




WNSA-Net: An Axial-Attention-Based Network for Schizophrenia Detection Using Wideband and Narrowband Spectrograms

Ling He , Jia Fu, Yuanyuan Li, Xi Xiong , and Jing Zhang 

Abstract—Schizophrenia is a severe mental disease that affects patients' thoughts, feelings, and behaviors. Speech signal has proven to be a biomarker in the early diagnosis of schizophrenia. Previous studies on schizophrenic speech detection are mainly based on manual feature extraction engineering, which requires domain knowledge for researchers and has difficulties extracting effective features. This work proposes an end-to-end architecture, called Axial-attention-based Network using Wideband and Narrowband Spectrograms (WNSA-Net), to detect schizophrenia. Specifically, we adopt both wideband and narrowband spectrograms as inputs to represent speech signals using fine time and frequency structures. Then dilated convolution blocks are employed to capture detailed and long-range information in spectrograms. Axial-attention blocks are introduced to augment the information in feature maps along the time and frequency axes. In addition, we employ a gate mechanism to fuse the output feature maps from all channels. Experimental results on the Schizophrenia dataset and its subdatasets show that schizophrenic patients have difficulties in expressing emotions. To validate the performance of our WNSA-Net, experiments are conducted on Schizophrenia dataset and open-access TORGO database, achieving 97.37% and 98.16% accuracy in detecting schizophrenia and dysarthria, respectively. The results show promise for the proposed method in the diagnosis of disordered speech.

Index Terms—Axial-attention, dilated convolution, gated multi-channel fusion, schizophrenia, wideband and narrowband spectrograms.

I. INTRODUCTION

SCHIZOPHRENIA is a severe mental disorder with a lifetime prevalence of approximately 1% [1]. Schizophrenic

Manuscript received 10 September 2021; revised 3 June 2022 and 18 July 2022; accepted 8 September 2022. Date of publication 26 September 2022; date of current version 12 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 81901389 and in part by SCU-Yibin Project under Grant 2020CDYB-27. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hema A. Murthy. (Corresponding author: Jing Zhang.)

Ling He, Jia Fu, and Jing Zhang are with the College of Biomedical Engineering, Sichuan University, Chengdu 610065, China (e-mail: ling.he@scu.edu.cn; jia_fu@stu.scu.edu.cn; jing_zhang@scu.edu.cn).

Yuanyuan Li is with the Mental Health Center, West China Hospital of Sichuan University, Chengdu 610041, China (e-mail: lyy510@sina.com).

Xi Xiong is with the School of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225, China (e-mail: flyxiongxi@gmail.com).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Biomedical Research Ethics Committee of West China Hospital of Sichuan University under Application No. 2017-271.

Digital Object Identifier 10.1109/TASLP.2022.3209941

patients are generally characterized by disordered thinking, impaired speech, and disorganized behaviors. Clinical experience demonstrates that patients with schizophrenia require long-term treatment, and early diagnosis is crucial for preventing the onset of psychosis and reducing the severity of the disease [2]. In the clinic, schizophrenia is diagnosed by neurologists based on patients' retrospective recall biases [3] or observing speech/behaviors in clinical interviews. The result of a diagnosis relies on the experience of the clinician and is time-consuming. Thus, an automatic and objective detection method for schizophrenia is highly desirable.

Patients with schizophrenia have the typical symptoms of incoherent speech and diminished emotional expression in speaking [4]. Studies [3], [5], [6], [7] have proven that speech can be viewed as a biomarker for the early detection of schizophrenia. Current studies for automatic schizophrenia detection using speech are mainly based on feature engineering techniques. The extracted features can be summarized into four categories: fluency-, pitch-, spectrum-, and intensity-related features. a) *Fluency-related features*: Fluency-related features are commonly used to describe the incoherence of expression. Rapcan et al. [7] extracted pause-related features from emotionally neutral recordings to differentiate schizophrenic patients and healthy controls combined with linear discriminant analysis. Tahir et al. [8] and Gosztolya et al. [9] analyzed a set of conversational cues based on temporal analysis, such as the number and duration of pauses, speaking percentage and rate, and response time, to differentiate the two groups in speech production fluency. Iter et al. [10] proposed two coherence metrics and a computational model for referential incoherence to distinguish schizophrenic subjects and healthy controls. b) *Pitch-related features*: Pitch (F0) is closely related to the emotional information of speech signals. Studies [7], [11], [12], [13] have extracted the statistics (such as the mean and variance) of pitch values to distinguish schizophrenic groups from controls. The results demonstrate that patients with schizophrenia have less variability in pitch in the speaking process. c) *Spectrum-related features*: Spectrum-related features, such as formants, Mel-frequency cepstral coefficient (MFCC), and linear prediction coefficient (LPC), are commonly calculated to estimate vocal tract characteristics during speech production process [14], [15], [16], [17]. Chhabra et al. [18] and Compton et al. [12] demonstrated that patients with schizophrenia

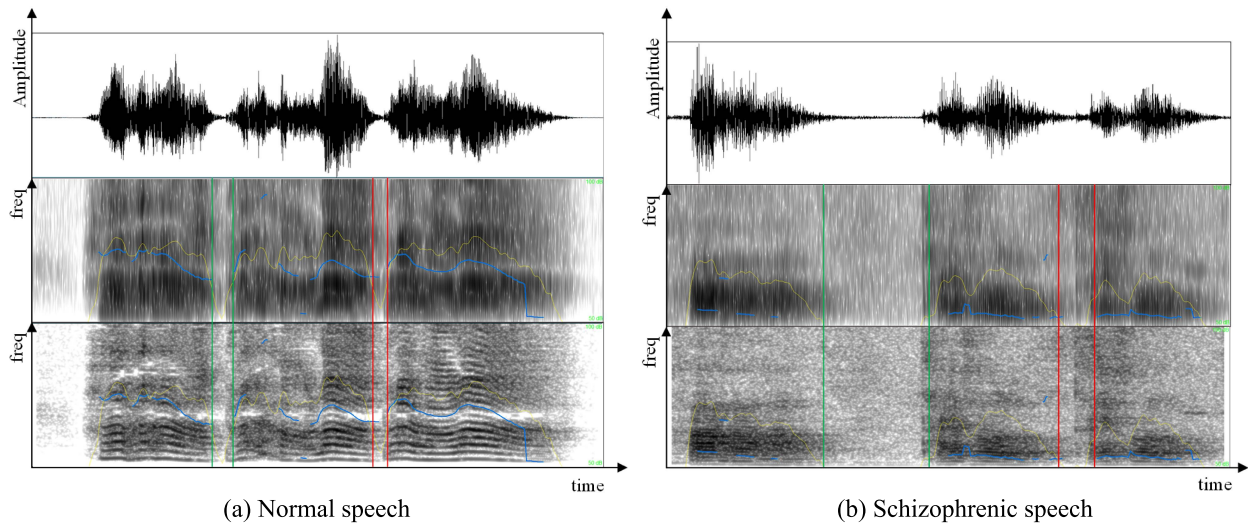


Fig. 1. Fluency-related features, pitch, intensity, and spectrograms for Mandarin utterance from healthy control and schizophrenic patient, which English translation is “Ha ha, it’s so great! It’s so great!”. The first to third rows in each sub-figure are the waveform, wideband spectrogram, and narrowband spectrogram of a speech signal, respectively.

have reduced formant dispersion, a smaller range of the second formant, and lower variations in energy in speech compared to healthy controls. d) *Intensity-related features*: Intensity is a crucial measure to capture emotion-related information in speech signals [7]. Studies [12], [19] have shown that patients with schizophrenia have reduced variability in the intensity of utterances. The four categories of features mentioned above are shown in Fig. 1 for utterances from schizophrenic patient and healthy subject. In each sub-figure, the length between two vertical lines with the same color means the duration of pause, the blue line means pitch contour, and the yellow line means intensity contour. It can be seen that there are significant differences between schizophrenic speech and normal speech. The features mentioned above can be employed to describe the characteristics in the speaking way of schizophrenic individuals and have much lower time complexity [20]. However, handcrafted feature extraction requires designers to have domain knowledge, and it is difficult to select effective features [21]. In addition, these handcrafted features are sensitive to the variability in data [22].

In recent years, many research efforts [23], [24], [25], [26], [27], [28], [29], [30] have been devoted to the pathological speech detection field using end-to-end neural networks. Pathological speech detection approaches generally comprise two essential parts: the input and the speech characteristics representation model. The input adopted in previous studies [23], [24], [25], [26], [27], [28] is usually the spectrogram or Mel-spectrogram that is calculated with a long window length, thus leading to good frequency resolution and poor time resolution of the input [31]. While both the time and frequency structures of speech signals are meaningful for detecting pathological speech. Concerning the representation of speech characteristics, previous works have proposed various methods for pathological speech detection, including convolutional neural networks (CNNs) [29], recurrent neural networks (RNNs) [27],

and residual-based models [30]. These methods treat the spectrogram or Mel-spectrogram as an image and consider rectangular kernels or deep network architectures to achieve the classification. The spatial distribution characteristics of spectrograms are neglected, as well as the meanings of time and frequency dimensions [23], [24], [32]. In addition, there are various shapes and scales of the target regions in the spectrograms or Mel-spectrograms caused by the divergent ways of speaking. These uncertainties may contribute to the limited performance of the networks [33], [34]. Thus, a novel deep learning architecture for schizophrenic speech detection is highly desirable.

