.edu.cn).

Longbiao Wang is with the Tianjin Key Laboratory of Cognitive Computingand Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China, and also with the Huiyan Technology Tianjin Company, Ltd., Tianjin 300350, China (e-mail: longbiao_wang@tju.edu.cn).

Jianwu Dang is with the Peng Cheng Laboratory, Shenzhen 518066, China, and also with the Tianjin Key Laboratory of Cognitive Computingand Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China (e-mail: jdang@jaist.ac.jp).

Digital Object Identifier 10.1109/TASLP.2023.3319276

I. Introduction

SPEECH production is a significantly complex process requiring the coordination of various muscles and motor control and also involves the brain [1], [2]. Disorders such as cerebral palsy(CP), amyotrophic lateral sclerosis (ALS), Parkinson's disease, stroke, or traumatic brain injuries can affect this process and result in a speech disorder [3], [4], [5], [6], [7]. These disorders affect the ability of a speaker to produce natural sounds and lead to decreased speech intelligibility and communication impairment, as observed in the cases of stuttering, speech apraxia, and dysarthria [8], [9]. Moreover, people with speech disorders who are also affected by physical disabilities may experience difficulties in remotely controlling the keyboard or mouse and other machine interfaces [10], [11], [12]. Automatic speech recognition (ASR) is an alternative method for alleviating this problem. However, the current ASR systems do not benefit people with speech impairments because their speech varies significantly from normal speech. Therefore, a suitable disordered ASR system is required for speakers with speech disorders. To achieve this, this study aimed to build a speaker-independent ASR system so that the maximum possible patients can benefit from this technology at a limited cost. Moreover, a dysarthric ASR task was introduced as a benchmark to measure the effectiveness of the proposed methods, where dysarthria is a clinical category of neurogenic motor speech disorders associated with muscle weakness [13].

The primary challenges faced by disordered ASR are the high variability in disordered speech and limited amount of available data. Speech disorders affect the articulation of speakers and lead to speech variations that are considerably different from normal speech. Such speech variations are characterised by unclear, unstable, and inaccurate pronunciations [7] and may partly include deletions, substitutions, insertions, and distortions of phonemes [14]. Previous studies have shown that variations in speech patterns occur even when individuals suffer from the same degree of speech disorders [15], [16]. Such high inter- and intraspeaker variabilities significantly reduce the robustness of disordered ASR. To build an ASR system that is suitable for most speakers with varying degrees of speech disorders, a large amount of disordered speech data is required for training. However, obtaining sufficient speech data is challenging because speakers with speech disorders usually struggle with articulation.

To address the problems of disordered ASR, adaptation methods have been extensively deployed and have demonstrated good performance. Two categories of adaptation methods have been explored: a) acoustic adaptation, which transfers knowledge from the model of normal speech to the model of disordered speech [17], [18], [19], [20], and b) pronunciation dictionary adaptation, which carefully designs a pronunciation dictionary for a specific speaker according to his/her articulation [14], [21]. Speech variations directly result in ill-suited acoustic models. Therefore, this study focused on acoustic adaptation. In acoustic adaptation, most previous studies focused on ASR for individual speakers; therefore, they did not consider inter-speaker variability. Our previous study [17] proposed a stage knowledge distillation for multiple speakers. However, only labelled speech was leveraged, which limited the improvement of the disordered ASR. Recently, unsupervised pre-training techniques have demonstrated a strong transfer ability, particularly in low-resource tasks [22], [23], [24], [25], [26], [27]. Adaptation methods based on these approaches have attracted considerable interest. In addition to the direct fine-tuning method, the approaches that add adapters to a pre-trained model have been widely studied [28], [29], [30]. Although these methods achieve parameter-efficient adaptation and competitive performance in low-resource tasks, their performance is worse than that when the entire model is fine-tuned. These methods are suitable for situations in which frequent adaptations are required. Moreover, although the adapters reduced the mismatch between normal and disordered speeches to a certain extent, our experiments showed limited improvements in disordered speech because of the high variabilities in speech.

When compared to machines, the human brain has an extraordinary ability to perceive speech, especially highly variable speech [31], [32]. Human speech processing is a useful tool of reference for machine speech recognition. Typical timescales associated with prominent rhythmic activity in the brain, are often defined as follows: delta, 1–4 Hz; theta, 4–8 Hz; alpha, 8–12 Hz; and gamma, above 30 Hz [33]. Previous research on cortical oscillations and speech processing found that the lower band (slow components) corresponds to comprehension at the syntactic level, while the higher band (rapid components) corresponds to comprehension at the word level [33], [34], [35]. In speech processing by human brain, the right-hemisphere auditory areas are sensitive to slowly varying temporal features, whereas the left-hemisphere auditory areas are sensitive to rapidly varying temporal features [36], [37]. This indicates that the right hemisphere is predominantly responsible for coding stable features of speech, and the left hemisphere predominantly responsible for coding the changeable features of speech [38]. The interaction of the two features affects the transmission and coordination of information between brain regions to complete the speech processing, where slowly and rapidly varying waves exist in each brain area [39], [40]. This processing corresponds to the decomposition of speech features in speech cognition. Speech features can be divided into stable features and changeable features. Similarly, the speech features input into a machine can be treated as a recombination of stable and changeable features. Stable features are the common parts of phonemes under the same perceptual category, which play a decisive role in recognition, but do not vary with context. Changeable features play a supplementary function to phoneme recognition, but are indecisive. The robustness of human speech processing capabilities partially results from the suitable decomposition and organic recombination of the stable and changeable features in a variety of contextual situations, especially for disordered speech. Accordingly, the decomposition and recombination of speech features are more conducive to learning varied speeches, such as dysarthric speech.

Inspired by human speech processing mechanisms, this study proposes a cognition-inspired feature decomposition and recombination network (CFDRN) to address the challenges of disordered ASR. In CFDRN, speech features are decomposed into slow and rapid components using our designed slow- and rapid-varying temporal processors. The slow-varying temporal processor is designed based on fast Fourier transform (FFT) to extract stable speech features. The rapid-varying temporal processor is designed based on multilayer perceptions (MLPs) to extract changing features. Then, correlation weights of the decomposed features are developed from the speech features. Finally, the features are recombined using a gated fusion module (GFM). Similar to the human speech cognitive function, the CFDRN is added into the layers of an original network that is pre-trained with unlabelled normal speech. The entire model is trained using dysarthric speech data. Experiments were conducted on the commonly used TORGO and UASpeech dysarthric corpora. The effectiveness of the proposed method was verified by comparing it with three popular unsupervised pre-training frameworks, wav2vec 2.0, HuBERT, and data2vec.

The primary contributions of this study are summarised as follows.

1) A novel CFDRN is proposed for disordered speech recognition. It is inspired by the human speech cognition mechanism, decomposing the features into slow and rapid components before reasonably recombining them. Experiments conducted under three unsupervised pre-training frameworks on two corpora validated the effectiveness of the CFDRN.
2) A slow-varying temporal processor was designed to extract stable features more effectively and efficiently. The extracted stable features were robust in the analysis of disordered speech and play a decisive role in disordered speech recognition.
3) A GFM was developed for the recombination of stable features and changeable features. The GFM improves the ASR performance, particularly for severe or moderate (S/M) dysarthric speakers.

The remainder of this article is organised as follows. Section II reviews advances in dysarthric speech recognition and unsupervised pre-training methods. Section III describes our proposed method. Section IV presents the experimental results and analyses. Section V provides a detailed discussion of the strengths and limitations of the proposed method. The conclusion and future work are presented in Section VI.

II. Related Work

A. Dysarthric Speech Recognition

Previous research on dysarthric speech recognition has attempted to deal with the large variations in dysarthric speech and limited amount of available data. Recently proposed methods can be classified into three broad categories: a) model architecture based approaches [41], [42], [43] that aid models in recognising dysarthric speech; b) data augmentation approaches that expand the speech data of typical or atypical speakers [44], [45], [46], [47], [48], [49], [50] or transforms normal speech to 'dysarthric-like' speech [51], [52], [53], [54]; c) embedding-based approaches [55], [56], [57] that add auxiliary features such as articulatory features into acoustic models; and d) adaptation approaches that transfer knowledge from an ASR for normal speech to one for dysarthric speech [17], [18], or transfer it further to a model for individual dysarthric speech [19], [20].

Our CFDRN was designed based on a technique that adapts an ASR for normal speech to one for dysarthric speech. In this category, certain studies have applied adaptation to acoustic models, while others have applied adaptation to pronunciation dictionaries. For example, for acoustic model adaptation, Lin et al. [17] proposed a staged knowledge distillation method that takes full advantage of a teacher model that learns from normal speech and avoids overfitting. For pronunciation dictionary adaptation, Mengistu et al. [14] analysed the pronunciation characteristics of dysarthric speakers and designed a suitable lexicon. Sawa et al. [21] utilised a phoneme-recognition model for pronunciation analysis and created an adaptive dictionary for individual dysarthric speakers. This study focused on acoustic model adaptation.

