An Efficient Deep Learning Based Method for Speech Assessment of Mandarin-Speaking Aphasic Patients

Seedahmed S. Mahmoud, *Senior Member, IEEE*, Akshay Kumar, Yiting Tang, Youcun Li, Xudong Gu, Jianming Fu, and Qiang Fang

***Abstract*—Speech assessment is an important part of the rehabilitation process for patients with aphasia (PWA). Mandarin speech lucidity features such as articulation, fluency, and tone influence the meaning of the spoken utterance and overall speech clarity. Automatic assessment of these features is important for an efficient assessment of the aphasic speech. Hence, in this paper, a standardized automatic speech lucidity assessment method for Mandarin-speaking aphasic patients using a machine learning based technique is presented. The proposed assessment method adopts the Chinese Rehabilitation Research Center Aphasia Examination (CRRCAE) standard as a guideline. Quadrature based high-resolution time-frequency images with a convolutional neural network (CNN) are utilized to develop a method that can map the relationship between the severity level of aphasic patients' speech and the three speech lucidity features. The results show a linear relationship with statistically significant correlations between the normalized true-class output activations (TCOA) of the CNN model and patients' articulation, fluency, and tone scores, i.e., 0.71 ($p < 0.001$), 0.60 ($p < 0.001$) and 0.58 ($p < 0.001$), respectively. The linearity of the proposed Mandarin aphasic speech assessment method and its significant correlation with the speech severity levels show the efficacy of the method in predicting the severity of impaired Mandarin speech. The outcome of this research envisages assisting speech-language pathologists in Mandarin-speech impairment assessment and promoting early support discharge; hence could alleviate the stress that the healthcare system is currently experiencing in China nationwide. The framework of the proposed Mandarin aphasic speech assessment method can be readily extended to other languages.**

***Index Terms*—Aphasia assessment, cerebrovascular accident, deep neural network, machine learning, Mandarin, speech impairment, speech lucidity.**

I. Introduction

IN RECENT years, cerebrovascular accident, also known as stroke, has become the second most common cause of death and the leading cause of ongoing disabilities worldwide [2], [3]. According to the government report provided by the Ministry of Health of People's Republic of China [4]–[6], there are 3 million new stroke incidents, and 1.5 million stroke-related deaths each year, and the total number of survivors has exceeded 8 million.

Aphasia is generally caused by a cerebrovascular accident and occurs in about one-third of the stroke patients [7], [8]. It is an acquired neurogenic language disorder in which an individual's ability to produce or comprehend language is compromised [9], [10]. Aphasia may cause impairments in both expressive and receptive language skills, including speaking, writing, reading, and listening [11]–[13]. Patients with aphasia (PWAs) often face significant communication difficulties, which may lead to feelings of frustration, loss of autonomy, and social isolation, among others [14]. A research investigation found that aphasia patients had the poorest quality of life (QOL), followed by cancer patients and Alzheimer patients [15], [16]. Studies indicated that commencing an intensive guided treatment at the acute or sub-acute stage of aphasia leads to the highest recovery [17], [18]**.** However, the aphasia assessment and rehabilitation are resource-intensive processes which require the presence of a speech-language pathologist (SLP). This requirement is difficult to fulfil due to the vast number of PWAs and limited resources. Financial difficulties and patients' health conditions also pose difficulty in aphasia assessment and rehabilitation.

Mandarin is the most widely spoken dialect in China. However, due to a large population and inadequate SLPs, Mandarin-speaking aphasic patients are often ineffectively guided in speech and language rehabilitation procedures. Speech and language disabilities cause a poor QOL for aphasic patients as well as for their caregivers. Therefore, it is crucial to develop automated solutions to assist PWA, which can reduce the burden on rehabilitation centers and SLPs national-wide. In-home automated speech therapy with limited supervision will enable PWAs to self-monitor their verbal output, as well as assist SLPs

Manuscript received December 30, 2019; revised May 10, 2020 and June 12, 2020; accepted July 18, 2020. Date of publication July 22, 2020; date of current version November 5, 2020. This work was supported by Li Ka Shing Foundation Cross-Disciplinary Research under Grant 2020LKSFG04C. *(Corresponding author: Qiang Fang.)*

Seedahmed S. Mahmoud, Yiting Tang, Youcun Li, and Qiang Fang are with the Department of Biomedical Engineering, College of Engineering, Shantou University, Shantou 515041, China (e-mail: mahmoud@stu.edu.cn; 17yttang@stu.edu.cn; youcunlee@163.com; qiangfang@stu.edu.cn).

Akshay Kumar is with the School of Engineering, Royal Melbourne Institute of Technology University, Melbourne, VIC 3053, Australia (e-mail: akshay.kumar@student.rmit.edu.au).