This work proposes an end-to-end framework, termed WNSA-Net, for schizophrenia detection based on speech signals. First, we explore both wideband and narrowband spectrograms with high time and frequency resolution to describe the overall characteristics of speech signals. Second, we introduce a multi-scale feature representation (*MSFR*) module with dilated convolution blocks to capture detailed information and long-range contexture in spectrograms, thus avoiding the manual feature extraction procedure. Third, we utilize the axial-attention block in the fluency- and affect-related feature augmentation (*FAFA*) module, which can propagate the height- and width-axis information of feature maps. The information can be transmitted along the time and frequency axes, reflecting incoherent expression and flat affect of schizophrenic speech. Finally, the gate mechanism is adopted to fuse the feature maps from multiple channels using learnable weights in the gated multi-channel fusion (*GMC-fusion*) module, which performs the feature selection function. Our main contributions are summarized as follows:

- 1) We propose a novel end-to-end framework for detecting schizophrenia using speech signals, called WNSA-Net. This model explores both wideband and narrowband spectrograms with fine time and frequency structures, aiming to describe the prosodic and acoustic features of utterances.

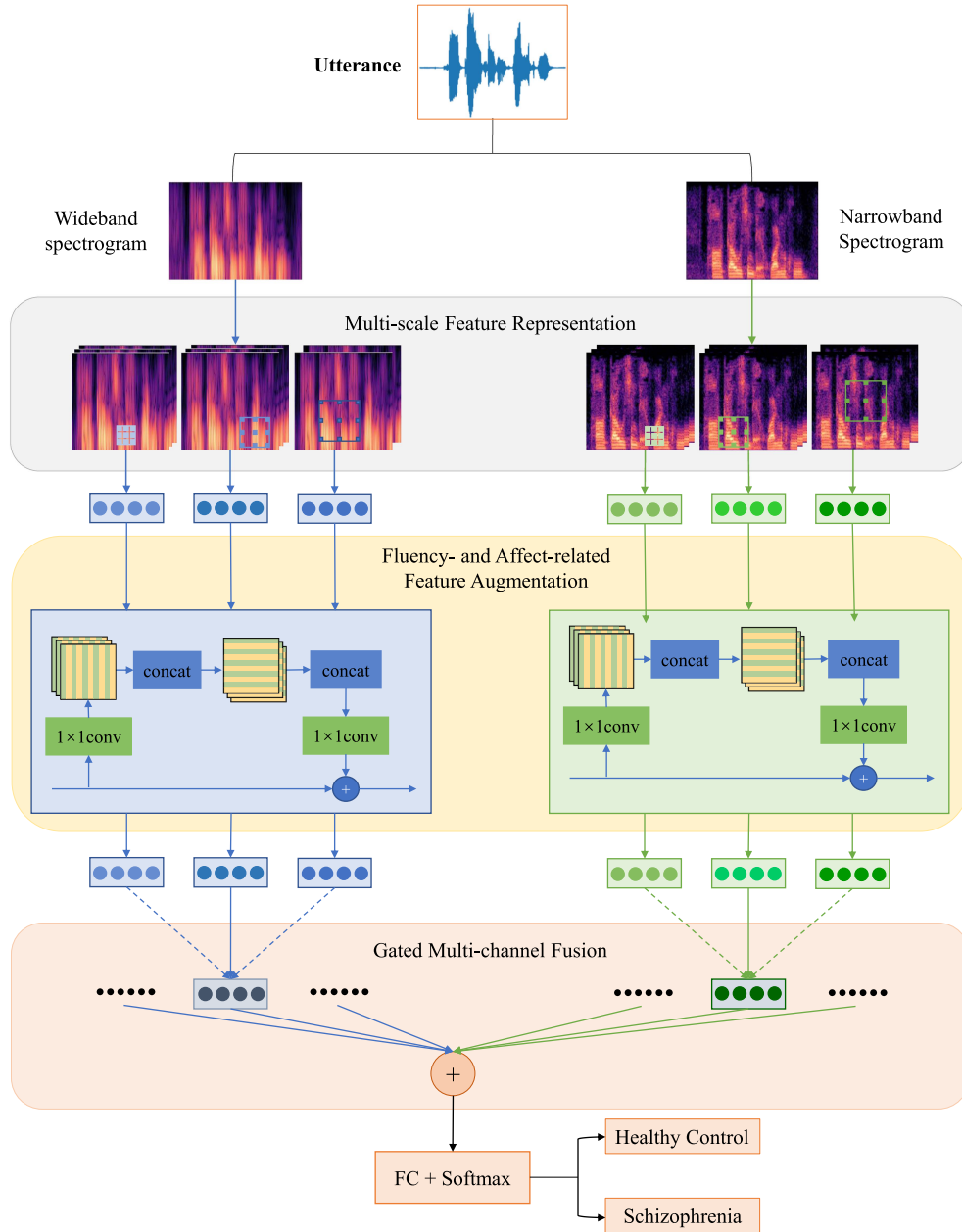


Fig. 2. The architecture of proposed WNSA-Net.

- 2) We introduce the dilated convolution into the MSFR module to extract the detailed information and long-range information of spectrograms.
- 3) To the best of our knowledge, this is the first attempt to employ the axial-attention architecture for speech classification tasks. The axial-attention block utilized in the FAFA module can propagate the information along the height- and width-axis, thus augmenting the time- and frequency-related features in spectrograms.
- 4) We evaluate WNSA-Net and its components on Schizophrenia dataset, and experimental results demonstrate the effectiveness of the proposed method. At the same time, our model is tested on the TORGO database [35] to validate the generalization.

The rest of this paper is organized as follows. Section II introduces the two branches and three components of the proposed neural network. Section III describes the dataset and experimental setup in detail. Section IV presents the experimental results on the schizophrenia dataset and open-access dysarthric speech database. Finally, the conclusion and limitations are presented in Section V.

II. METHODS

In this Section, we describe the proposed WNSA-Net in detail. Fig. 2 shows the architecture of WNSA-Net with two branches, wideband branch, and narrowband branch. In each branch, there

are three key components. 1) Multi-scale feature representation (*MSFR*) module: Dilated convolutions are employed to extract feature maps with different receptive fields, aiming to cover the information that reflects fluency-, pitch-, spectral-, and intensity-related characteristics in the spectrogram. 2) Fluency- and affect-related feature augmentation (*FAFA*) module: Axial-attention blocks are utilized to augment the time and frequency domain information in spectrograms and capture nonlocal features. 3) Gated multi-channel fusion (*GMC-fusion*) module: Feature maps from multiple channels are fused with a gate mechanism.

A. Data Representation

The spectrogram is a widely used tool to represent the spectral distribution of the signal in a spatial distribution [36]. In a spectrogram, time- and frequency-related features can be employed to express acoustic and prosodic features of utterances that reflect the characteristics in the pronunciation process. Concerning the length of window employed to generate a spectrogram, there are two categories of spectrograms: wideband spectrograms and narrowband spectrograms. 1) Wideband spectrogram: It is created using a window which is typically approximately 3 ms, resulting in good time resolution and poor frequency resolution [31], [37]. The wideband spectrogram is characterized by vertical striations, reflecting the dark horizontal bands that are formant frequencies [38]. The tracks of formants can be identified in the wideband spectrogram, which reflects the resonances of the vocal tract in the speaking process. 2) Narrowband spectrogram: This spectrogram is generated by a window which is about 20 ms [31], [37]. The length of the window is likely to be larger than the pitch, resulting in the effectiveness of capturing each harmonic separately in a spectrogram.

Schizophrenia is characterized by disordered thinking and impaired cognition. Patients with schizophrenia usually have a lack of interest in activity and trouble in emotional expression, which can be manifested as the small range of formants, little or no variation in the intensity, tone, and pitch of the speech in wideband and narrowband spectrograms [12], [18], [19]. In addition, incoherent or illogical thinking in schizophrenia leads to nonfluent verbal performance when they speak. Speech spoken by schizophrenic patients exhibits an increasing number and duration of pauses [8], [9], which can also be seen in the wideband and narrowband spectrograms. In this work, the two spectrograms mentioned above are adopted as the input of the proposed WNSA-Net. We calculate the spectrogram of each speech sample using the *librosa* package in Python. The wideband spectrogram $S_w \in R^{N \times f_{n_w}}$ is calculated with a 2 ms window and a 1 ms skip. The narrowband spectrogram $S_n \in R^{N \times f_{n_n}}$ is calculated with a 25 ms window and a 10 ms skip. Here, N is used to represent the number of points in short-time Fourier transform (STFT). f_{n_w} and f_{n_n} represent the number of frames for each speech signal using the window length of the wideband spectrogram and narrowband spectrogram, respectively. To encode the features in the spectrogram, the wideband spectrogram and narrowband spectrogram are converted into feature maps $X_w \in R^{N \times f_{n_w} \times C_0}$ and $X_n \in R^{N \times f_{n_n} \times C_0}$ using 1×1 convolution, in which C_0 represents the number of filters in the convolutional layer.