Previous studies have only leveraged labelled speech, which limits the improvement in dysarthric ASR. Recent studies have focused on adapting an unsupervised pre-training model to the target task to obtain a model from the maximum amount of unlabelled data. Direct fine-tuning of the source model is the most commonly used approach, although the performance may be affected by a large mismatch between the source and target data. Therefore, the technique of inserting an adapter into a pre-trained model has garnered interest. The effectiveness of this technique has been verified in several tasks, such as natural language processing [28] and visual recognition [29] tasks. In the speech recognition field, Tomanek et al. proposed a residual adapter for accented and atypical speech recognition [30]. These methods achieve parameter-efficient adaptation and competitive performance on low-resource tasks; however, their performance is worse than that obtained when fine-tuning the entire model. Nevertheless, these methods are suitable for situations in which frequent adaptations are required. Fan et al. [58], [59] proposed an adaptation at the pre-training stage, followed by fine-tuning of the entire model for child speech recognition. This exhibited a higher cost than the adaptation approach reported in this study. Moreover, the experiments in this study demonstrated that these model adaptation approaches are not so effective for speaker-independent dysarthric ASR.

B. Unsupervised Pre-Training Methods

In the speech processing field, the unsupervised pre-training method learns speech representations from a large amount of unlabelled speech and then fine-tunes the model for the downstream target task [23], [24], [25], [26], [27]. Previous studies demonstrated the competitive performance of this technique in dealing with a variety of speeches, particularly in the case of limited available data [22]. Based on the training objective, these methods can be classified as follows: a) generative learning techniques, such as wav2vec 2.0 [24], which recover masked frames by contrastive objectives; b) discriminative learning methods, such as HuBERT [25], wavLM [60], and data2vec [26]. These models predict discrete targets of masked regions from a large amount of speech. HuBERT uses an acoustic discovery system instead of the contrastive learning used in wav2vec 2.0. WavLM further improves HuBERT by transforming the input features. It solves full-stack downstream speech tasks. Data2vec [26] predicts contextualised latent representations through a teacher-student mode. It uses the same learning method for either speech, natural language processing, or computer vision. The unsupervised pre-training method eliminates the requirement for labelled data and has been proven to be effective in low-resource speech tasks [22].

Based on previous work, this study aimed to explore an effective adaptation method for dysarthric ASR using unsupervised pre-training models.

III. Proposed Cognition-Inspired Feature Decomposition and Recombination Network (CFDRN) for Dysarthric ASR

To solve the problems of dysarthric ASR, this study utilised an unsupervised pre-training technique to train a speech representation model using normal speech. Then, we added a CFDRN to the layers of the speech representation model. Finally, we trained the entire network to adapt the speech representation model for dysarthric ASR.

A. Model Architecture

The proposed model architecture uses a Transformer model [61] as its backbone. The pre-trained model is composed of feature extraction and context networks. The feature extraction network comprises a stack of convolutional neural networks, while the context network comprises a Transformer encoder, which consists of blocks of self-attention and feedforward layers. As shown in Fig. 1, the blue parts of the model are first initialised using the unsupervised pre-training techniques with a large amount of normal speech data. Then, the proposed CFDRN is added after the self-attention and feedforward layers of each block, as shown by the yellow parts of the model in the figure. A linear layer is added after the last layer of the CFDRN to predict characters (shown in the green parts of the figure). Finally, the entire model, including the CFDRN modules and others, is updated. Note that we did not distinguish between the training of the CFDRN modules and that of the other modules, we

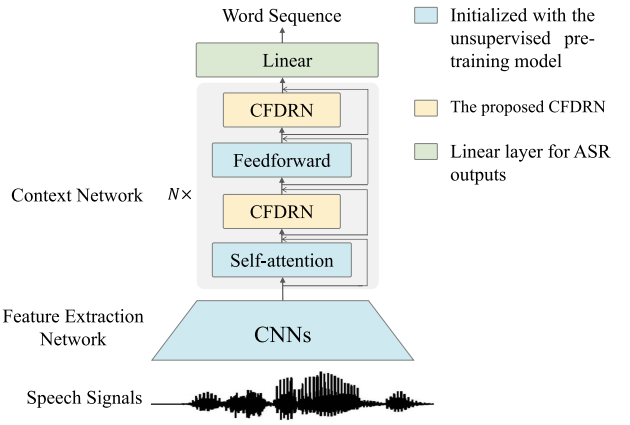

Fig. 1. Overall model architecture of the proposed method.

implemented the functionality of CFDRN through initialization and model architecture alone.

B. The Architecture of CFDRN

The high variabilities in disordered speech lead to difficulties in ASR. The human brain has the extraordinary ability to deal with variations. According to previous studies on human speech perception, the human brain tends to process slow-/rapid-varying temporal features separately in the right-/left-hemisphere auditory areas, respectively [36], [37]. The interaction between these features influences information transmission and coordination across brain regions during speech processing, with the presence of slow and fast oscillations within each brain area [39], [40]. Inspired by this process, we propose the CFDRN. In the CFDRN, we divided the speech features of each layer into two groups, one has more slow components, and the other has more rapid components, and proposed the gated fusion modules (GFM) to realise their interaction through modulation via different activation functions. The architecture of the CFDRN is illustrated in Fig. 2. The central part of the figure shows the components of the CFDRN. The left and right sides of the figure show details of the slow- and rapid-varying temporal processors, respectively. The top of the figure shows the details of the proposed GFM.

Given a hidden speech representation $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_N | \mathbf{h}_i \in \mathbb{R}^D, 1 \leq i \leq N\}$ of length N with D dimensions, we can split $\mathbf{H}$ by an α proportion along the dimension of the feature channels into sets of pre-processed stable features $\mathbf{H}^C$ and changeable features $\mathbf{H}^A$. The shape of $\mathbf{H}^C$ is $N \times D^C$, where $D^C = \alpha \times \mathrm{D}$ and the shape of $\mathbf{H}^A$ is $N \times D^A$, where $D^A = (1-\alpha) \times \mathrm{D}$. $\alpha = 0.5$ indicates that slow and rapid components are equally important in ASR. $\alpha > 0.5$ indicates that the slow components are dominant in ASR, and vice versa. $\mathbf{H}^C$ is fed into the slow-varying temporal processor $\mathcal{P}_{\text{Slow}}$ to extract stable features $\mathbf{F}^C$, whereas $\mathbf{H}^A$ is fed into the rapid-varying temporal processor $\mathcal{P}_{\text{Rapid}}$ to extract changeable features $\mathbf{F}^A$. This process is expressed as follows:

$$\mathbf{H} = [\mathbf{H}^C; \mathbf{H}^A], \tag{1}$$

$$\mathbf{F}^C = \mathcal{P}_{\text{Slow}}(\mathbf{H}^C), \tag{2}$$

$$\mathbf{F}^A = \mathcal{P}_{\text{Rapid}}(\mathbf{H}^A). \tag{3}$$

We use $\mathbf{F}^{RC}$ to denote the recombined features, which are calculated using the GFM $\mathcal{M}_{\text{Fusion}}$. This module selects and combines the stable and changeable features $\mathbf{F}^C$ and $\mathbf{F}^A$, respectively, by referring to their original hidden speech representations $\mathbf{H}$. The formula for this process is as follows.

$$\mathbf{F}^{RC} = \mathcal{M}_{\text{Fusion}}(\mathbf{H}^C, \mathbf{H}^A, \mathbf{F}^C, \mathbf{F}^A). \tag{4}$$

The recombined features $\mathbf{F}^{RC}$ are the outputs of the CFDRN. Hidden speech representation is decomposed and selectively recombined to deal with the variabilities in dysarthric speech, similar to the human speech cognition.

C. Extractions of Slow and Rapid Components in Speech Features

The Fourier transform (FT) is the link between the time domain and frequency domain [62]. However, the result of FT is affected by the frame size. If a signal has a frame size of 20 ms, the frequencies below 50 Hz cannot be shown in the frequency domain because of its frequency resolution. Neural networks, which are good at modelling long-term signals, make up for this by considering information across the frames. Because the neural network has a recurrent mechanism, it can remember sequences over a wider span, and hence compensate for the limitations caused by the frame size of the FT. Therefore, we use FFT with a recurrent framework to extract slow-varying temporal features.

Specifically, in the slow-varying temporal processor, the pre-processed stable features $\mathbf{H}^C$ are mapped to a lower-dimensional space by down-projection. The lower-dimensional features are denoted as $\mathbf{L}^C$. The advantages of this step are that it removes redundant information from the features, saves storage space and reduces computations. Then, we apply a 1D FFT to the frames of $\mathbf{L}^C$. The outputs of the FFT are complex-values, including real numbers and imaginary numbers. In the frequency domain, we use two MLPs to transform the outputs of the FFT, and an inverse Fourier transform to transform the features into the time domain. To reduce the computation budget, the MLP is designed as a bottleneck framework, which first reduces and then increases the dimensionality of the features. Finally, the time-domain transformed features are projected onto a space with equal dimensions of $\mathbf{H}^C$. The features after up-projection are stable features $\mathbf{F}^C$.