Xudong Gu and Jianming Fu are with the 2nd Hospital of Jiaxing, Zhejiang 314000, China (e-mail: jxgxd@hotmail.com; fjm_7758@163.com).

Digital Object Identifier 10.1109/JBHI.2020.3011104

to prescribe appropriate therapy. In this paper, a standardized automatic speech lucidity assessment method for Mandarin-speaking aphasic patients using an efficient machine learning based technique and the Chinese Rehabilitation Research Center Aphasia Examination (CRRCAE) standard as a guideline is presented. The proposed solution envisages supporting in-home aphasia rehabilitation and the early supported discharge (ESD) scheme recommended by the World Health Organization (WHO). Also, the method will help to automate the manual speech lucidity assessment part of the CRRCAE standard; hence, the effort required from SLPs for using the CRRCAE standard will reduce. In addition, the suitability of deep learning for the speech impairment assessment has been compared with a widely used classical machine learning technique. The deep learning-based method utilizes a high-resolution time-frequency (TF) image at the network input nodes to improve the performance of the proposed speech lucidity assessment method.

This paper is organized as follows. Section II presents the related research on aphasia and methods employed in literature for an automatic aphasic speech assessment. In Section III, the Mandarin speech data collection and CRRCAE standard are discussed. In Section IV, a comparative investigation between classical machine learning (CML) algorithms and a deep neural network (DNN) algorithm to select a suitable algorithm for the development of an automatic Mandarin-speech lucidity assessment method is presented. Section V introduces the methodology of the proposed aphasic speech lucidity assessment method. Validation results and discussion of the findings are presented in Sections VI and VII, respectively.

II. Related Work

Several studies have been conducted on aphasia and speech impairment in China and worldwide. Researchers in [19] studied the association between post-stroke depression, aphasia, and physical independence in Chinese stroke patients at a 3-month follow-up. Their research found that aphasia was associated with physical dependence, though the association was statistically insignificant. In [20], research using Cantonese AphasiaBank database was conducted to analyze samples from important event narrative task produced by speakers with fluent aphasia and their controls in terms of topic and vocabulary. The study aimed to identify topics of interest to elderly speakers and the lexical items employed in talking about these topics, which may be used as potential training materials for language rehabilitation [20].

Several efforts on research towards automatic speech assessment for patients with aphasia and speech disorder have been made worldwide [21]–[23]. For example, researchers in [24]–[26] used classical machine learning (CML) techniques such as support vector machines (SVM) and Gaussian Mixture Model (GMM) to assess speech disorder automatically. In [26], SVM was used to distinguish between normal and pathological voices, while in [24], researchers used SVM and random forests in the task of Parkinson's disease classification. In [27], an aphasia speech recognition method based on MFCC features and dynamic time warping (DTW) algorithms was proposed. Their performance results showed that the average recognition rate of 10 numerical numbers was 78%, and that of 4 words was 77.5%. Also, a computer-assisted speech recognition system based on discrete wavelet transform (DWT) and artificial neural network for patients with aphasia and dysarthria was developed in [28]. Their preliminary clinical trials showed that patients' pronunciation level significantly improved after treatment ($p < 0.025$) [28]. However, there is a lack of studies on automatic assessment and rehabilitation solutions for Mandarin-speaking aphasic patients. Besides, most of these studies do not follow any well-known standard, such as CRRCAE for Chinese aphasic patients, to enable SLPs to reliably and efficiently assess a patient's disability level.

Overall, the major issues being faced in aphasia assessment and rehabilitation, especially in China, are 1) least attention is given to research on accurate speech impairment severity level assessment methods, 2) lack of reliable and standardized automatic tools to be used during assessment and rehabilitation for Mandarin aphasic patients [29], 3) the limitations of the Mandarin based CRRCAE standard which assesses the overall speech clarity rather than targeting specific lucidity features. In addition, utilization of the CRRCAE standard is time-consuming and subjective [30]. In this paper, we introduce a novel method that combines machine learning with a high time-frequency resolution imaging technique to develop an automatic speech lucidity assessment method that can be used among Mandarin-speaking aphasic patients. This research aims to address and resolve issues being faced in aphasia assessment and rehabilitation, especially in China.

III. Mandarin Speech Data Collection and the CRRCAE Standard

For the development and validation of the proposed aphasic speech lucidity assessment method, twelve patients with aphasia (PWA) (5 female, mean age: 61.8 ± 14.4 years) were recruited in this study from the 2nd Hospital of Jiaxing, Zhejiang province, China and from the 1st Affiliated Hospital of Shantou University, Guangdong province, China (see Table I). The data collection was approved by the ethics committees of the hospitals, and the procedures were strictly followed to ensure the investigation complied with the declaration of Helsinki. The criteria for excluding the patients were as follows: 1) patients with dementia, mental disorder and inability to communicate; 2) patients with severe uncorrelated visual and hearing impairment; 3) individuals who were addicted to tobacco and alcohol; 4) recent symptoms of respiratory tract infection such as rhinitis, tonsillitis, pharyngitis and other diseases that affect pronunciation; and 5) patients with a history of chronic throat or lung diseases. Table I shows the detailed information of the recruited aphasic patients.