B. Multi-Scale Feature Representation (MSFR)

Patients with schizophrenia are reported to have structural brain changes, such as reduced gray matter density in the medial and lateral temporal lobes [39], [40], great variability of the thalamus and third ventricle volumes [41], [42], and disrupted connections between brain regions [43], [44]. These changes affect cognitive and memory abilities, leading to defects in thought, information, and speech processing. Patients usually have trouble concentrating and expressing emotion, manifested as a set of negative symptoms, such as disrupted thought, incoherent speech, and blunted affect. These symptoms can be reflected in the way they speak. Previous studies [7], [8], [9], [11], [12], [13], [18], [19] have demonstrated that fluency-, pitch-, spectrum-, and intensity-related features are effective in discriminating schizophrenic patients from healthy controls. Concerning the scale of information used for feature extraction, the features can be classified into two categories: detailed features and long-range features. Detailed features referring to the extraction are conducted within a small scale, such as the value of pitch period, the frequency of the first and second formants, and the values of intensity. Long-range features refer to the extraction being based on a large scale, such as the number and duration of pauses, the contours of pitch and formants, and the variation of intensity values. The two categories of features can reflect the speaking way from incoherence and affect perspectives, which are both essential for schizophrenic speech detection.

This work utilizes dilated convolution [45] to capture contextual information in speech signals. The dilated convolution can keep the output resolution without upsampling by expanding receptive fields [45]. To some extent, the contexture captured is dependent on the size of receptive fields. Small receptive fields are stimulated by fine details, and large receptive fields are stimulated by coarse details [46]. In the *MSFR* module, a dilated convolution operation is conducted on the feature maps that are converted from spectrograms. The dilated convolution blocks have different dilation ratios to capture information in multi-scale receptive fields, which can deliver the detailed information and long-range information of speech signals.

The *MSFR* module is composed of a convolutional block and three stages of 2D residual blocks. The convolutional block comprises a 2D convolutional layer followed by a batch normalization (BN) layer and rectified linear unit (ReLU) activation, aiming to accelerate the training and overcome the vanishing gradient problem. For each 2D residual block, a dilated convolution layer is used to substitute the conventional convolution layer. In the three stages, the numbers of 2D residual blocks are three, three, and two, respectively.

Given the feature maps $X_w \in R^{N \times f_{n_w} \times C_0}$ and $X_n \in R^{N \times f_{n_n} \times C_0}$, the outputs produced by the *MSFR* module are obtained by concatenating the outputs of the three stages on the wideband branch and narrowband branch, which can be formally expressed as:

$$F_w = \{F_{w1}^{MSFR}(X_w; \Theta); F_{w2}^{MSFR}(X_w; \Theta); F_{w3}^{MSFR}(X_w; \Theta)\}, \quad (1)$$

$$F_n = \{F_{n1}^{MSFR}(X_n; \Theta); F_{n2}^{MSFR}(X_n; \Theta); F_{n3}^{MSFR}(X_n; \Theta)\}, \quad (2)$$

where $F_w \in R^{N \times f_{n_w} \times (C_1 + C_2 + C_3)}$ and $F_n \in R^{N \times f_{n_n} \times (C_1 + C_2 + C_3)}$ represent the output of wideband and narrowband branch, respectively. F_{wl}^{MSFR} and F_{nl}^{MSFR} represents the *MSFR* module in the l th stage, Θ denotes the parameters of *MSFR* module, and C_l , $l = 1, 2, 3$ indicates the number of filters in the dilated convolutional layer in the l th stage.

C. Fluency- and Affect-Related Feature Augmentation (FAFA)

Incoherent expression and blunted affect are prominent symptoms of schizophrenia. Utterances spoken by schizophrenic patients have significant differences from those spoken by healthy individuals. The differences are mainly manifested in time and frequency domain features, such as the number and duration of pauses, the range of pitch values, and the frequency of formants. To highlight the differences in feature maps, this section aims to encode the information along the time and frequency axes sequentially.

This work utilizes the position-sensitive axial-attention block [47] to propagate the information in feature maps from the *MSFR* module. The axial-attention block can learn the correlation between the specific instant/frequency and the whole spectrogram, and it can exploit positional information of the feature map to capture the spatial features. Time and frequency domain features are activated with different weights, which are determined by the effectiveness of features for the schizophrenic speech detection task.

Specifically, the feature maps $F_w \in R^{N \times f_{n_w} \times (C_1 + C_2 + C_3)}$ and $F_n \in R^{N \times f_{n_n} \times (C_1 + C_2 + C_3)}$ produced by the *MSFR* module are fed into the axial-attention blocks. Taking $F_w \in R^{N \times f_{n_w} \times (C_1 + C_2 + C_3)}$ as an example, we first flatten the feature map F_w into a sequence along the height-axis. The output at position $o = (i, j)$, $y_{wo_h} \in R^{N \times f_{n_w} \times d_{out}}$, can be computed by projecting the input with the height-axis axial-attention layer as:

$$a_p = \text{softmax}_p \left(q_{wo}^T k_{wp} + q_{wo}^T r_{wp-o}^q + k_{wp}^T r_{wp-o}^k \right), \quad (3)$$

$$y_{wo_h} = \sum_{p \in R_{m(o) \times 1}} a_p (v_{wp} + r_{wp-o}^v), \quad (4)$$

where d_{out} denotes the number of output channels. $R_{m(o) \times 1}$ is the whole location lattice for the height-axis axial-attention layer. Queries $q_{wo} = W_{wQ} F_w$, keys $k_{wp} = W_{wK} F_w$, and values $v_{wo} = W_{wV} F_w$ are all linear projections of the input feature map F_w , $\forall o \in R$. W_{wQ} , W_{wK} , and W_{wV} are all learnable matrices. The learnable r_{wp-o}^q , r_{wp-o}^k , and r_{wp-o}^v are the positional encodings for queries, keys, and values. softmax_p represents a softmax function applied to all possible $p = (a, 1)$ positions, which in this case is also the whole 1D lattice.

Then, the feature map from the height-axis axial-attention layer is fed into the width-axis axial-attention layer, and the information is propagated along the width-axis. Finally, we obtain the feature maps F_{wA}^i , and F_{nA}^i , $i = 1, 2, 3$ by adding

the outputs of the width-axis axial-attention layer with the input feature maps F_w^i and F_n^i , respectively.

D. Gated Multi-Channel Fusion (GMC-Fusion)

To effectively combine the encoded feature, we employ the gated fusion technique [48] to fuse the output feature maps from three channels on the wideband branch and narrowband branch. The *GMC-fusion* module is similar to the gate mechanism in the long short-term memory (LSTM) block, which contains two categories of gates, memory gates m_w^i and m_n^i and reset gates r_w^i and r_n^i . Specifically, we obtain the fused feature representation f_w^i and f_n^i by:

$$G_w^i = (1 - m_w^i) \odot F_{wA}^i + \sum_{j \in \{1, 2, 3\} \setminus \{i\}} \alpha_w^j m_w^j \odot F_{wA}^j, \quad (5)$$

$$G_n^i = (1 - m_n^i) \odot F_{nA}^i + \sum_{j \in \{1, 2, 3\} \setminus \{i\}} \alpha_n^j m_n^j \odot F_{nA}^j, \quad (6)$$

$$F_{wo}^i = r_w^i \odot \tanh(G_w^i) + (1 - r_w^i) \odot F_{wA}^i, \quad (7)$$

$$F_{no}^i = r_n^i \odot \tanh(G_n^i) + (1 - r_n^i) \odot F_{nA}^i, \quad (8)$$

where \odot denotes the Hadamard product. α_w^j and α_n^j are learnable to adjust the relative ratio of the memory gate, which controls the information flow of features from the other levels j combined with the current level i .

To generate the final feature map for the classification task, feature maps F_{wo}^i and F_{no}^i from three levels on wideband branch and narrowband branch are aggregated. The convolutional layer followed by the sigmoid function is applied to obtain the classification results, which can be expressed as:

$$P = \sigma \left(\text{Conv} \left(\sum_{i=1}^3 F_{wo}^i + \sum_{i=1}^3 F_{no}^i \right) \right), \quad (9)$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\text{Conv}(\cdot)$ denotes the 3×3 convolutional operation. P represents the predicted probability of the healthy class. A binary cross-entropy loss function calculates the loss of speech sample m as follows:

$$\text{Loss} = Y(m) \log P(m) + (1 - Y(m)) \log (1 - P(m)). \quad (10)$$

where $Y(m)$ and $P(m)$ are the actual label and predicted label of the sample m , respectively.