We attempt to explain mathematically how our method extracts slow-varying temporal features given the limitation of the frame size. The process can be formulated as follows:

$$\mathcal{P}_{Slow}(\mathbf{H}^C) = \mathcal{F}^{-1}(\mathcal{G}(\mathcal{F}(\mathbf{H}^C))), \tag{5}$$

where $\mathcal{F}$ is the FFT function, and $\mathcal{F}^{-1}$ is the inverse FFT function. $\mathcal{G}$ is a nonlinear system, consisting of multiple stacked linear layers and activation layers. To simplify the calculation, we consider the case in which $\mathcal{G}$ consists of one linear layer and one activation layer, which can be formulated as:

$$\mathcal{G}(\mathbf{x}) = \mathcal{A}(\mathbf{W} * \mathbf{x} + \mathbf{b}), \tag{6}$$

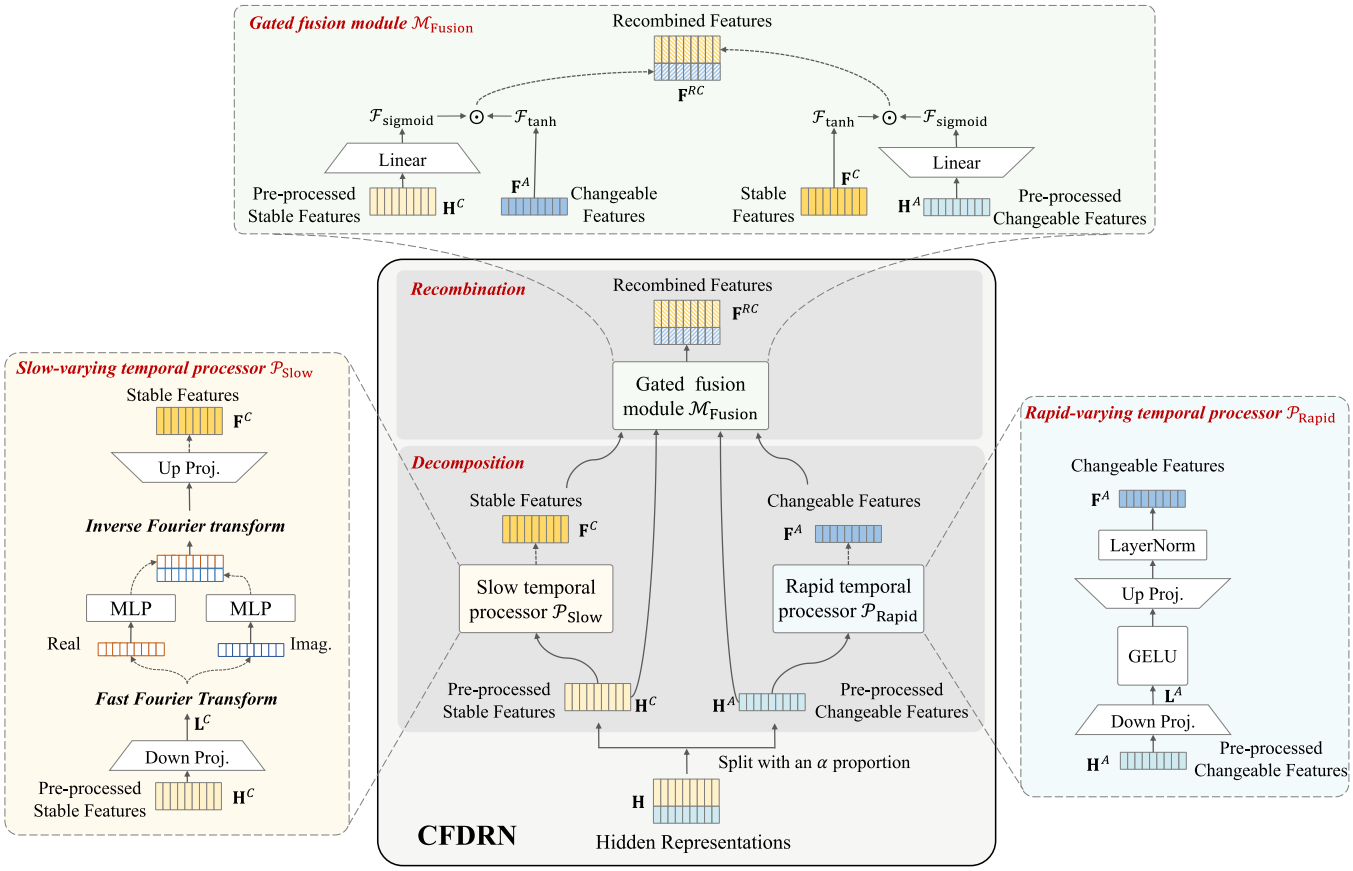

Fig. 2. Architecture of the proposed cognition-inspired feature decomposition and recombination network (CFDRN).

where $\mathcal{A}$ is the activation function. $\mathbf{W}$ is the weight applied to input vector $\mathbf{x}$ and $\mathbf{b}$ is the bias. In our case, $\mathcal{G}$ is a nonlinear system acting in the frequency domain, and hence $\mathbf{W}$ and $\mathbf{b}$ are complex-values.

The output $\mathcal{P}_{Slow}(\mathbf{H}^C)$ consists of slow-varying features when they satisfy the following inequality:

$$\sum_{i=1}^{N-1}\{[\mathcal{P}_{Slow}(\mathbf{h}_{i+1}^C)-\mathcal{P}_{Slow}(\mathbf{h}_i^C)]-(\mathbf{h}_{i+1}^C-\mathbf{h}_i^C)\}<0, \quad (7)$$

where $\mathbf{h}_i^C$ is the i-th frame of pre-processed $\mathbf{H}^C$. By combining (5–6) and inequality (7), the inverse Fourier transform of the bias $\mathbf{b}$ is the impulse function $\delta(t)$, and its variation is infinite. Therefore, if expression (7) is valid, it must satisfy $\mathbf{b}=\mathbf{0}$. Then, the inequality 5 can be written as follows:

$$\mathcal{P}_{Slow}(\mathbf{H}^C)=\mathcal{F}^{-1}(\mathcal{A}(\mathbf{W}(\mathcal{F}(\mathbf{H}^C)))). \quad (8)$$

For activation function $\mathcal{A}$, we use the ReLU function to apply a nonlinear transform. Therefore, activation function $\mathcal{A}$ serves as a filter that filters out unwanted frequency components. This process is controlled by the optimization of weight $\mathbf{W}$. When the elements of $\mathbf{W}(\mathcal{F}(\mathbf{H}^C))<=0$, the corresponding frequency component is removed. When the elements of $\mathbf{W}(\mathcal{F}(\mathbf{H}^C))>0$, (8) can be written as:

$$\mathcal{P}_{Slow}(\mathbf{H}^C)=\mathcal{F}^{-1}(\mathbf{W}(\mathcal{F}(\mathbf{H}^C))). \quad (9)$$

According to the linearity of the FT, the outputs $\mathcal{P}_{Slow}(\mathbf{H}^C)$ are slow-varying features when they satisfy the following inequality:

$$\sum_{i=1}^{N-1}\{(\mathbf{W}_{i+1}*\mathbf{h}_{i+1}^C-\mathbf{W}_i*\mathbf{h}_i^C)-(\mathbf{h}_{i+1}^C-\mathbf{h}_i^C)\}<0, \quad (10)$$

where $\mathbf{W}_i$ is the weight of the i-th frame. From the previous definition, we know that $\mathbf{W}=\mathrm{A}e^{j\boldsymbol{\theta}}$, where $\mathrm{A}=\|\mathbf{W}\|$, and $\boldsymbol{\theta}$ is the phase. Therefore, when A is less than 1, inequality (10) is valid.

If the nonlinear system applied on the frequency domain satisfies

$$\begin{cases}\mathbf{b} & =\mathbf{0},\\ \|\mathbf{W}\| & <1,\end{cases} \quad (11)$$

we can extract slow-varying temporal features from the system. The above mathematical proof is referenced from [62]. In our slow-varying temporal processor, we did not use bias vectors, and initialised the weights as very small values close to zero to ensure the processor has the ability to extract slow-varying temporal features. Then, the processor learns to perceive dysarthric speech during training.

In the rapid-varying temporal processor, we used an MLP to extract the details of the hidden speech representation. The pre-processed changeable features $\mathbf{H}^A$ are first down-projected

by a linear layer. After a nonlinear transformation, they are up-projected to the original dimensions, similar to the slow-varying temporal processor. We used GELU activation to perform the nonlinear transformation. The changeable features $\mathbf{F}^A$ are extracted after layer normalization.