Thirty-four healthy subjects (11 females, mean age: 21.5 ± 3.1 years) were also recruited from Shantou University (STU), China. Healthy participants were part of this investigation to help in setting a speech lucidity features benchmark for the assessment method. Particularly, their speech data were used in this paper to (1) compare the performance of various classical

TABLE I
APHASIC PATIENTS' INFORMATION

Patient ID	Gender, Male/ Female	Age, Yrs.	Days since stroke	Cardinal symptom	Native dialect
P01	F	29	274	Broca aphasia	Mandarin
P02	F	62	91	Broca aphasia, Dysarthria	Mandarin
P03	M	66	91	Broca aphasia	Mandarin
P04	F	57	122	Broca aphasia, Dysarthria	Mandarin
P05	M	56	8	Broca aphasia	Mandarin
P06	M	59	24	Broca aphasia	Teochew dialect
P07	F	70	1096	Anomic aphasia	Teochew dialect
P08	M	77	61	Dysarthria	Jiaxing dialect
P09	M	87	730	Dysarthria	Mandarin
P10	F	49	30	Dysarthria	Jiaxing dialect
P11	M	67	122	Combined aphasia	Jiaxing dialect
P12	M	62	152	Transcotical motor aphasia	Jiaxing dialect

TABLE II
SELECTED WORDS FROM CRRCAE STANDARD

Noun	Pinyin pronunciation	English Meaning
楼房	lou2 fang2	building
牙刷	ya2 shua1	toothbrush
钟表	zhong1 biao3	horologe
火	huo3	fire
电灯	dian4 deng1	lamp
椅子	yi3 zi1	chair
月亮	yue4 liang4	moon
自行车	zi4 xing2 che1	bicycle
鱼	yü2	fish
西瓜	xi1 gua1	watermelon
Verb	**Pinyin pronunciation**	**English Meaning**
写	xie3	write
哭	ku1	cry
游泳	you2 yong3	swim
坐	zuo4	sit
敲	qiao1	knock
穿衣	chuan1 yi1	dress
跳舞	tiao4 wu3	dance
喝水	he1 shui3	drink water
睡	shui4	sleep
飞	fei1	fly

machine learning and deep neural network (DNN) classification algorithms in automatic speech recognition (ASR), (2) evaluate the suitability of a quadrature time-frequency method for the classification of the speech data using DNN model in ASR and (3) train the weights of the DNN model to develop the proposed automatic speech lucidity assessment method.

Mandarin is the native dialect (mother tongue) for five recruited patients while the second dialect for the others. In metropolitan areas in China, Mandarin has already become very common. Though there is a slight difference between a person whose mother tongue is Mandarin and a person whose mother tongue is a local dialect when they speak in Mandarin, the difference becomes insignificant when it comes to the pronunciation of an isolated Mandarin word compared to speaking a complete sentence. In this paper, the collected speeches from all patients were Mandarin vowels, isolated Mandarin words, and combined Mandarin words (see Table II). Thus, a Mandarin speech lucidity assessment method for aphasic patients will serve a wide range of users whose mother tongue can be Mandarin or can be other dialects.

From CRRCAE standard, six Mandarin vowels, ten nouns, and ten verbs keywords were selected (see Table II). The speeches of healthy participants and aphasic patients while they were uttering the vowels and the selected words were recorded using Lenovo B613 voice recording pen with a sampling rate of 48000 samples/s. The voice recorder has a low-noise and high-fidelity sound. Healthy participants' speech data were recorded at Shantou University (STU) in a vacant office space where all external noise sources were eliminated. The recording environment of the patients (9 out of 12 patients) from the 2nd Hospital of Jiaxing was a professional designed speech therapy room with soundproof walls. The other three patients' speech was recorded in a vacant ward located in a quiet corner of the corridor in the 1st Affiliated Hospital of Shantou University. Hence the data collection environment for healthy subjects and patients was largely consistent. Furthermore, in order to remove any inconsistency left in the recorded speech data, the data were manually inspected thoroughly to remove any noisy and unqualified speech samples. Overall 4% of the speech samples were removed.