III. EXPERIMENTAL DATABASES AND SETUP

A. Corpus Description

In this work, we applied the proposed method to the Schizophrenia dataset (also named SCZ dataset) to validate the effectiveness, which is recruited by the Psychiatry Department of the Mental Health Center, Sichuan University. The SCZ dataset comprise 34 first-episode drug-naïve patients with schizophrenia (15 males and 19 females) and 38 healthy controls (20 males and 18 females). The age of the schizophrenic group ranges from 18

to 60 (mean 39.2 ± 12.4), and the age of the control group ranges from 21 to 57 (mean 41.7 ± 10.6).

All participants are asked to read text with neutral, positive, and negative sentiments. The Chinese utterances and its IPA format for each emotional state are listed in Appendix A. All recordings were recorded in 16-bit format using SONY ICD-TX650, and the sampling frequency is 44.1 kHz. Specifically, SCZ dataset comprises 720 utterances (340 schizophrenic patients and 380 healthy controls) with a neutral sentiment, 569 utterances (271 schizophrenic patients and 298 healthy controls) with a positive sentiment (emotional state of happiness), and 216 utterances (102 schizophrenic patients and 114 healthy controls) with a negative sentiment (emotional state of anger).

To validate the effectiveness of our model in different emotions, we divide the SCZ dataset into three subdatasets, termed Subdataset I, Subdataset II, and Subdataset III, which comprise speech signals with neutral, positive, and negative sentiments of the SCZ dataset, respectively.

B. Implementation Details

Considering the importance of high frequency energy to detect disordered speech [49], [50], we utilize high sampling frequency of speech signals in the experiments. All recordings are framed and converted to wideband spectrograms and narrowband spectrograms using the STFT method, which is implemented using the librosa package in Python. The number of points in the STFT method is set as 512 in this study. To improve the invariance of the network to noise and ensure the consistency of the inputs, spectrograms are augmented by random cropping, random rotation, random rescaling, adding Gaussian noise, frequency masking, and time masking [51]. The inputs of WNSA-Net are all 256×256 .

In the training stage, we use the Adam [52] optimizer with a learning rate of 0.01 to optimize the parameters. The model is trained with a minibatch of 16 for 200 epochs. Dropout [53] with a rate of 0.1 is used to avoid overfitting problems. All experiments are implemented using the PyTorch framework [54]. Five-fold cross-validation is employed in the experiments to estimate the performance of WNSA-Net.

In the test stage, the patch size equals the training patch size. Augmentation to wideband and narrowband spectrograms is also utilized to improve the robustness of classification. To quantitatively evaluate the classification results, we calculate the mean accuracy, precision, recall, and F1-score, which can measure the performance on each class of data. The accuracy is the ratio of correctly predicted samples to total samples. The precision is the ratio of correctly predicted positive samples to total predicted positive samples. The recall refers to the probability of positive samples that are classified correctly. The F1-score is the harmonic mean of precision and recall, which can help balance the metric across positive and negative samples [55].

Table I shows the proposed WNSA-Net architecture details. In this architecture, a spectrogram is first embedded into a feature map using a convolutional layer with a filter size of 7×7 . In the MSFR module for each branch, there are three convolutional layers with a filter size of 3×3 and a dilation ratio of 1, 2, and

TABLE I
THE PROPOSED WNSA-NET ARCHITECTURE DETAILS

Layer	Dimension
Conv	7×7 (16 filters)
MSFR_conv1	3×3 (16 filters)
MSFR_conv2	5×5 (16 filters)
MSFR_conv3	7×7 (16 filters)
FAFA_conv1	1×1 (64 filters)
FAFA_conv2	1×1 (16 filters)
GMC-fusion_conv	3×3 (48 filters)
FC	$1 \times 1 \times 8$, $1 \times 1 \times 2$ (2 hidden layers)

TABLE II
THE CLASSIFICATION RESULTS ON FOUR SCHIZOPHRENIA DATASETS USING THE PROPOSED WNSA-NET

Dataset	Accuracy	Precision	Recall	F1-score
Subdataset I	0.9583	0.9503	0.9778	0.9627
Subdataset II	0.9643	0.9393	0.9870	0.9603
Subdataset III	0.9688	0.9375	1.0000	0.9667
SCZ dataset	0.9737	0.9499	0.9925	0.9695

3, respectively. In addition, the number of filters is all set as 16. In the FAFA module, there are four axial-attention layers. In the fully connected (FC) neural network, there are two hidden layers with 8 and 2 neurons, respectively. The output of this network is the probability of each class for the input sample.

IV. RESULTS AND DISCUSSION

This work proposes a deep learning architecture, termed WNSA-Net, to perform schizophrenia detection based on speech signal analysis. In this Section, we validate the performance of our WNSA-Net and verify the effectiveness of each key component in the proposed model using utterances of the Schizophrenia dataset with neutral, positive, and negative sentiments separately and mixed. Comparisons with existing works are also performed. Additionally, we validate the generalization of WNSA-Net using the TORGO database, which records speech signals of dysarthric patient.

A. The Classification Results of the Proposed WNSA-Net

To evaluate the performance of the proposed WNSA-Net, the model is tested using SCZ dataset and its three subdatasets in this subsection. In all experiments, the dataset is split into training and test sets, where the training set comprises utterances spoken by 56 individuals and the test set comprises utterances spoken by 16 individuals. Experimental results are shown in Table II.

Table II shows the accuracy, precision, recall, and F1-score of the proposed WNSA-Net. Experimental results demonstrate that the proposed WNSA-Net achieves accuracies of 0.9583–0.9737 and F1-scores of 0.9603–0.9695 on the schizophrenic speech classification task.

B. Effectiveness of Wideband and Narrowband Branch

To verify the importance of dual branches in the proposed WNSA-Net, we conduct controlled experiments to investigate

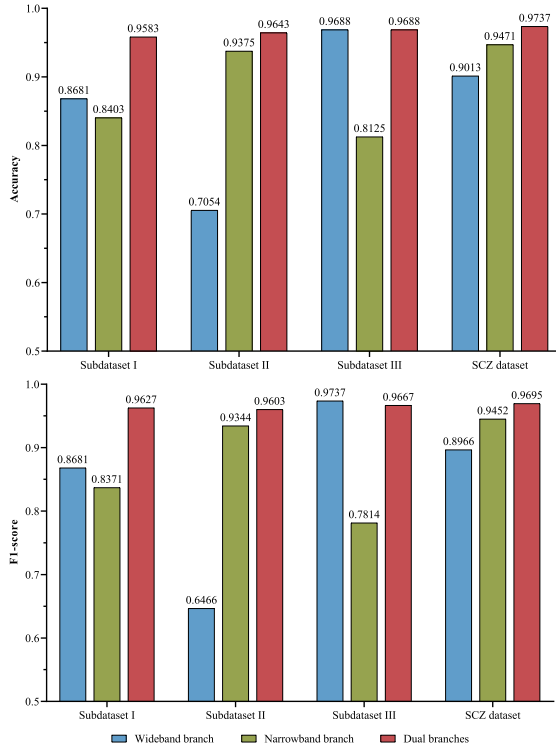


Fig. 3. The results of the proposed WNSA-Net using a single branch and dual branches as the input.

the performance among single branches and dual branches. Experimental results are performed in Fig. 3.

Wideband branch: The network is generated by removing the narrowband branch of WNSA-Net. Specifically, the wideband spectrogram is calculated by the FFT method with a 2 ms Hamming window. The dilated convolution block is used to extract multi-scale features of spectrograms. Then, the time and frequency domain information are augmented by axial-attention blocks. These features from three channels are fused with a gate mechanism.

Narrowband branch: The network is generated by removing the wideband branch of WNSA-Net. Specifically, the narrowband spectrogram is calculated by the FFT method with a 25 ms Hamming window. Then, the operations on the spectrograms are the same as the wideband branch.

Dual branches: It is the proposed WNSA-Net that has both narrowband branch and wideband branch.

The experimental results shown in Fig. 3 demonstrate that WNSA-Net with dual branches has robust performances in classifying speech spoken by patients with schizophrenia and healthy controls. WNSA-Net with dual branches shows 2.66%–7.24% on accuracy and 2.43%–7.29% on F1-score over WNSA-Net with a single branch in classifying schizophrenic patients and controls on the SCZ dataset.