D. Gate Fusion Module

The interaction between slow- and rapid-varying temporal features plays an important role in speech processing in the brain. This interaction is related to the modulation of the amplitude and phase of the two features [40]. Inspired by this, we propose the GFM, using two activation functions to constrain the extraction of slow- and rapid-varying temporal features. Specifically, we use a linear layer followed by the sigmoid function to generate a correlation coefficient from one set of features, and we use the tanh function applied to the other set of features to describe the positive and negative effects of the interaction. When stable features with correlation are extracted, the correlation coefficients are obtained from the pre-processed changeable features $\mathbf{H}^A$. The sigmoid function constrains the value of the correlation coefficient between 0 and 1, and the tanh function is applied to the pre-processed stable features $\mathbf{H}^C$. This usage of two activations is similar to that in [63], [64]. In brief, the obtained recombined stable features $\mathbf{R}^C$ and changeable features $\mathbf{R}^A$ are expressed as follows:

$$\mathbf{R}^C = \mathcal{F}_{\text{sigmoid}}(\mathbf{W}^{D^A \times D^C} * \mathbf{H}^A) \odot \mathcal{F}_{\tanh}(\mathbf{F}^C), \tag{12}$$

$$\mathbf{R}^A = \mathcal{F}_{\text{sigmoid}}(\mathbf{W}^{D^C \times D^A} * \mathbf{H}^C) \odot \mathcal{F}_{\tanh}(\mathbf{F}^A), \tag{13}$$

where $\mathbf{W}^{D^A \times D^C}$ and $\mathbf{W}^{D^C \times D^A}$ are learnable weights applied to the pre-processed features $\mathbf{H}^A$ and $\mathbf{H}^C$, respectively. $*$ denotes a convolution operator and $\odot$ denotes an element-wise multiplication operator. $\mathcal{F}_{\text{sigmoid}}$ is a sigmoid function, and $\mathcal{F}_{\tanh}$ is a tanh function. The recombined stable features obtained by multiplying the outputs of the two activations. The recombined changeable features are also obtained in a similar manner. Finally, the recombined features $\mathbf{F}^{RC}$ are obtained by concatenating the recombined stable and changeable features along the feature channels.

IV. Experimental Results and Analysis

A. Datasets

The proposed methods were validated using two commonly used dysarthria datasets: TORGO [65] and UASpeech .[1] Moreover, the unsupervised speech representation model was trained on the LibriSpeech ASR corpus [66], which contains 1,000 h of reading speech with a 16 kHz sampling rate.

The TORGO database consists of data from eight dysarthric speakers with varying degrees of CP or ALS and seven healthy control speakers. Speech files in the dataset were recorded using a microphone array and head-worn microphone with a 16 kHz sampling rate. The recordings lasted for 7 h, with 3 h of dysarthric speech and 4 h of normal control speech. There are 16,432 utterances in total, of which 4,171 are multiword utterances and 12,261 are single-word utterances. The UASpeech corpus comprises data from 15 dysarthric speakers with CP and 13 healthy control speakers. Speech files in the corpus were recorded using seven microphones at a 16 kHz sampling rate. The original recording contained long silent segments at the beginning and end of the audio recording. After cleaning the data following [55], the recordings lasted for 78.5 h, with 47.8 h of dysarthric speech and 30.7 h of normal control speech. There are 143,550 single-word utterances. Dysarthric speakers in these datasets were classified into four severity levels: severe, severe to moderate, moderate and mild.

The training, development, and test sets were divided according to the speaker, considering the severity levels of dysarthria. A speaker can only be included in one set. Therefore, the speakers in the three sets were not the same. Each set must contain speakers with various degrees of dysarthria. Specifically, for the TORGO database, the training set contained three speakers with S/M dysarthria (F01, M04, and M05) and all healthy control speakers. The development set included one speaker each with severe dysarthria (M01) and mild dysarthria (F04). The remaining two speakers with S/M dysarthria (M02, and F03) and one speaker with mild dysarthria (M03) were included in the test set. For the UASpeech corpus, the training set contained four speakers with S/M dysarthria (F03, M12, M07, and M05), two speakers with mild dysarthria (M09 and M14), and all healthy control speakers. The development set contained two speakers with S/M dysarthria (M01 and F02) and two speakers with mild dysarthria (M11 and M10). The test set contained three speakers with S/M dysarthria (M04, M16, and F04) and two speakers with mild dysarthria (M08, and F05).

B. Experimental Settings

All experiments were conducted using the open-source sequence modelling toolkit, Fairseq [67]. The proposed method was validated using three widely used unsupervised pre-training approaches: wav2vec 2.0, HuBERT, and data2vec. WavLM was not included in this study because we only focused on speech recognition tasks. When compared to wavLM, HuBERT has shown competitive performance in low-resource cases [60]. We chose the base models of the three approaches comprising approximately 95 M parameters as the baseline, which have been released on the official website.[2] The context network contained 12 transformer blocks with model dimensions of 768, inner dimensions (feedforward layer) of 3,072, and eight attention heads. All models were trained on a single NVIDIA V100 GPU. The batch size was set to 200 s. The Adam optimiser [68] was adopted with a warm-up learning rate similar to that in [24]. The learning rates were set to 1e-4 and 5e-5 for the TORGO and UASpeech datasets, respectively. For wav2vec 2.0 and data2vec, the models for the TORGO and UASpeech datasets were trained with 45 k and 90 k iterations, respectively. For HuBERT, the model for the TORGO and UASpeech datasets was trained with 60 k and 100 k iterations, respectively. A greedy search was

[1][Online]. Available: http://ifp-08.ifp.uiuc.edu/protected/UASPEECH

[2][Online]. Available: https://github.com/facebookresearch/fairseq

TABLE I
COMPARISON OF DYSARTHRIC SPEECH RECOGNITION PERFORMANCE (WORD ERROR RATE%) OF DIFFERENT APPROACHES ON THE TORGO AND UASPEECH DATASETS

ID	Pre-training	Method	Model Size (M)	TORGO			UASpeech		
				Dev	Test	Avg.	Dev	Test	Avg.
1	Wav2vec 2.0	Baseline	94.4	22.77	25.55	24.58	51.35	32.67	40.61
2	Wav2vec 2.0	PE-Adapter [28]	103.9	23.02	25.25	24.47	47.22	30.09	37.38
3	Wav2vec 2.0	RA [30]	104.7	21.64	23.15	22.62	51.91	33.16	41.14
4	Wav2vec 2.0	AdaptFormer [29]	103.9	22.51	24.72	23.96	52.52	32.96	41.28
5	Wav2vec 2.0	CFDRN (Ours)	106.4	**19.19**	**21.97**	**21.00**	**44.71**	**28.26**	**35.25**
6	HuBERT	Baseline	94.5	21.44	24.83	23.65	51.51	31.30	39.89
7	HuBERT	PE-Adapter [28]	104.0	19.75	23.96	22.50	51.36	31.76	40.09
8	HuBERT	RA [30]	104.8	19.36	23.47	22.04	52.83	33.39	41.66
9	HuBERT	AdaptFormer [29]	104.0	21.92	24.86	23.83	50.28	30.79	39.08
10	HuBERT	CFDRN (Ours)	106.5	**17.92**	**21.72**	**20.40**	**47.97**	**29.68**	**37.46**
11	Data2vec	Baseline	93.2	18.48	21.13	20.21	47.89	29.78	37.48
12	Data2vec	PE-Adapter [28]	102.7	17.47	21.09	19.83	48.45	28.51	36.99
13	Data2vec	RA [30]	103.4	19.41	19.71	19.61	48.43	29.16	37.35
14	Data2vec	AdaptFormer [29]	102.7	17.44	20.92	19.71	47.57	28.86	36.82
15	Data2vec	CFDRN (Ours)	105.2	**14.60**	**18.17**	**16.93**	**46.21**	**28.07**	**35.79**

Bold entities indicate that method achieved the best performance compared to methods with similar experimental Settings.

performed during the evaluation. A language model was not used in our experiments.

C. Results

We trained the ASR model separately on the two datasets. We evaluated the proposed CFDRN by comparing it with three mainstream unsupervised pre-training techniques: wav2vec 2.0, HuBERT, and data2vec, on the TORGO and UASpeech dysarthria datasets. Because of the high inter- and intraspeaker variabilities, each speaker has different impacts on ASR, even when they have the same degree of dysarthria. Analysing more speakers will help us discover the more general advantages and disadvantages of the proposed methods. Therefore, for a more objective comparison and analysis, we provided the word error rate (WER) for both the development and test sets. The baselines were pre-trained using the three unsupervised pre-training methods and directly fine-tuned to the dysarthric ASR task. We compared three popular adaptation approaches [28], [29], [30] to fine-tune the pre-training model by plugging a supplementary network into it. The principal difference between these approaches lies in the architecture of the supplementary networks. The parameter-efficient adapter (PE-Adapter) [28] adds linear layers after the self-attention layer and after each block of the encoder. The residual adapter (RA) [30] adds a bottleneck network after the feedforward layer. AdaptFormer [29] adds a scaled bottleneck network parallel to the feedforward layer. To achieve competitive performance on dysarthric ASR, the compared methods applied the adaptation by updating the entire model.