The six Mandarin vowels considered in this paper are: ā, ō, ē, ī, ū, and ǖ [31]. Each vowel was repeated with an average of 3 times per aphasic patient (with a minimum record of 1 time and a maximum of 7 times per patient based on their medical condition at the time of recording) and 6 times per healthy participant. Vowels are voiced speech where air from the lungs is modulated by the vocal cords and results in a quasi-periodic excitation, unlike unvoiced speech consonant where air from the lungs passes through a constriction in the vocal tract and becomes turbulent, a noise-like excitation. In [32], the categories of Mandarin vowels in terms of articulatory positions and acoustic properties were examined. The study shows that the changes in articulation do not necessarily change the acoustic output. For instance, the tongue height does not completely correspond to the vowel height informant values. Aphasia assessment for patients with high severity level can be difficult. Hence, Mandarin vowels are suitable to be considered in their situation.

The 20 keywords (ten nouns and ten verbs) included in this research have been taken from the CRRCAE standard [30] and have been shown in Table II. The selected words belong to

food, everyday objects, and activities categories. The Mandarin language has four tones, which are represented by numbers, as shown in Table II. Each word was repeated with an average of 3 times per aphasic patient and 5 times per healthy participant. Also, five of the twelve patients had recorded vowels only due to their medical condition. Referring to the Japanese Standard Aphasia Examination, CRRCAE was compiled in 1990, bearing in mind the Chinese language and Chinese culture [30]. Researchers in [33] used this standard to assess 151 healthy people. Their research finding showed that the CRRCAE standard is applicable to adult aphasic patients in different regions of China who speak Mandarin. Besides, their investigation showed that there is no significant difference in speech scores when considering participants of different occupations, genders, ages, and cultural levels ($p > 0.05$) [33]. In [34], the reliability and validity of CRRCAE standard were analyzed on 20 aphasic patients. The results indicated that CRRCAE has good reliability and validity and could be used as a quantitative indicator in speech and language rehabilitation of aphasia patients [34]. In addition, the results of [34] showed a very high correlation between CRRCAE and Aphasia Quotient (AQ) of Western Aphasia Battery (WAB) ($r = 0.948, p < 0.01$), indicating good validity of CRRCAE standard. However, as mentioned earlier, the CRRCAE standard is time-consuming, subjective and used to assess the overall speech clarity rather than targeting specific lucidity features [30].

IV. Selection of a Machine Learning Method

The proposed automatic Mandarin-speech lucidity assessment method relies immensely on the accuracy of the chosen machine learning method. Therefore, in this section, a comparative study will be introduced to assist the selection of a suitable machine learning technique. Classical machine learning algorithms utilizing conventional ASR features, as inputs, for the classification of speech signals will be compared with a deep neural network (DNN) algorithm utilizing quadrature based high-resolution TF images at its input nodes [35]. Mandarin vowels and the selected isolated keywords speech data from healthy participants are used in this comparison.

A. Classical Machine Learning

A typical Mandarin ASR system consists of a pre-processing stage that extracts unique features from the recorded vowels/words speech, and a classifier that assigns the computed features to a class of one of the Mandarin vowels: ā, ō, ē, ī, ū and ǖ and the twenty isolated keywords (see Table II).

A total of 51 features based on Mel Frequency Cepstral Coefficients (MFCCs), the formant frequencies, and signal energy were calculated for the classification [36]–[38]. The feature vectors of all data samples were standardized by subtracting mean and dividing by the standard deviation of each of the data sample's feature vector, to avoid some features from potentially dominating others due to a large magnitude. Several classical machine learning algorithms were evaluated for the classification of the speech signals; specifically, quadratic support vector machine (QSVM), a radial basis function (RBF) kernel SVM, linear discriminant analysis (LDA), random forest and k-nearest neighbors (kNN) were evaluated. Two separate datasets were constructed: one containing all vowels and words speech data (26 classes) of healthy participants (named *vowels + words* dataset onwards) and the other containing only words speech data (20 classes) of healthy participants (named *only-words* dataset onwards). Five-fold cross-validation was used to estimate the performance of the classifiers to classify the two datasets. Fig. 2 shows the step-by-step methodology employed for the performance evaluation of the classical machine learning algorithms to classify *vowels + words* and *only-words* datasets.

B. Deep Neural Network With High-Resolution TF Image Set

In the last few years, deep learning based algorithms, precisely the convolutional neural networks (CNN), have given outstanding classification results in the field of computer vision [39]. For CNN based classification of speech signals, a TF representation of the signals in the form of an image can be used as an input to the CNN model. However, in some of the speech signals such as in Mandarin speech, we may be confronted by multiple components with narrow separation in time or frequency or both; in such a case, many time-frequency distributions (TFDs) fail to reveal the true structure of the signal, as many frequency components will overlap. This will result in image distortion and resolution degradation and, consequently, a poor image classification performance [40], [41]. A TFD that reveals the exact multi-component structure of a speech signal can improve the classification of speech signals using CNN. Therefore, a TFD with a joint time-frequency resolution is essential. In [35], we found that the Hyperbolic T-distribution (HTD) [1] outperforms wavelet transform, Wigner-Ville distribution (WVD), Choi-Williams distribution (CWD) and the Exponential T-distribution (ETD), in terms of resolution, signal-to-noise (SNR) performance and cross-terms reduction. Hence, the HTD can produce high-resolution TF images of Mandarin speech signals, which were utilized as inputs to the CNN model for the speech signal classification.