The network with wideband spectrogram achieves good performances on Subdataset III. There are significant differences between schizophrenic speech and normal speech in temporal features. Utterances of anger emotional state from healthy controls show a faster speech rate [56], while schizophrenic

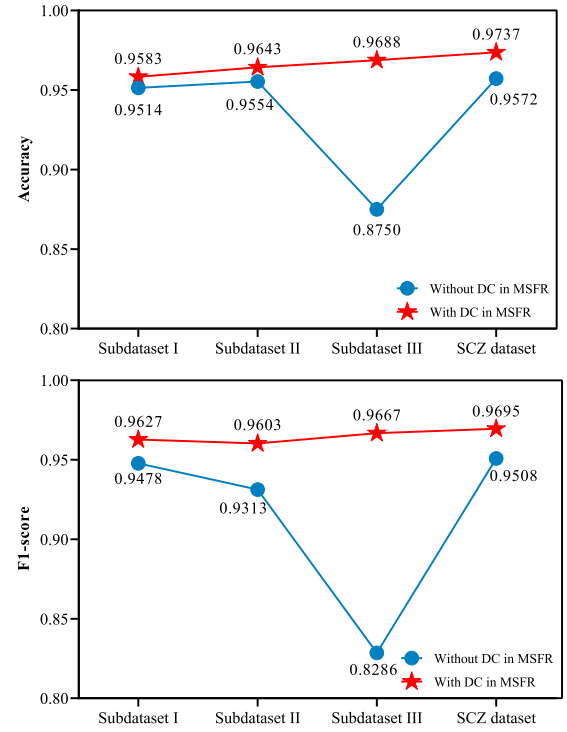


Fig. 4. Performance of MSFR module in the WNSA-Net.

patients with negative symptoms often have hesitant speech with long pauses [7]. Thus, the wideband spectrogram with fine time structures can perform good results in detecting schizophrenia.

The network with narrowband spectrogram achieves high classification accuracy and F1-score on Subdataset II, owing to its good frequency information representation ability. Schizophrenic patient with negative symptoms exhibits blunted vocal affect, which is manifested as the abnormal variability in fundamental frequency, formants, and intensity for the utterances of a happy emotional state [5]. The narrowband spectrogram can represent these characteristics well. **The network with narrowband spectrogram achieves 93.75% classification accuracy on Subdataset II.**

WNSA-Net with dual branches utilizes both wideband and narrowband spectrograms as the input, contributing to the comprehensive description of the characteristics of speech signals. The schizophrenic group is reported to have significant differences from the healthy group in fluency-, vocal- and glottal-related features [7], [8], [9], [11], [12], [13], [18], [19]. Thus, WNSA-Net with dual branches has better performances on the classification task.

C. Effectiveness of the Dilated Convolution in the MSFR Module

To verify the effectiveness of the dilated convolution (DC) in the MSFR module in our WNSA-Net, one controlled experiment is implemented in this subsection. Experimental results are shown in Fig. 4.

Without DC in the *MSFR* module: The network is generated by removing the DC in the *MSFR* module of WNSA-Net. Specifically, the wideband and narrowband spectrograms of recordings are converted to feature maps and fed into the *MSFR* module with conventional convolutions. Then, axial-attention blocks are employed to augment the information along the height- and width-axis. Feature maps from the wideband branch and narrowband branch are combined via the feature fusion module.

With DC in the *MSFR* module: It is the proposed WNSA-Net that utilizes dilated convolutions in the *MSFR* module.

As shown in Fig. 4, the use of DC in the *MSFR* module improves the classification accuracy by 1.65% and the F1-score by 1.87% on the SCZ dataset. The reason lies in that DC blocks in the *MSFR* module use different dilation ratios to aggregate multi-scale context in the feature maps. The *MSFR* module without DC is simulated by local features in a fixed scale receptive field, contributing to the limitations in multi-scale feature extraction. DC blocks in the *MSFR* module of each branch have different dilation ratios, which gain different scales of receptive fields. A DC block with a small dilation ratio can be utilized to extract detailed features, such as the formant frequency of vowels, pitch period, and energy values. The DC block with a large dilation ratio can capture long-range information, such as the variation of energy and intensity, the bandwidth of formants, and the range of pitch values. Studies [7], [12] show that schizophrenic speech and normal speech are divergent in detailed and long-range information, such as intensity, speaking rate, and the contour of vocal pitch. Especially for anger emotional states, owing to the significant differences in time- and frequency-domain [7], [56], the fusion of multi-scale features extracted from DC blocks can improve the results of schizophrenic speech detection.

D. Effectiveness of Axial-Attention Block

To verify the effectiveness of the axial-attention blocks in the *FAFA* module, one controlled experiment is conducted. The results are shown in Fig. 5.

Without the *FAFA* module: The network is generated by removing the *FAFA* module of WNSA-Net. Specifically, the feature maps from the *MSFR* module are fused with the gate mechanism directly and then fed into the FC layers for classification.

With the *FAFA* module: It is the proposed WNSA-Net that utilizes the *FAFA* module with axial-attention blocks.

The results shown in Fig. 5 denote that the network with the *FAFA* module shows 1.22%–3.52% on F1-score over that without the *FAFA* module on the schizophrenic speech detection task. The reason lies in the two advantages of the *FAFA* module. One is the ability to extract global features [47], [57]. The *FAFA* module can extract the context in the whole feature map using two axial-attention layers, which are helpful for the representation of schizophrenic speech characteristics. The other is that the *FAFA* module can enhance the information along the width- and height-axis [47], [57]. The axial-attention block in the *FAFA* module can propagate the information along the time and frequency axes. The feature map through the axial-attention block is weighted according to the contributions of classifying

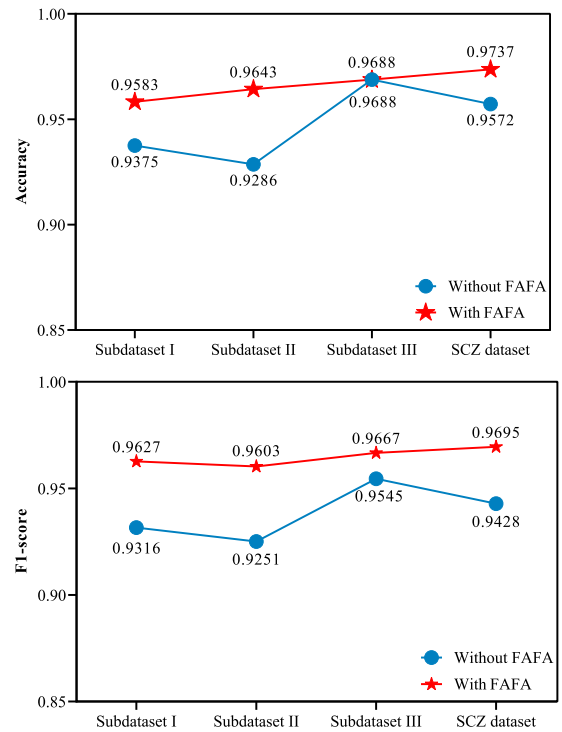


Fig. 5. Performance of *FAFA* module in the WNSA-Net.

schizophrenic speech and control. The improvement of the *FAFA* module on the Subdataset III is not very significant, which may be attributed to significant differences in time- and frequency-domain information between schizophrenic speech and normal speech. The feature augmentation operation is not essential for this situation. While for neutral and happy emotional states, pause-, pitch-, and formant-related features are important for detecting schizophrenia. The importance of the information in feature maps is not the same. Thus, the use of the axial-attention block improves the performance of classifying patients with schizophrenia and healthy controls.

E. Effectiveness of Gated Multi-Channel Fusion

To verify the effectiveness of the *GMC-fusion* module, two comparisons are implemented. The results are shown in Fig. 6.

Add: The network is generated by utilizing an add operation to substitute for the *GMC-fusion* module in WNSA-Net. Specifically, two representations are extracted from two branches via the *MSFR* module. Then, features are augmented via axial-attention blocks in the *FAFA* module. Finally, we simply add the outputs from the *FAFA* module together and put it into the FC layers.

Concatenate: Different from Add, the outputs from the *FAFA* module are combined via feature concentration.

With the *GMC-fusion* module: It is the proposed method that uses *GMC-fusion* module.

The experimental results shown in Fig. 6 demonstrate that our method with the *GMC-fusion* module achieves the best performance in all cases. The method with the *GMC-fusion* module shows an improvement of 1.95%–8.24% on F1-score over the method with Add. The fusion method using the add operation

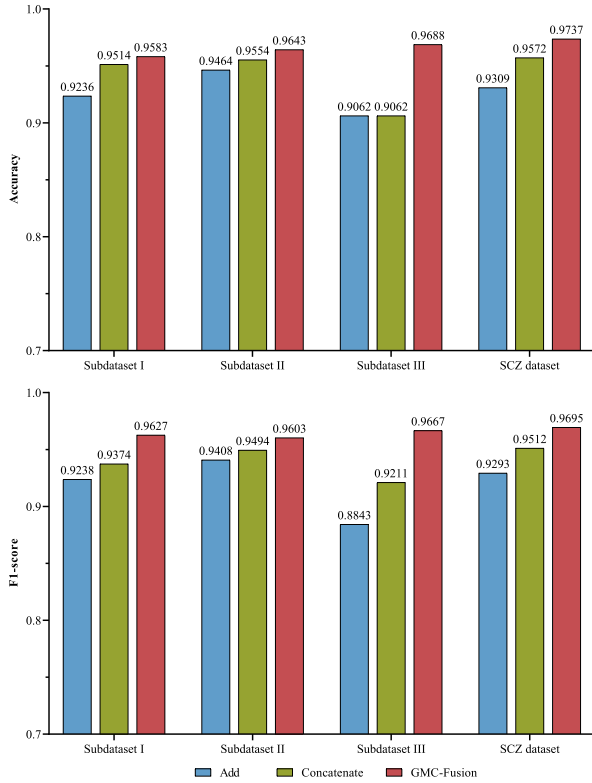


Fig. 6. Performance of *GMC-fusion* module in the WNSA-Net.

treats all feature maps from six paths equally, contributing to the fact that it cannot emphasize the effective features for the classification of schizophrenic speech and healthy controls. Meanwhile, the results demonstrate that the *GMC-fusion* method shows an improvement of 1.09%–4.56% on F1-score over the method with Concatenate. The fusion method using Concatenate operation enriches the channels of feature maps, but it retains abundant features and is unable to concentrate on the effective features for the classification. In the *GMC-fusion* module, the weights of six channels are learnable, ensuring that the features can be effectively selected. Thus, the *GMC-fusion* module outperforms the module with Add and Concatenate operations on the schizophrenic detection task.