1) Primary Results of the CFDRN: Table I compares the dysarthric ASR performances (WER%) of the different approaches on the TORGO and UASpeech corpora. The results show that the data2vec model is more suitable for initializing dysarthric ASR than the wav2vec 2.0 and HuBERT models (comparing ID-1, ID-6, and ID-11). The three compared approaches were not beneficial for all datasets, possibly because of the variations in dysarthric speech across speakers. Specifically, the PE-Adapter pre-trained with HuBERT degraded the performance of ASR on the UASpeech corpus. RA pre-trained with either wav2vec 2.0 or HuBERT performed worse than the baseline on the UASpeech corpus. AdaptFormer pre-trained with wav2vec 2.0 and HuBERT was not efficient on the TORGO and UASpeech datasets, respectively. The results suggest the necessity for a dysarthric ASR to insert a well-designed network into the pre-training model, which was the focus of this study. When compared to the wav2vec 2.0-based baseline, the CFDRN achieved WER reductions (WERRs) of 14.56% and 13.20% on the TORGO and UASpeech datasets, respectively. When compared to the HuBERT-based baseline, the CFDRN achieved relative WERRs of 13.74% and 6.09% on the TORGO and UASpeech datasets, respectively. When compared to the data2vec-based baseline, the CFDRN achieved relative WERRs of 16.23% and 4.51% on the TORGO and UASpeech datasets, respectively. Evidently, the proposed CFDRN achieved significant improvements in dysarthric ASR when compared to the three pre-training approaches. The improvements were probably achieved because of the decomposition and recombination of features, which allows more effective feature extraction by the model.

2) Results on Different Degrees of Dysarthria: Fig. 3 shows the ASR performance for speakers with different degrees of dysarthria, that is, speakers with S/M dysarthria, and speakers with mild dysarthria. The results were obtained from the development sets and test sets. The different adaptation approaches that were used along with the three pre-training approaches were compared. For speakers with different degrees of dysarthria, the ASR performance with CFDRN was more stable than that with the compared methods. In the TORGO dataset, the improvements in the ASR for S/M dysarthric speakers were more obvious than those in the ASR for mild dysarthric speakers. In the UASpeech corpus, we observed that some of the compared approaches degraded the performance of ASR. As expected, CFDRN was effective in both S/M and mild dysarthria cases. Overall, the results for speakers with different degrees of dysarthria demonstrate that the proposed CFDRN is robust to significant variations in dysarthric speech.

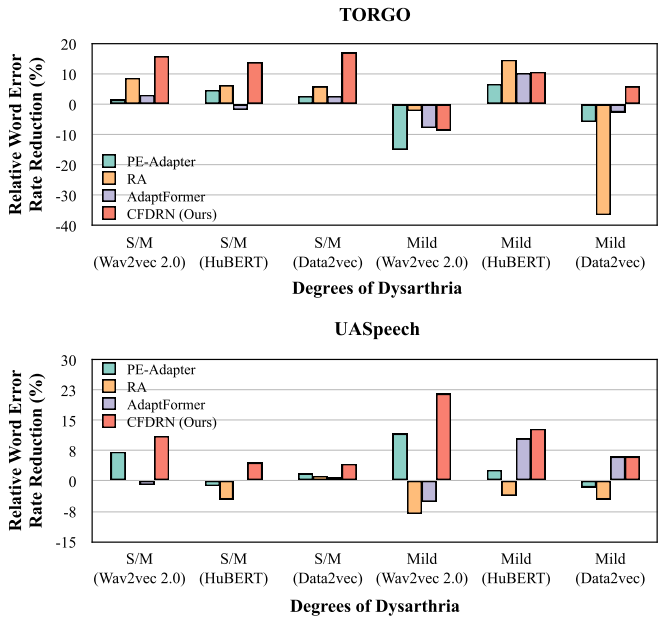

Fig. 3. ASR performance (word error rate(WER)%) for speakers with different degrees of dysarthria on the TORGO and UASpeech datasets. 'S/M' indicates speakers with severe or moderate dysarthria.

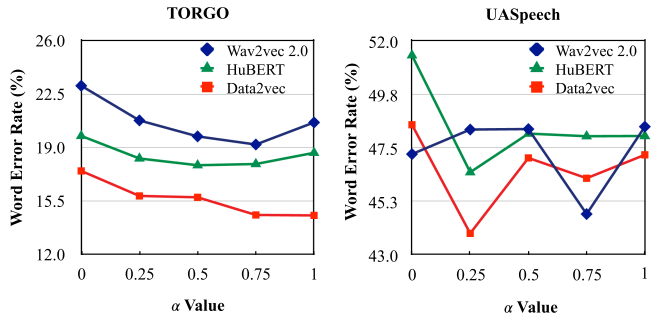

Fig. 4. Effect of different α values on dysarthric ASR performance. The results were obtained from the development sets of TORGO and UASpeech datasets.

D. Ablation Study

The proposed CFDRN comprises three key steps: a) the hidden speech representation is split into slow and rapid components in a specified proportion; b) a slow-varying temporal processor was designed that adopts FFT to extract stable features; and c) a GFM is proposed to recombine the stable and changeable features. Ablation studies were conducted to validate the effectiveness of each component of the proposed CFDRN.

1) Different α Values: Fig. 4 shows the ASR performance for different α values, where α is the proportion of the dimensions of slow and rapid components. $\alpha = 1$ implies that we only used the slow-varying temporal processor to adapt the model from normal to dysarthric speech. The results were obtained from the development sets of the TORGO and UASpeech datasets. We can observe that the CFDRN improves ASR performance with different α values. Dysarthric ASR achieves the best results in most cases when $\alpha = 0.75$. This suggests that the slow-varying temporal features dominate in ASR. The ASR performance marginally fluctuates when α is adjusted between zero and one. This validates the necessity of decomposing the features into stable and changeable features.

TABLE II
EFFECT OF FAST FOURIER TRANSFORM ON THE SLOW-VARYING TEMPORAL PROCESSOR

Pre-training	Network	TORGO				UASpeech			
		S/M	Mild	Test	Avg.	S/M	Mild	Test	Avg.
Wav2vec 2.0	Attention	35.33	4.20	24.86	24.03	62.93	15.11	30.95	39.73
Wav2vec 2.0	RNN	32.62	4.12	22.55	22.27	59.60	14.23	29.96	37.59
Wav2vec 2.0	FFT	30.65	4.09	**21.97**	**21.00**	56.80	12.39	**28.26**	**35.25**
HuBERT	Attention	34.52	4.52	25.35	23.62	61.93	13.02	29.98	38.20
HuBERT	RNN	32.46	3.98	23.84	22.12	60.74	13.18	29.45	37.66
HuBERT	FFT	29.78	3.96	**21.72**	**20.40**	59.95	13.60	**29.28**	**37.46**
Data2vec	Attention	26.14	3.69	19.43	17.98	59.42	13.44	29.02	37.11
Data2vec	RNN	24.93	3.88	18.86	17.28	58.07	12.27	28.26	35.85
Data2vec	FFT	24.60	3.47	**18.17**	**16.93**	58.15	12.06	**28.07**	**35.79**

Bold entities indicate that method achieved the best performance compared to methods with similar experimental Settings.

TABLE III
EFFECT OF THE GATED FUSION MODULE

Pre-training	GFM	TORGO				UASpeech			
		S/M	Mild	Test	Avg.	S/M	Mild	Test	Avg.
Wav2vec 2.0	w/o	31.40	4.01	22.71	21.45	60.83	13.80	29.88	38.01
Wav2vec 2.0	w/	30.65	4.09	**21.97**	**21.00**	56.80	12.39	**28.26**	**35.25**
HuBERT	w/o	31.53	4.12	23.02	21.57	60.38	13.23	30.44	37.50
HuBERT	w/	29.78	3.96	**21.72**	**20.40**	59.95	13.60	**29.68**	**37.46**
Data2vec	w/o	26.25	3.61	19.05	18.02	57.69	11.61	**27.63**	**35.33**
Data2vec	w/	26.03	3.61	**18.93**	**17.89**	58.15	12.06	28.01	35.79

Bold entities indicate that method achieved the best performance compared to methods with similar experimental Settings.

2) Slow-Varying Temporal Processor With the Fast Fourier Transform: Table II lists the effect of the FFT in the slow-varying temporal processor. The slow-varying temporal processor attempts to extract stable features, which change slowly. Therefore, the slow-varying temporal processor requires the ability to extract features over a long period. To validate the use of the FFT, we compared two widely used networks that satisfy our requirements: attention mechanism and recurrent neural network (RNN). We can observe that the FFT outperforms the other two methods in most cases, although the RNN achieves competitive performance when compared with the FFT for data2vec. The ASR for speakers with S/M dysarthria gained more help from the FFT than from the other approaches. The table shows that the RNN outperformed the attention mechanism on both datasets under the three unsupervised pre-training frameworks. Although the RNN exhibits a marginal advantage on some subsets, it requires more parameters, resulting in slower inference.