In Quadrature TFD, the continuous TFD of the analytic signal $z(t)$associated with the original real signal *s(t)* can be expressed as follows [1]:

$$\rho\left(t,f\right) = F_{\tau\to f}\left[G\left(t,\tau\right) *_{(t)} K_z\left(t,\tau\right)\right] \tag{1}$$

where $K_z(t,\tau) = z(t+\tau/2)z^*(t-\tau/2)$ is the instantaneous autocorrelation product, F is the Fourier transform, $G(t,\tau)$ is the time-lag kernel and $*_{(t)}$ denotes time convolution. The kernel for the HTD is given by [1]

$$G\left(t,\tau\right) = R_\sigma\left(t\right) = \frac{k_\sigma}{\cosh^{2\sigma}\left(t\right)} \tag{2}$$

where σ is a real positive number and k_σ is a normalization factor given by

$$k_\sigma = \int_{-\infty}^{\infty} \frac{1}{\cosh^{2\sigma}\left(t\right)} dt = \frac{\Gamma\left(2\sigma\right)}{2^{2\sigma-1}\Gamma^2\left(\sigma\right)} \tag{3}$$

in which Γ represents the gamma function. The TFD images of the word *lou2 fang2* (Building) along with their time-domain

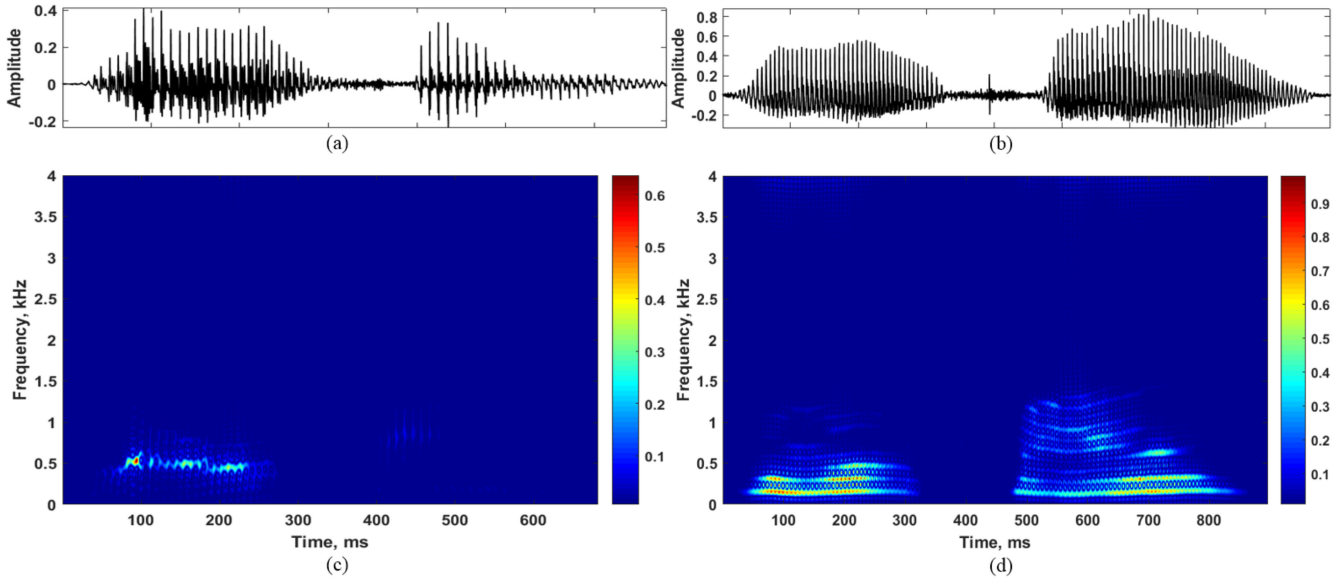

Fig. 1. Raw time-domain waveform of (a) a healthy participant and (b) a patient with aphasia when they spoke a Mandarin noun *lou2 fang2* (Building). (c, d) Time-frequency transform of speech data of a Mandarin noun *lou2 fang2* (Building) of a healthy participant and an aphasic patient using Hyperbolic T-distribution (HTD), respectively [1].

waveforms have been shown in Fig. 1(c, d) and Fig. 1(a, b), respectively.