F. Comparison With Existing Works

To comprehensively evaluate the performance of the proposed WNSA-Net, we implement four deep neural networks (Axialnet [47], AlexNet [58], BiLSTM [59], and SE-ResNet50 [60]) and our model on the SCZ dataset. Axialnet and SE-ResNet50 are both based on the attention mechanism, and AlexNet and BiLSTM are commonly used for classification and speech processing tasks. The classification results on the SCZ dataset using the four deep neural networks and our method are shown in Table III.

The experimental results shown in Table III demonstrate the effectiveness of our WNSA-Net. As shown in Table III, our method outperforms Axialnet by 7.53% on accuracy and 6.95% on F1-score, attributed to the additional dilated convolutional blocks can improve the detailed and long-range contextual extraction. The proposed WNSA-Net outperforms AlexNet

TABLE III
THE RESULTS OF CLASSIFYING SCHIZOPHRENIC PATIENTS AND HEALTHY CONTROLS ON THE SCZ DATASET USING FOUR DEEP NEURAL NETWORKS AND OUR PROPOSED WNSA-NET

Network	Accuracy	Precision	Recall	F1-score
Axialnet	0.8984	0.8324	0.9796	0.9000
AlexNet	0.8444	0.8451	0.8163	0.8304
BiLSTM	0.9492	0.9068	0.9932	0.9481
SE-ResNet50	0.9619	0.9412	0.9776	0.9600
WNSA-Net (ours)	0.9737	0.9499	0.9925	0.9695

and BiLSTM methods, the reason lies in that our proposed method augments the fluency- and affect-related features by adding axial-attention blocks. Compared with SE-ResNet50, our method has a slight improvement in classification accuracy, owing to that SE-ResNet50 utilizes short-connection and attention schemes to extract low-level and high-level features and emphasize the effective channels. The results in Table III highlight the importance of the representation, augmentation, and selection of the features related to the speaking way for the schizophrenic speech detection task.

G. Further Validation of the Proposed WNSA-Net Using the TORGO Database

To further verify the effectiveness and generalization of the model, the proposed WNSA-Net is tested on an open-access database, the TORGO database [35]. It is a corpus that is commonly used for detecting dysarthric speech spoken by patients with cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). The TORGO database comprises seven non-dysarthric speakers and eight dysarthric patients. There are three severity levels of dysarthria, containing three patients with very low level, two patients with low level, and three patients with medium level. The utterances in TORGO database are classified as three categories, non-words, words, and sentences.

Schizophrenic patients have abnormalities in the position and movement of tongue and jaw [11], leading to the typical characteristics of acoustic features, such as fluency- and affect-related features [7]. Similarly, CP and ALS affect the functions of the brain and nervous system. Patients with CP or ALS have disruptions in the neuro-motor interface, resulting in difficulty controlling vocal articulators. Studies [66], [67], [68], [69] have proven that the utterances spoken by CP and ALS patients have a limited range of pitch, longer percent pause times, and high noise-to-harmonic ratios compared with speech produced by healthy controls. In [61], [62], [63], [64], [65], time- and frequency-domain features are extracted to achieve the classification of dysarthric speech and non-dysarthric speech. Glottal parameters based on time- and frequency-domain, principal component analysis (PCA) [61], [62], [63] and acoustic features, such as the spectrogram, speech rate, voice quality, and rhythm [64], [65], are extracted. The glottal and acoustic features are combined with SVM classifiers and CNN- and LSTM-based neural networks to classify dysarthric patients and healthy controls. The classification results on the TORGO database in existing works [61], [62], [63], [64], [65] are listed in Table IV.

TABLE IV
THE RESULT OF ALS DETECTION ON THE TORGO DATABASE USING STATE-OF-THE-ART METHODS AND THE PROPOSED WNSA-NET

	Data Type	Features (Inputs)	Classifiers (Networks)	Accuracy	Precision	Recall	F1-score
[61]	Non-words, words, and sentences	Raw speech	CNN + MLP	0.7883	-	0.7624	-
			CNN + LSTM	0.7115	-	0.6617	-
		Glottal flow	CNN + MLP	0.8112	-	0.7526	-
			CNN + LSTM	0.7541	-	0.6968	-
[62]	Non-words Words Sentences	OpenSMILE-2 + glottal	SVM	0.9352	-	-	-
				0.9429	-	-	-
				0.9138	-	-	-
[63]	Sentences	OpenSMILE-1 OpenSMILE-2 Glottal-1 Glottal-2 OpenSMILE-1 + Glottal-1 OpenSMILE-1 + Glottal-2 OpenSMILE-2 + Glottal-1 OpenSMILE-2 + Glottal-2	SVM	0.8354	-	-	-
				0.8861	-	-	-
				0.7276	-	-	-
				0.7734	-	-	-
				0.8558	-	-	-
				0.8517	-	-	-
				0.8981	-	-	-
				0.8991	-	-	-
[64]	Words Sentences Words and sentences	Mel-Spectrogram	CNN-based	0.6800	-	-	-
				0.6200	-	-	-
				0.6600	-	-	-
[65]	Sentences	Speech rate + Pitch + Voice quality	RF	0.7890	0.7690	0.8330	0.8000
			SVM	0.8150	0.8650	0.7500	0.8040
			MLP	0.8230	0.8310	0.8500	0.8240
		Speech rate + Pitch + Voice quality + Rhythm	RF	0.8150	0.7880	0.8670	0.8250
			SVM	0.8230	0.8100	0.8500	0.8290
			MLP	0.8570	0.8640	0.8600	0.8570
Ours	Non-words, words, and sentences	Wideband spectrogram + Narrowband Spectrogram	WNSA-Net	0.9816	0.9926	0.9825	0.9872

OpenSMILE-1: RMS-energy, MFCCs (12), zero-crossing rate, pitch, voicing probability

OpenSMILE-2: log-energy, MFCCs (13), Mel-spectrum (26), pitch, jitter, shimmer, zero-crossing rate, voicing probability, spectral flux, roll-off points, spectral centroid, position of spectral maximum and minimum

Glottal-1: time- and frequency-domain glottal parameters

Glottal-2: PCA-based glottal parameters

To validate the performance of our WNSA-Net, we conduct experiments on the TORGO database to classify dysarthric speech and healthy controls. The classification results are given in Table IV. The experimental results in Table IV show that the proposed WNSA-Net achieves 98.16% classification accuracy on TORGO database, which outperforms state-of-the-art methods. The reason lies in that the pathological speech of the TORGO database has significant differences from the healthy speech in the glottal closure and the movements of speech organs, which can be manifested in time- and frequency-domain glottal parameters and acoustic features. WNSA-Net is designed by considering the relations between each component and speech characteristics. The *MSFR* module can extract multi-scale information that reflects the speaking way. The axial-attention blocks in the *FAFA* module can highlight the weights of time- and frequency-related features. The *GMC-fusion* module gives larger weights to the feature maps that are more helpful for the classification of dysarthric speech and healthy controls. Thus, WNSA-Net performs well in the dysarthric articulation detection task.

V. CONCLUSION

Speech evaluation is an important method in the early diagnosis of psychological and neurological diseases. This work proposes a novel end-to-end framework, termed WNSA-Net,

for schizophrenic speech and dysarthric articulation detection. This network comprises two branches (wideband branch and narrowband branch) and three key modules (*MSFR*, *FAFA*, and *GMC-fusion*). The importance and effectiveness of each branch and module are investigated concerning the correlation with speech characteristics. Experimental results on the Schizophrenia dataset and TORGO Database have demonstrated that the proposed WNSA-Net can provide effective and robust information for the early diagnosis of schizophrenia and dysarthric articulation. However, due to the limitation of data resources, we did not evaluate the performance on large-scale dataset, which is left to future work. Our method only achieves the classification of patients and healthy controls, which yields another future direction in the extension to assess the severity level of schizophrenic and dysarthric patients. Additionally, the proposed model needs intensive computational resources and a large memory footprint. Future work will seek to find a lightweight network that can be used on mobile devices to detect disordered speech.