3) Gated Fusion Module: Table III lists the performance of the CFDRN-based ASR with and without the proposed GFM. The results show that the proposed GFM improved the results in almost all cases of high variability (S/M dysarthria) and achieved competitive performance in cases of lower variability (mild dysarthria). Thus, the GFM demonstrates the potential for

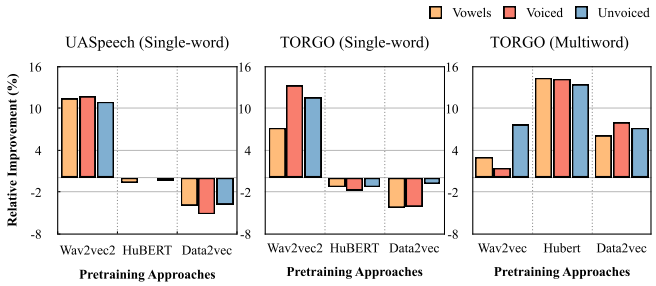

Fig. 5. Relative improvements (%) on vowels, voiced consonants, and unvoiced consonants when comparing the CFDRN-based ASRs with and without the GFM. The performance on single-word and multiword utterances are presented. The average performances of the development and test sets are shown.

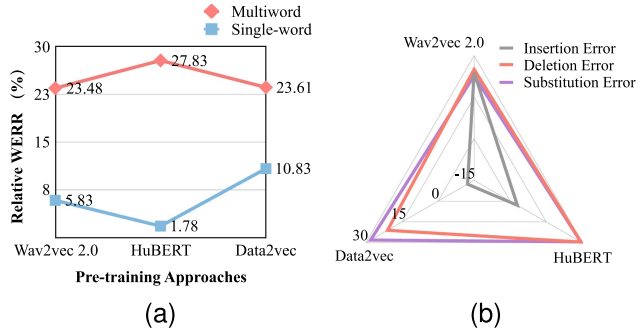

Fig. 6. Analysis diagram of the CFDRN-based dysarthric ASR, in cases of multiword and single-word recognition, when compared to the baseline. (a) Relative WER reduction for multiword and single-word utterances. (b) Relative correction (%) of three types of recognition errors in the ASR for multiword dysarthric speech.

application in the speaker-dependent ASR for low-intelligibility speech.

Fig. 5 shows the relative improvements on vowels, voiced consonants, and unvoiced consonants, comparing the CFDRN-based ASRs with and without the GFM. When recognizing single-word utterances, the improvements on the UASpeech and TORGO datasets are similar. The ASR performance gains from the GFM under the wav2vec 2.0 framework. However, the CFDRN without the GFM achieved marginally better performance than that with the GFM when using the HuBERT and data2vec pre-training approach. We can observe that when recognizing multiword utterances, the use of GFM improves the performance in all cases. From previous research [24], [25], [26], wav2vec2 is a generative learning techniques, which is good at extracting contextual information in speech, whereas HuBERT and data2vec are discriminative learning techniques, which are good at extracting discriminative features from speech clusters. Therefore, one possible reason the GFM decreased ASR performance on single-word utterances is that the HuBERT and data2vec are more tolerant of speech variability. Further feature manipulation leads to redundant information. When the amount of context information increases, the recognition performance can be improved by GFM.

E. Analysis

The subsequent section presents a further analysis of the CFDRN in dysarthric ASR focusing on speech recognition errors in specific categories of phonemes, thereby providing a basis for the further improvement of dysarthric speech recognition technology.

1) Speech Recognition Error Analysis: Fig. 6(a) shows the relative WERR of the CFDRN-based dysarthric ASR when compared to the baselines for multiword and single-word recognition. The improvement in the multiword ASR was obviously greater than that in the single-word ASR. This is probably because the single-word ASR lacks sufficient contextual information, making it challenging to identify distorted phonemes.

There are three types of recognition errors: insertion, deletion, and substitution errors. Most errors in single-word speech recognition errors are undoubtedly substitution errors. Therefore, for

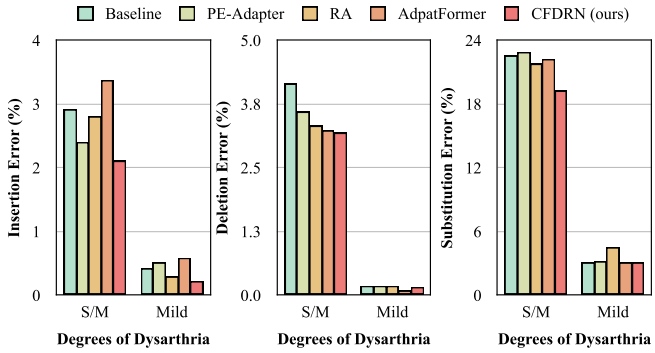

Fig. 7. Results (%) of three recognition errors in multiword dysarthric ASR under the wav2vec 2.0 pre-training framework. The results are the averages of the development and test sets.

an in-depth understanding of the CFDRN, we analysed the relative correction of the three types of errors in multiword dysarthric ASR, as shown in Fig. 6(b). The figure shows that deletion and substitution errors are significantly decreased using our CFDRN with three pre-training approaches, whereas insertion errors are increased when compared to the baseline. The results demonstrate that the CFDRN extracts phoneme features more effectively, including stable features that contain ambiguous phoneme features and changeable features that contain distinguishing details. Insertion errors in dysarthric speech are usually caused by imprecise articulation and improper breathing [69]. The increased insertion errors could have been caused by the excessively rich details of the changeable features. These details included mispronunciation and acoustic noise caused by the incoordination of respiratory and articulatory organs. That information was misrecognised as inserted phonemes by our system.

Fig. 7 shows the results of the three recognition errors in the multiword dysarthric ASR under the wav2vec 2.0 pre-training framework. The errors are related to the degrees of dysarthria. The three recognition errors in speakers with S/M dysarthria are significantly higher than those in speakers with mild dysarthria. Substitution errors occurred more frequently than other errors. All adaptation approaches contribute to reducing deletion errors,

TABLE IV
SEVEN COMMON HIGH-FREQUENCY RECOGNITION ERRORS OF COGNITION-INSPIRED FEATURE DECOMPOSITION AND RECOMBINATION NETWORK-BASED ASR

Rank	Recognition Error	Articulatory Error Type	Freq. (%)
1	/θ/ → /s/	Backing	1.41
2	/ʃ/ → /s/	Fronting	1.25
3	/m/ → /n/	Backing	1.16
4	/a/ → /ʌ/	Mid-vowel raising	0.94
5	/ɔɪ/ → /a/	Monophthongization	0.88
6	/ʊ/ → /ʌ/	Mid-vowel lowering	0.75
7	/ʧ/ → /t/	Fronting	0.71

"/x/ → /y/" indicates the ASR misrecognizes sound /x/ as sound /y/.

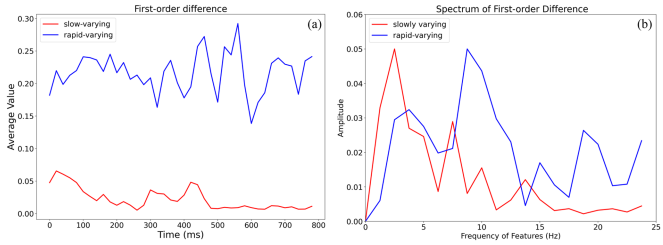

Fig. 8. Visualization of average values of the first-order differences and their spectrum of the extracted features. These features are the outputs of the slow- and rapid-varying processors. The processors were placed in the CFDRN, and the CFDRN was positioned after the first attention layer of the encoder layers. (a) First order difference. (b) Spectrum of first-order difference.

but not reducing other errors. For mild dysarthria, CFDRN achieved the lowest insertion and substitution error rates, and AdaptFormer achieved the lowest deletion error rate. For S/M dysarthria, the CFDRN achieved the lowest error rates for all three types of recognition errors. This indicates that the CFDRN has a stronger ability to extract variations in speech than the other approaches.

2) Misrecognised Phoneme Analysis: In addition to the significant variations in pronunciation of dysarthric speakers, previous research has shown that they also make highly consistent articulatory errors [70], which may lead to highly consistent recognition errors in the ASR. Table IV lists the seven common high-frequency speech recognition errors in the CFDRN-based ASR. The frequency was calculated according to the transcriptions and predictions were produced by the models pre-trained with wav2vec 2.0, HuBERT, or data2vec on both datasets. High-frequency speech recognition errors reflect the commontypes of articulatory errors in speakers with dysarthria. The articulatory error types were classified according to [11], [14], [71]. The table shows that the most confusing sounds for speech recognition are /θ/, /ʃ/, /m/, /a/, ɔɪ/, /ʊ/, and /ʊ/. Variations in dysarthric speech are closely associated with deviant pronunciation. Recognition errors for consonants are mainly of the backing or fronting error type, in which the articulatory place of the consonant is incorrect. Recognition errors for vowels primarily occur at the mid-vowel tongue height during monophthongisation. In particular, fricative consonants pose a challenge for dysarthric ASR. One possible reason is that dysarthric speech is usually accompanied by undesirable acoustic noise due to improper breathing, and the noise is similar to the fricative consonants that degrade ASR performance. Moreover, the recognition of mid-vowels is difficult in dysarthric ASR because the poor flexibility of the tongue of the speaker leads to deviations when pronouncing significantly high or low vowels. This study suggests that more attention should be focused on correcting these consonants and vowels to improve dysarthric ASR.