State-of-the-art computer vision CNN architectures are data-hungry architectures and have hundreds of thousands of trainable parameters. Hence training them with small datasets such as here is challenging and can lead to overfitting [39]. Thus, transfer learning (TL) was utilized for training the CNN model in this research. The key idea behind transfer learning is to fit the CNN model with the speech data distribution $P_s(x, y)$ from a pre-trained CNN model distribution $P_p(x, y)$ that has some common knowledge structure with the $P_s(x, y)$ instead of fitting from random parameters, where x and y are the data sample and its label, respectively, and thereby resulting in better classification performance.

State-of-the-art pre-trained CNN model ResNet-34 was utilized for $P_p(x, y)$ which was fine-tuned with healthy participants' speech TFD images [42]. Two separate models for the two datasets: *vowels + words* dataset (26 classes) and *only-words* dataset (20 classes) were trained. In order to utilize the weights of the pre-trained ResNet-34 model for transfer learning, it is imperative to transform the input to the same format the pre-trained model was originally trained on [42]. Therefore, all TFD RGB color images were resized to $224 \times 224 \times 3$ pixels and normalized as per ImageNet dataset characteristics, before feeding them to the pre-trained ResNet-34 CNN model [42]. The fine-tuning of the pre-trained model was carried out using cyclical learning rates with a maximum learning rate set to 0.03 [43]. The optimum value for the maximum learning rate was estimated through a learning rate range test [43]. ADAM was used as an optimizer with default parameters for $\beta 1 = 0.9$ and $\beta 2 = 0.999$ with a cross-entropy loss function [44]. Weight decay with an empirically chosen multiplying factor 0.01 was used to prevent overfitting [45]. Five-fold cross-validation was adopted to estimate the performance of the classifier in classifying the two datasets. The CNN model for the classification of *vowels + words* dataset was trained for 20 epochs, and the model for the classification of *only-words* dataset was trained for 15 epochs, both with a batch size of 128. Fig. 2 shows the step-by-step methodology employed for the performance evaluation of the CNN model to classify *vowels + words* and *only-words* datasets. All models were trained on NVIDIA Tesla

TABLE III
PERFORMANCE EVALUATION OF QSVM, RBF-SVM, LDA, RANDOM FOREST, KNN, AND CNN CLASSIFICATION ALGORITHMS FOR THE CLASSIFICATION OF *ONLY-WORDS* (20 CLASSES) AND *VOWELS + WORDS* (26 CLASSES) DATASETS IN TERMS OF ACCURACY, PRECISION AND RECALL VALUES ON FIVE-FOLD CROSS-VALIDATION.

Dataset	Classification Method	Accuracy (%)	Precision (%)	Recall (%)
Only-words	Quadratic SVM	93.81	94.09	93.82
	RBF SVM	92.40	92.94	92.42
	LDA	**95.74**	**95.77**	**95.75**
	Random Forest	93.45	93.49	93.47
	kNN	80.22	83.56	80.24
	CNN	**99.23**	**99.22**	**99.24**
Vowels + words	Quadratic SVM	92.05	92.28	92.07
	RBF SVM	90.00	90.54	89.99
	LDA	**93.95**	**93.97**	**93.96**
	Random Forest	92.18	92.18	92.17
	kNN	76.49	80.84	76.50
	CNN	**97.39**	**97.49**	**97.42**