APPENDIX A DATASET

In this work, utterances in three emotional states are used to evaluate the performance of WNSA-Net. The text in Mandarin and IPA format are given in Table A1.

TABLE A1
TEXT FOR SPEECH RECORDINGS IN MANDARIN, IPA, AND ENGLISH FORMAT

Sentiment	pinyin	IPA	English
Neutral	yuè jì, bèi chēng wéi huā zhōng huáng hòu, yòu chēng yuè yuè hóng. tā yī nián sì jì dū kě yǐ kāi huā, huā duǒ yī bān shì hóng sè huò fēn sè de, yě yǒu bái sè hé huáng sè de. hóng sè de yuè jì huā dài biǎo chún jié de ài, rén mēn duō bǎ tā zuò wéi ài qíng de xīn wù, ài de dài míng cí, shì qíng rén jié de shǒu xuān huā huì. yuè jì de huā hěn dà, chéng fā sǎn xīng, xiāng qì nóng yù. kě yǐ yòng lái guān shāng, yě kě yǐ yòng zuò yào cái hé shí cái. zhōng guó shì yuè jì de yuán chān dì zhī yī, wǒ guó yǐ jīng yǒu liǎng qiān yú nián de yuè jì huā zhōng zhī lì shǐ le.	yè tèi, pèi tʂʰɛŋ uéi xuā tʂʰɔŋ xuán xòu, iòu tʂʰɛŋ yè yè xóng. tʰā i tʰān sì tèi tū kʰɛ i tʰāi xuā, xuā tuǒ i tʰān sì xóng sè xuǒ fēn sè tʂ, iè iòu pái sè xé xuán sè tʂ. xóng sè tʂ yè tèi xuā tǎi piǎo tʂʰún tei tʂ ài, rén mēn tuǒ pǎ tʰā tsuò uéi ài tei tʂ tʂ cín uù, ài tʂ tǎi mǐn tʂʰí, qì tei tʂ rén tei tʂ sǒu yǎn xuā xui. yè tèi tʂ xuā xěn tà, tʂʰéŋ fā sǎn cǐn, cǐāŋ tei tʰi nòng iù. kʰɛ i tʰiòng lái kuān shāng, iè kʰɛ i tʰiòng tsuò iào tʂʰái xé qì tʂʰái. tʂʰɔŋ kuó qì yè tèi tʂ yán tʂʰǎn tì tǐ i tʂ, uò kuó i tʂ tʰiòng iòu liǎng tei tʰiòng iù nián tʂ yè tèi xuā tʂʰɔŋ tǐ lì sǐ lǚ.	Rose, is known as the queen of flowers, also known as the China rose. It can bloom all year round. The flowers are usually red or pink, but also white and yellow. The red rose represents pure love. People often take it as a token of love, a synonym of love, and it is the preferred flower for Valentine's day. Rose flowers are large, divergent, and fragrant. It can be used for viewing, as well as for medicinal materials and food materials. China is one of the places of origin of roses. China has a history of planting roses for more than 2000 years.
Negative	“pā” de yī shēng, duǒ duǒ bǎ mā mā zuì xǐ huān de wǎn dǎ suì le. mā mā tīng dào shēng yīn lián máng gǎn guò lái, kàn dào dì shàng de suǐ piàn, qì dé dèng yuán liǎo yǎn jīng, dà shēng hòu dào: “gēn nǐ shuō liǎo duǒ shǎo cǐ liǎo, bù xǔ wǎn wǒ de wǎn! kàn bā, wǎn bèi dǎ suì le! nǐ zhēn de shì yào qì sǐ wǒ!”	“pʰā” tʂ i tʂʰɛŋ, tuǒ tuǒ pǎ mā mā tsui tʂ xuān tʂ uǎn tǎ suì lǚ. mā mā tʰi tʰiòng tǎo sǐn lián máng kǎn kuò lái, kʰàn tǎo tǐ sǎn tʂ suǐ pʰiàn, tei tʂ té tèn yán liǎo iǎn tei tʂ, tǎ sǐn xǒu tǎo: “kēn nǐ guò liǎo tuǒ gǎo tʂʰi liǎo, pǔ cǔ uán uò tʂ uǎn! kʰàn pā, uǎn pèi tǎ suì lǚ! nǐ tʂʰɛn tʂ qì iào tei tʂ sǐ uò!”	With a “pa”, Duoduo broke her mother's favorite bowl. Mother heard the sound and rushed over. Seeing the debris on the ground, she was so angry that she widened her eyes and shouted, “how many times have I told you! Don't touch my bowl! See, the bowl was broken out! You really want to piss me off!”
Positive	kǎo shì juǎn zǐ fā xià lái le, fēi fēi kǎo liǎo jiù shí bā fēn, tā gāo xīng de huān hū: “hā hā, tài hào lái! tài hào lái!” shuō wán, fēi fēi kuài de pǎo huí jiā, yī jìn jiā mén jiù dà shēng de hǎn dào: “mā mā, mā mā, wǒ kǎo liǎo jiù shí bā fēn!” mā mā tīng liǎo, yǎn jīng xiào dé mǐ chéng liǎo yī tiáo fēng, bào zhù fēi fēi qīn liǎo yī kǒu, kāi xīn de shuō: “wǒ mēn fēi fēi kě zhēn bàng!”	kʰǎo qì tʂyàn tʂí fā xià lái lǚ, fēi fēi kʰǎo liǎo tiù qì pā fēn, tʰā kǎo cǐn tʂ xuān xū: “xā xā, tʰāi xǎo lái! tʰāi xǎo lái!” guō uán, fēi fēi kʰuài tʂ pʰǎo xuǐ tei tʂ, i tʂ tʂ tʂ mēn tei tʂ tǎ sǐn tʂ xǎn tǎo: “mā mā, mā mā, uò kʰǎo liǎo tiù qì pā fēn!” mā mā tʰi tʰiòng liǎo, iǎn tei tʂ xiào té mǐ tʂʰɛŋ liǎo i tʰiào fēng, pǎo tʂ fēi fēi tʂʰi liǎo i tʰiào, kʰāi cǐn tʂ guō: “uò mēn fēi fēi kʰɛ tʂʰɛn pàn!”	The examination paper was handed out, and Feifei got 98 points. He cheered happily, “Ha ha, it's so great! It's so great!” With that, Fei Fei ran home quickly. As soon as he entered the house, he shouted loudly, “Mom, mom, I got 98 points on the exam!” Mother listened, her eyes narrowed with laughter, hugged Feifei, kissed him, and said happily, “our Feifei are great!”

REFERENCES

- [1] A. K. Pagsberg, “Schizophrenia spectrum and other psychotic disorders,” *Eur. Child Adolesc. Psychiatry*, vol. 22, no. 1, pp. 3–9, 2013.
- [2] H. Häfner et al., “Early detection and secondary prevention of psychosis: Facts and visions,” *Eur. Arch. Psychiatry Clin. Neurosci.*, vol. 254, no. 2, pp. 117–128, 2004.
- [3] D. M. Low, K. H. Bentley, and S. S. Ghosh, “Automated assessment of psychiatric disorders using speech: A systematic review,” *Laryngoscope Invest. Otolaryngol.*, vol. 5, no. 1, pp. 96–116, 2020.
- [4] S. Mitra, T. Mahintamani, A. R. Kavoov, and S. H. Nizamie, “Negative symptoms in schizophrenia,” *Ind. Psychiatry J.*, vol. 25, no. 2, 2016, Art. no. 135.
- [5] A. S. Cohen, K. R. Mitchell, and B. Elvevåg, “What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments,” *Schizophrenia Res.*, vol. 159, no. 2sol:3, pp. 533–538, 2014.
- [6] A. Parola, A. Simonsen, V. Bliksted, and R. Fusaroli, “Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis,” *Schizophrenia Res.*, vol. 216, pp. 24–40, 2020.
- [7] V. Rapcan, S. D'Arcy, S. Yeap, N. Afzal, J. Thakore, and R. B. Reilly, “Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia,” *Med. Eng. Phys.*, vol. 32, no. 9, pp. 1074–1079, 2010.
- [8] Y. Tahir, D. Chakraborty, J. Dauwels, N. Thalmann, D. Thalmann, and J. Lee, “Non-verbal speech analysis of interviews with schizophrenic patients,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5810–5814.
- [9] G. Gosztolya, A. Bagi, S. Szalóki, I. Szendi, and I. Hoffmann, “Identifying schizophrenia based on temporal parameters in spontaneous speech,” *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3408–3412.
- [10] D. Iter, J. Yoon, and D. Jurafsky, “Automatic detection of incoherent speech for diagnosing schizophrenia,” in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol.: From Keyboard Clin.*, 2018, pp. 136–146.
- [11] F. Bernardini et al., “Associations of acoustically measured tongue/jaw movements and portion of time speaking with negative symptom severity in patients with schizophrenia in Italy and the United States,” *Psychiatry Res.*, vol. 239, pp. 253–258, 2016.
- [12] M. T. Compton et al., “The aprosody of schizophrenia: Computationally derived acoustic phonetic underpinnings of monotone speech,” *Schizophrenia Res.*, vol. 197, pp. 392–399, 2018.
- [13] R. Gold et al., “Auditory emotion recognition impairments in schizophrenia: Relationship to acoustic features and cognition,” *Amer. J. Psychiatry*, vol. 169, no. 4, pp. 424–432, 2012.
- [14] O. C. Ai, M. Hariharan, S. Yaacob, and L. S. Chee, “Classification of speech dysfluencies with MFCC and LPCC features,” *Expert Syst. Appl.*, vol. 39, no. 2, pp. 2157–2165, 2012.