V. DISCUSSION

In previous sections, we introduced the proposed method and analysed the performance in dysarthric speech recognition. This section discusses the findings, strengths, and limitations of this study.

Dysarthric ASR is challenged by the high variability of disordered speech and limited amount of available dysarthric speech data. Human speech processing mechanisms provide excellent references to address these problems. Speech features are decomposed and processed separately using the left and right hemispheres of the brain [36], [37]. The robustness of human speech processing capabilities partially results from the suitable decomposition and organic recombination of features. Inspired by this, we proposed the CFDRN. For the CFDRN, we designed slow- and rapid-varying temporal processors for feature decomposition. We proposed a GFM for feature recombination. Previous experiments and analyses have validated the effectiveness of the proposed CFDRN for dysarthric ASR tasks. To further understand the CFDRN, we explored possible explanations for some of the results observed in our experiments.

A. Functions of Slow-Varying and Rapid-Varying Temporal Processors

The slow- and rapid-varying temporal processors were designed to decompose features into stable and changeable features. Stable features describe the common parts where speakers pronounce the same phonemes, which are stable features that play a decisive role in ASR. Changeable features supplement the ASR process, which are indecisive.

Fig. 8 shows the average values of the first-order difference and their spectrum in the extracted features. The first-order difference reflects the rate of change between two adjacent frames. A lower value indicates a slower change. We visualised the outputs of the processors in the first encoder layers. The figure shows that the output of the slow-varying temporal processor has a smaller first-order difference. This indicates that the slow-varying temporal processor tends to extract stable features, whereas the rapid-varying temporal processor tends to extract changeable features, similar to human speech processing. The spectrum of the first-order difference reflects the frequency component of the change between two adjacent frames. The dominant peak of the slow-varying component is about 2.5 Hz, and that of the rapid-varying component is about 8.25 Hz. These two components demonstrate distinguishable properties and showed similarities to speech processing in the brain. In summary, the visualization of the first-order difference shows

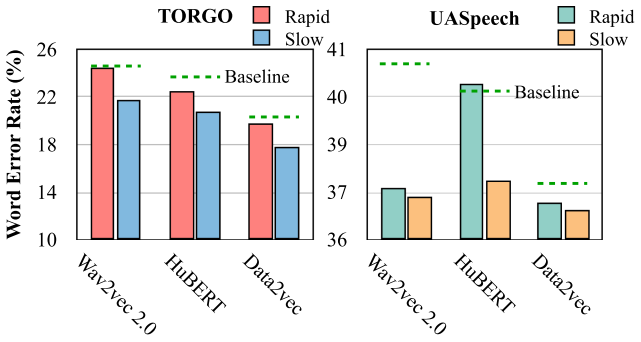

Fig. 9. Comparison of the ASR performance of the CFDRN-based ASR that uses either a rapid-varying temporal processor or a slow-varying temporal processor. The results represent the averages of the development and test sets for the TORGO and UASpeech datasets.

that the proposed CFDRN can simulate human speech processing to a certain extent.

Previous studies determined that the right side of the human brain plays a dominant role in phonological cognition, whereas the left side is a helper. Experimental demonstrations have shown only marginal degradation in the speech recognition accuracy in people who were paralysed on either the left or right side of the brain [72]. Fig. 9 shows the ASR performance of a CFDRN-based ASR that uses either the slow- or rapid-varying temporal processor. The ASR performance was better when only the slow-varying temporal processor was used than when only the rapid-varying temporal processor was used. Moreover, the performance improved even when only one processor was used. These findings are consistent with those on speech cognition processing in the human brain. The extracted stable features play a decisive role in the ASR, and the extracted changeable features supplement the ASR.

B. The Strengths and Limitations of the CFDRN

This article proposed the novel CFDRN, which applies human speech processing mechanisms to machine speech recognition. In contrast to previous studies on disordered ASR, the exploration in this study was from the perspective of feature decomposition and recombination to solve the high variability problem. According to the results and analysis in Sections IV-D and IV-E, the strengths of the CFDRN are as follows.

1) The CFDRN, which is inspired from the human speech cognition process, extracts detailed variations of speech more effectively and efficiently by first decomposing features into stable and changeable features, and then organically recombining them. It is robust to dysarthric speech with varying degrees of intelligibility.
2) The CFDRN achieves significant improvements in both dysarthric ASR of multiword and single-word utterances.
3) The CFDRN extracts more distinguishing features for ASR and significantly reduces deletion and substitution errors.

The limitation of the CFDRN is that it does not consider the effect of acoustic noise due to improper breathing, leading to increased insertion errors in certain cases as described in Section IV-E.

VI. Conclusion and Future Work

This study aimed to develop speaker-independent dysarthric speech recognition. To alleviate the data scarcity problem, we utilised unsupervised pre-training techniques to initialise the model and adapt it from normal to dysarthric speech. The proposed CFDRN could address variabilities in disordered speech. The CFDRN was added to each layer of the model for effective and efficient adaptation. The features in the CFDRN were decomposed into stable and changeable features, processed separately, and recombined selectively, similar to human speech cognition. We designed the slow- and rapid-varying temporal processors to handle the stable and changeable features. A GFM was developed to recombine these features. Experiments were conducted using the widely used TORGO and UASpeech dysarthria datasets. The effectiveness of the proposed method was validated using the wav2vec 2.0, HuBERT, and data2vec unsupervised pre-training techniques. The advantages and limitations of the proposed methods were analysed and discussed in detail. We determined that confusion in phoneme recognition often occurs between phonemes with close articulatory places because of the difficulty experienced by disordered speakers in accurately controlling them. According to the findings, in the future, this study suggests enhancing the method by adding a carefully designed language model for dysarthric speakers to avoid the effect of acoustic noise.

References

[1] B. Butterworth, "Speech production," in *Lexical Representation Process*. Cambridge, MA, USA: MIT Press, 1989, Art. no. 108.
[2] A. Asaei, M. Cernak, and H. Bourlard, "Perceptual information loss due to impaired speech production," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2433–2443, Dec. 2017.
[3] S. Liu et al., "Recent progress in the CUHK dysarthric speech recognition system," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2267–2281, 2021.
[4] T. L. Whitehill and V. Ciocca, "Speech errors in cantonese speaking adults with cerebral palsy," *Clin. linguistics Phonetic*, vol. 14, no. 2, pp. 111–130, 2000.
[5] T. Makkonen, H. Ruottinen, R. Puhto, M. Helminen, and J. Palmio, "Speech deterioration in amyotrophic lateral sclerosis (ALS) after manifestation of bulbar symptoms," *Int. J. Lang., Commun. Disord.*, vol. 53, no. 2, pp. 385–392, 2018.
[6] S. Scott and F. Caird, "Speech therapy for Parkinson's disease," *J. Neurol. Neurosurgery Psychiatry*, vol. 46, no. 2, pp. 140–144, 1983.
[7] R. D. Kent, "Research on speech motor control and its disorders: A review and prospective," *J. Commun. Disord.*, vol. 33, no. 5, pp. 391–428, 2000.
[8] E. Yílmaz, V. Mitra, C. Bartels, and H. Franco, "Articulatory features for ASR of pathological speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 2958–2962.
[9] R. D. Kent and Y.-J. Kim, "Toward an acoustic typology of motor speech disorders," *Clin. linguistics Phonetics*, vol. 17, no. 6, pp. 427–445, 2003.
[10] Z. Yue, E. Loweimi, H. Christensen, J. Barker, and Z. Cvetkovic, "Acoustic modelling from raw source and filter components for dysarthric speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2968–2980, 2022.
[11] Y. Lin, L. Wang, J. Dang, S. Li, and C. Ding, "End-to-end articulatory modeling for dysarthric articulatory attribute detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7349–7353.
[12] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2010, pp. 70–78.

[13] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian, and J. R. Duffy, "Acoustic studies of dysarthric speech: Methods, progress, and potential," *J. Commun. Disord.*, vol. 32, no. 3, pp. 141–186, 1999.

[14] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 4924–4927.

[15] J. R. Duffy, *Motor Speech Disorders E-Book: Substrates, Differential Diagnosis, and Management*. Amsterdam, The Netherlands: Elsevier Health Sci., 2019.

[16] H. P. Rowe, S. E. Gutz, M. F. Maffei, K. Tomanek, and J. R. Green, "Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective," *Front. Comput. Sci.*, vol. 4, 2022, Art. no. 770210.