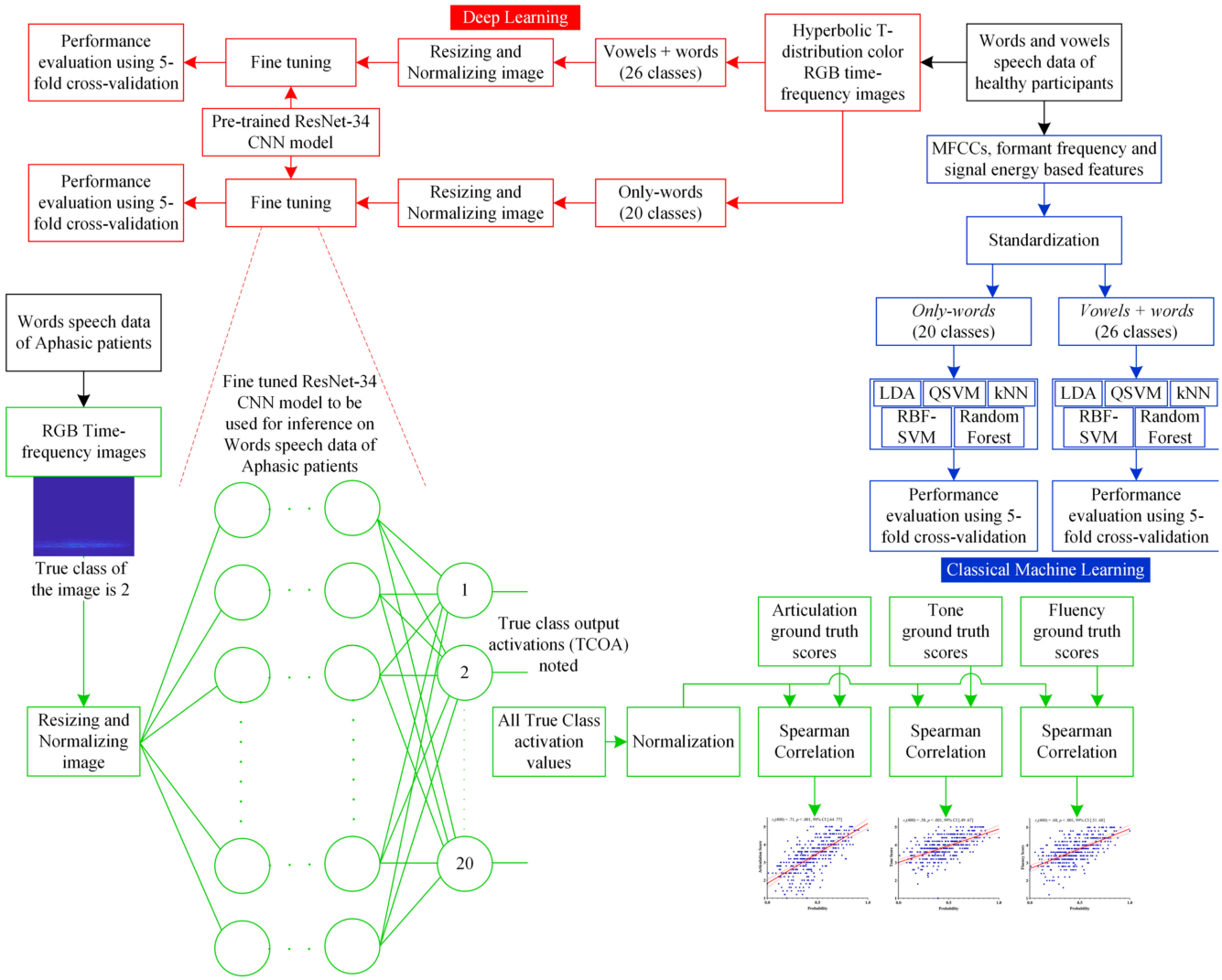

Fig. 2. Step-by-step methodology employed for the performance evaluation of the CNN model which utilizes Hyperbolic T-distribution color RGB time-frequency images (in red), and classical machine learning algorithms which utilize MFCCs, formant frequency and signal energy based features (in blue), for the classification of *vowels + words* and *only-words* datasets. The bottom half of the image (in green) shows the methodology adopted for carrying out correlation analyses between the normalized true-class output activations (TCOA) of aphasic patients' words speech data when put through ResNet-34 CNN model trained on healthy participants' *only-words* dataset, and the ground truth scores of the three fundamental features of Mandarin language, namely articulation, fluency, and tone.

P40 GPU in fastai, a PyTorch based deep neural networks library [46].

C. Result Comparison

This section presents a comparison among the performance of various classification algorithms employed in this study to classify *vowels + words* and *only-words* datasets. Accuracy, precision, and recall were used as the performance evaluation metrics for the comparison whose results have been shown in Table III. The comparative results show that the CNN based classification method performed best among all the classification algorithms employed for the classification of the two datasets, in terms of the three performance evaluation metrics. The performance of LDA was observed to be the highest among the classical machine learning (CML) algorithms considered. Confusion matrixes of *only-words* dataset classification using CNN and LDA have been shown in Fig. 3. Confusion matrixes of *only-words* dataset classification using all other classifiers employed in this study and confusion matrixes of vowels+ words dataset classification have been supplied as supplementary material. Based on these results, the ResNet-34 CNN model with HTD TF images as input was considered for further analysis. This model will be used in the next section to develop the proposed assessment method.

V. Aphasic Speech Lucidity Assessment Method

A challenging and engaging rehabilitation program helps stroke patients recover faster from the disability due to stroke [47]. Assessment of the speech disability levels plays a crucial role when it comes to personalizing the rehabilitation program.