- [15] J.-W. Lee, H.-G. Kang, J.-Y. Choi, and Y.-I. Son, "An investigation of vocal tract characteristics for acoustic discrimination of pathological voices," *Biomed. Res. Int.*, vol. 2013, 2013, Art. no. 758731.
- [16] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, Jan. 2006.
- [17] W. N. Chan, N. Zheng, and T. Lee, "Discrimination power of vocal source and vocal tract related features for speaker segmentation," *IEEE-ACM Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1884–1892, Aug. 2007.
- [18] S. Chhabra, J. C. Badcock, M. T. Maybery, and D. Leung, "Voice identity discrimination in schizophrenia," *Neuropsychologia*, vol. 50, no. 12, pp. 2730–2735, 2012.
- [19] D. Chakraborty et al., "Prediction of negative symptoms of schizophrenia from emotion related low-level speech signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 6024–6028.
- [20] T. Georgiou, Y. Liu, W. Chen, and M. Lew, "A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision," *Int. J. Multimedia Inf. Retrieval*, vol. 9, no. 3, pp. 135–170, 2020.
- [21] J. Zhang and Y. Wu, "A new method for automatic sleep stage classification," *IEEE Trans. Biomed. Circuits Syst.*, vol. 11, no. 5, pp. 1097–1110, Oct. 2017.
- [22] K. A. Nugroho, "A comparison of handcrafted and deep neural network feature extraction for classifying optical coherence tomography (OCT) images," in *Proc. IEEE 2nd Int. Conf. Inform. Comput. Sci.*, 2018, pp. 1–6.
- [23] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepaudioNet: An efficient deep model for audio based depression classification," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, 2016, pp. 35–42.
- [24] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy-Switz*, vol. 22, no. 6, 2020, Art. no. 688.
- [25] N. Trinh and O. Darragh, "Pathological speech classification using a convolutional neural network," in *Proc. Ir. Mach. Vis., Image Process.*, 2019, pp. 28–30.
- [26] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 446–450.
- [27] K. G. Dávid Sztahó and T. M. Gábor, "Deep learning solution for pathological voice detection using LSTM-based autoencoder hybrid with multi-task learning," *Biosignals*, vol. 4, pp. 135–141, 2021.
- [28] S. Souli, R. Amami, and S. B. Yahia, "A robust pathological voices recognition system based on dcnn and scattering transform," *Appl. Acoust.*, vol. 177, 2021, Art. no. 107854.
- [29] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41034–41041, 2018.
- [30] M. A. Mohammed et al., "Voice pathology detection and classification using convolutional neural network model," *Appl. Sci.-Basel*, vol. 10, no. 11, 2020, Art. no. 3723.
- [31] R. Reddy, V. Ramachandra, N. Kumar, and N. C. Singh, "Categorization of environmental sounds," *Biol. Cybern.*, vol. 100, no. 4, pp. 299–306, 2009.
- [32] L. Deng, O. Abdel-Hamid, and D. Yu, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6669–6673.
- [33] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE 4th Int. Conf. 3-D Vis.*, 2016, pp. 565–571.
- [34] G. Wang et al., "A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2653–2663, Aug. 2020.
- [35] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The toro database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, 2012.
- [36] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Noida, UP, India: Pearson Educ. India, 2006.
- [37] H. Chaurasiya, "Time-frequency representations: Spectrogram, cochleogram and correlogram," *Procedia Comput. Sci.*, vol. 167, pp. 1901–1910, 2020.
- [38] S. Cheung and J. S. Lim, "Combined multi-resolution (wide-band/narrowband) spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1991, pp. 457–460.
- [39] S. M. Lawrie, A. M. McIntosh, J. Hall, D. G. Owens, and E. C. Johnstone, "Brain structure and function changes during the development of schizophrenia: The evidence from studies of subjects at increased genetic risk," *Schizophrenia Bull.*, vol. 34, no. 2, pp. 330–340, 2008.
- [40] L. E. DeLisi, K. U. Szulc, H. C. Bertisch, M. Majcher, and K. Brown, "Understanding structural brain changes in schizophrenia," *Dialogues Clin. Neurosci.*, vol. 8, no. 1, p. 71, 2006.
- [41] S. P. Brugger and O. D. Howes, "Heterogeneity and homogeneity of regional brain structure in schizophrenia: A meta-analysis," *JAMA Psychiatry*, vol. 74, no. 11, pp. 1104–1111, 2017.
- [42] E. Antonova, T. Sharma, R. Morris, and V. Kumari, "The relationship between brain structure and neurocognition in schizophrenia: A selective review," *Schizophrenia Res.*, vol. 70, no. 2–3, pp. 117–145, 2004.
- [43] M. E. Shenton, C. C. Dickey, M. Frumin, and R. W. McCarley, "A review of MRI findings in schizophrenia," *Schizophrenia Res.*, vol. 49, no. 1–2, pp. 1–52, 2001.
- [44] M. Kubicki, R. W. McCarley, and M. E. Shenton, "Evidence for white matter abnormalities in schizophrenia," *Curr. Opin. Psychiatry*, vol. 18, no. 2, pp. 121–134, 2005.
- [45] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–13.
- [46] O. F. Lazareva, T. Shimizu, and E. A. Wasserman, *How Animals See the World: Comparative Behavior, Biology, and Evolution of Vision*. London, U.K.: Oxford Univ. Press, 2012.
- [47] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 108–126.
- [48] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10502–10511.
- [49] N. V. Naranjo, E. M. Lara, I. M. Rodríguez, and G. C. García, "High-frequency components of normal and dysphonic voices," *J. Voice*, vol. 8, no. 2, pp. 157–162, 1994.
- [50] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Front. Psychol.*, vol. 5, 2014, Art. no. 587.
- [51] 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.
- [52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.
- [53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [54] A. Paszke et al., "Automatic differentiation in pytorch," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–4.
- [55] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and ROC: A family of discriminant measures for performance evaluation," in *Proc. Australas. Joint Conf. Artif. Intell.*, 2006, pp. 1015–1021.
- [56] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech - A review," in *Toward Robotic Socially Believable Behaving Systems - Volume I - Modeling Emotions (Intelligent Systems Reference Library)*, A. Esposito and L. C. Jain, Eds. Berlin, Germany: Springer, 2016, vol. 105, pp. 205–238.
- [57] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–11.
- [58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, pp. 1097–1105.
- [59] S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional long short-term memory networks for relation classification," in *Proc. 29th Pacific Asia Conf. Lang., Inf., Comput.*, 2015, pp. 73–78.
- [60] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [61] N. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67745–67755, 2020.
- [62] N. Narendra and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3403–3407.
- [63] N. Narendra and P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Commun.*, vol. 110, pp. 47–55, 2019.

- [64] S. Prakash, "Deep learning-based detection of dysarthric speech disability," May 10, 2020. [Online]. Available: <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1204/reports/custom/report42.pdf>
- [65] A. Hernandez, E. J. Yeo, S. Kim, and M. Chung, "Dysarthria detection and severity assessment using rhythm-based metrics," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 25–29.
- [66] J. R. Green et al., "Bulbar and speech motor assessment in ALS: Challenges and future directions," *Amyotrophic Lateral Sclerosis, Frontotemporal Degeneration*, vol. 14, no. 7/8, pp. 494–500, 2013.
- [67] J. Tomik, B. Tomik, and M. Wiatr, "The evaluation of abnormal voice qualities in patients with amyotrophic lateral sclerosis," *Neurodegenerative Dis.*, vol. 15, no. 4, pp. 225–232, 2015.
- [68] K. M. Allison, Y. Yunusova, T. F. Campbell, J. Wang, J. D. Berry, and J. R. Green, "The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to ALS," *Amyotrophic Lateral Scler., Frontotemporal Degeneration*, vol. 18, no. 5/6, pp. 358–366, 2017.
- [69] R. Delorey, H. Leeper, and A. Hudson, "Measures of velopharyngeal functioning in subgroups of individuals with amyotrophic lateral sclerosis," *J. Med. Speech-Lang. Pathol.*, vol. 7, no. 1, pp. 19–31, 1999.