[17] Y. Lin, L. Wang, S. Li, J. Dang, and C. Ding, "Staged knowledge distillation for end-to-end dysarthric speech recognition and speech attribute transcription," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 4791–4795.

[18] J. Deng et al., "Bayesian parametric and architectural domain adaptation of LF-MMI trained TDNNS for elderly and dysarthric speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4818–4822.

[19] F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7424–7428.

[20] R. Takashima, T. Takiguchi, and Y. Ariki, "Two-step acoustic model adaptation for dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6104–6108.

[21] Y. Sawa, R. Takashima, and T. Takiguchi, "Adaptation of a pronunciation dictionary for dysarthric speech recognition," in *Proc. IEEE Glob. Conf. Life Sci., Technol.*, 2022, pp. 631–635.

[22] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised automatic speech recognition: A review," *Speech Commun.*, vol. 139, pp. 76–91, 2022.

[23] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3465–3469.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Int. Conf. Adv. Neur. Inf. Process. Syst.*, 2020, pp. 12449–12460.

[25] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[26] A. Baevski, W. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 1298–1312.

[27] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[28] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.

[29] S. Chen et al., "Adaptformer: Adapting vision transformers for scalable visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 16664–16678.

[30] K. Tomanek, V. Zayats, D. Padfield, K. Vaillancourt, and F. Biadsy, "Residual adapters for parameter-efficient ASR adaptation to atypical and accented speech," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6751–6760.

[31] J. P. Rauschecker and S. K. Scott, "Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing," *Nature Neurosci.*, vol. 12, no. 6, pp. 718–724, 2009.

[32] J. R. Binder et al., "Human temporal lobe activation by speech and nonspeech sounds," *Cereb. Cortex*, vol. 10, no. 5, pp. 512–528, 2000.

[33] J. Obleser and C. Kayser, "Neural entrainment and attentional selection in the listening brain," *Trends Cogn. Sci.*, vol. 23, no. 11, pp. 913–926, 2019.

[34] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: Emerging computational principles and operations," *Nature Neurosci.*, vol. 15, no. 4, pp. 511–517, 2012.

[35] B. Zhao, J. Dang, and G. Zhang, "EEG source reconstruction evidence for the noun-verb neural dissociation along semantic dimensions," *Neuroscience*, vol. 359, pp. 183–195, 2017.

[36] E. Formisano, F. De Martino, M. Bonte, and R. Goebel, "'Who' is saying 'What'? Brain-based decoding of human voice and speech," *Science*, vol. 322, no. 5903, pp. 970–973, 2008.

[37] D. Poeppel, "The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'," *Speech Commun.*, vol. 41, no. 1, pp. 245–255, 2003.

[38] D. A. Abrams, T. Nicol, S. Zecker, and N. Kraus, "Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech," *J. Neurosci.*, vol. 28, no. 15, pp. 3958–3965, 2008.

[39] B. Berger et al., "Dynamic regulation of interregional cortical communication by slow brain oscillations during working memory," *Nature Commun.*, vol. 10, no. 1, 2019, Art. no. 4242.

[40] E. Başar, C. Başar-Eroğlu, S. Karakaş, and M. Schürmann, "Brain oscillations in perception and memory," *Int. J. Psychophysiology*, vol. 35, no. 2/3, pp. 95–124, 2000.

[41] B. F. Zaidi, S. A. Selouani, M. Boudraa, and M. Sidi Yakoub, "Deep neural network architectures for dysarthric speech analysis and recognition," *Neural Comput. Appl.*, vol. 33, no. 15, pp. 9089–9108, 2021.

[42] E. Hermann and M. Magimai-Doss, "Handling acoustic variation in dysarthric speech recognition systems through model combination," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4788–4792.

[43] X. Xie, R. Ruzi, X. Liu, and L. Wang, "Variational auto-encoder based variability encoding for dysarthric speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4808–4812.

[44] L. Wu, D. Zong, S. Sun, and J. Zhao, "A sequential contrastive learning framework for robust dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 7303–7307.

[45] L. Prananta, B. M. Halpern, S. Feng, and O. Scharenborg, "The effectiveness of time stretching for enhancing dysarthric speech for improved dysarthric speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 36–40.

[46] M. Geng et al., "Investigation of data augmentation techniques for disordered speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 696–700.

[47] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3586–3589.

[48] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 471–475.

[49] D. S. Park et al., "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2613–2617.

[50] Y. Takashima, T. Takiguchi, and Y. Ariki, "End-to-end dysarthric speech recognition using multiple databases," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6395–6399.

[51] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Synthesizing dysarthric speech using multi-speaker Tts for dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7382–7386.

[52] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5836–5840.

[53] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 6009–6013.

[54] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 4115–4119.

[55] F. Xiong, J. Barker, and H. Christensen, "Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition," in *Proc. IEEE 13th ITG-Symp. Speech Commun.*, 2018, pp. 1–5.

[56] Z. Yue, E. Loweimi, Z. Cvetkovic, H. Christensen, and J. Barker, "Multimodal acoustic-articulatory feature fusion for dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7372–7376.

[57] Z. Yue, E. Loweimi, and Z. Cvetkovic, "Raw source and filter modelling for dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7377–7381.

[58] R. Fan and A. Alwan, "DRAFT: A novel framework to reduce domain shifting in self-supervised learning and its application to children's ASR," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 4900–4904.

[59] R. Fan, Y. Zhu, J. Wang, and A. Alwan, "Towards better domain adaptation for self-supervised models: A case study of child ASR," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1242–1252, Oct. 2022.

[60] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.
[61] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5884–5888.
[62] A. V. Oppenheim, A. S. Willsky, S. H. Nawab, and J.-J. Ding, *Signals and Systems*, vol. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 1997.
[63] A. Van den Oord et al., "Conditional image generation with pixelcnn decoders," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4797–4805.
[64] A. V. D. Oord et al., "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Workshop Speech Synth. Workshop*, 2016.
[65] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, 2012.
[66] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.
[67] M. Ott et al., "fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Demonstrations*, 2019, pp. 48–53.
[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int Conf. Learn. Representations*, 2015, pp. 1–15.
[69] K. T. Mengistu and F. Rudzicz, "Comparing humans and automatic speech recognition systems in recognizing dysarthric speech," in *Proc. Can. Conf. Artif. Intell.*, 2011, pp. 291–300.
[70] R. Whurr, "Clinical management of dysarthric speakers," *J. Neurol., Neurosurgery, Psychiatry*, vol. 51, no. 11, 1988, Art. no. 1467.
[71] J. W. Lobel, "Vowel fronting, raising, and backing in Luzon and north-central Sulawesi," *WACANA, J. Humanities Indonesia*, vol. 22, no. 1, 2022, Art. no. 7.
[72] G. Hickok et al., "Bilateral capacity for speech sound processing in auditory comprehension: Evidence from wada procedures," *Brain Lang.*, vol. 107, no. 3, pp. 179–184, 2008.

Yuqin Lin received the bachelor's degree from Northeast Normal University, Changchun, China, in 2018. She is currently working toward the Ph.D. degree with the College of Intelligence and Computing, Tianjin University, Tianjin, China. Her research interests include speech recognition, speech production, and speech perception.

Longbiao Wang (Member, IEEE) received the Dr. Eng. degree from the Toyohashi University of Technology, Toyohashi, Japan, in 2008. From 2008 to 2012, he was an Assistant Professor with the faculty of engineering with Shizuoka University, Shizuoka, Japan. From 2012 to 2016, he was an Associate Professor with the Nagaoka University of Technology, Nagaoka, Japan. He is currently a Professor, the Director of the Tianjin Key Laboratory of Cognitive Computing and Application and Vice Dean of the School of Artificial Intelligence, Tianjin University, Tianjin, China. His research interests include robust speech recognition, speaker recognition, acoustic signal processing, and natural language processing.

Yanbing Yang received the bachelor's degree from the Hebei University of Technology, Tianjin, China, in 2019. She is currently working toward the master's degree with the College of Intelligence and Computing, Tianjin University, Tianjin. Her research focuses on speech recognition.

Jianwu Dang (Member, IEEE) received the graduation and M.S. degrees from Tsinghua University, Beijing, China, in 1982 and 1984, respectively, and the Ph.D. degree from Shizuoka University, Shizuoka, Japan, in 1992. From 1984 to 1988, he was a Lecture with Tianjin University, Tianjin, China. From 1992 to 2001, he was a Senior Researcher with ATR Human Information Processing Laboratories, Kyoto, Japan. From 1998, he joined the University of Waterloo, Waterloo, ON, Canada, as a visiting Scholar for one year. Since 2001, he has been a Professor with Japan Advanced Institute of Science and Technology, Nomi, Japan. From 2002 to 2003, he joined the Institute of Communication Parlee, Center of National Research Scientific, France, as a Research Scientist. Since 2009, he has been with Tianjin University. His research interests include speech science, brain science, and speech signal processing. He built MRI-based bio-physiological models for speech and swallowing, and endeavors to apply these models on clinics.