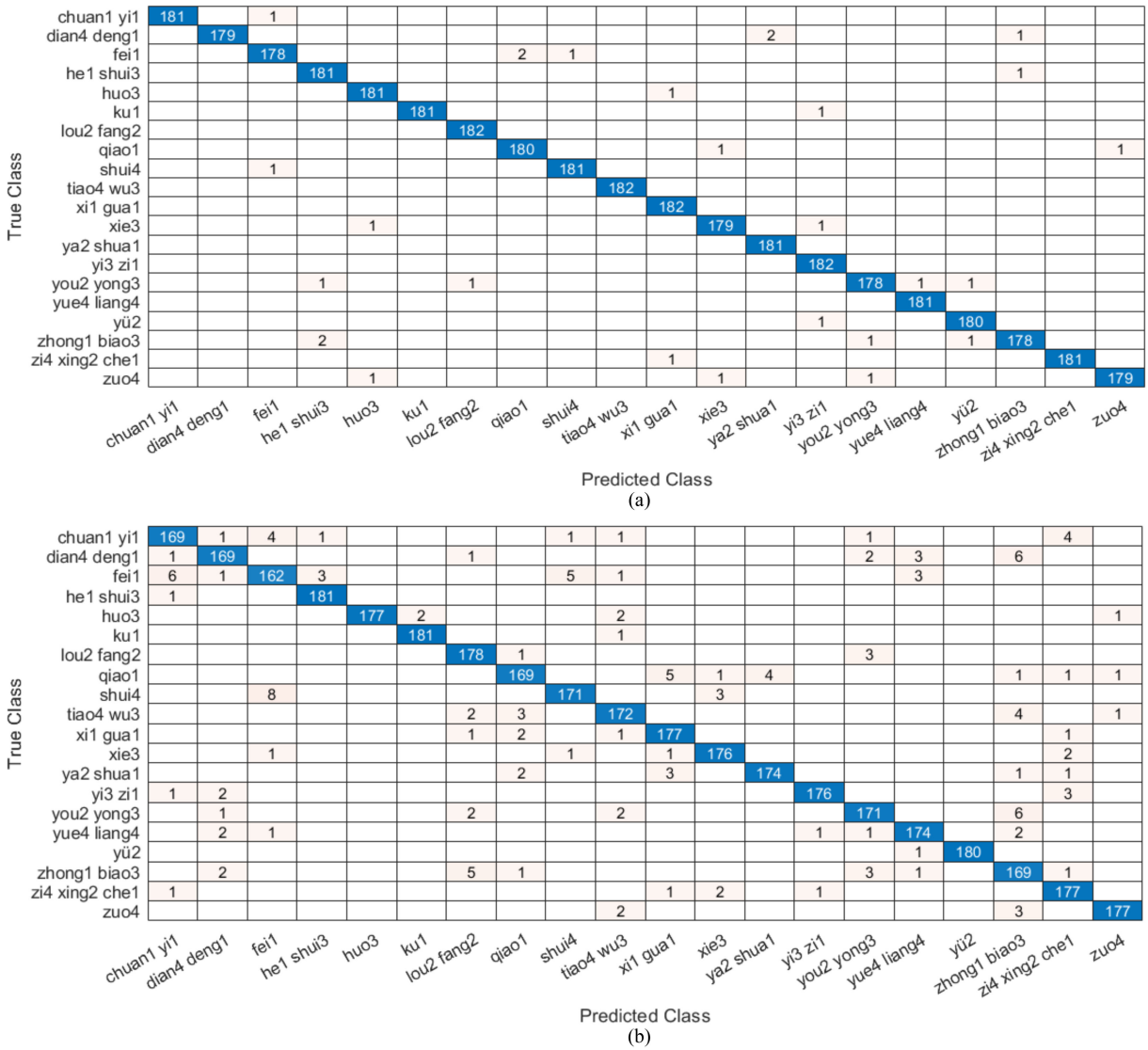

Fig. 3. Confusion matrix of *only-words* dataset classification using (a) convolutional neural networks and (b) linear discriminant analysis.

The conventional assessment and rehabilitation techniques of aphasia [30], [48]–[50] are resource-intensive and thus are unsustainable due to the growing number of PWA. Therefore, an automatic aphasia disability assessment method using the CRRCAE standard as a guideline is presented in this section, which can mitigate the issues mentioned above. In principle, the proposed method compares the speech of an aphasic patient with the speech of healthy speakers and objectively scores the speech lucidity of the aphasic speech. This procedure is carried out using the ResNet-34 CNN model trained on health people Mandarin speech and without the involvement of an SLP. The general idea of the method has been depicted in Fig. 2 (in green). The method is presented in the next subsections.

A. Extraction of Ground Truth

An important problem in this work involves constructing the ground-truth labels of the speech dataset of PWA in the absence of standard assessment scores. Conventionally, as a part of the assessment procedure, speech-language pathologists give a score to the speech of a PWA. However, due to the limited availability of SLPs, it is difficult to carry out this task at a considerably large scale. Nevertheless, a previous study has shown that an untrained listener of a language can evaluate the lucidity of speech close to a specialist level [51].

Therefore, three scholars who had Mandarin as their first language were recruited from the language department of Shantou University, China, to score the speech data of the PWA. Mandarin words have three fundamental features that define the overall lucidity of the spoken words, namely articulation, fluency, and tone [52], [53]. The scholars scored the recorded speech samples of the PWA on each of the three-fundamental features on a 5-point Likert scale. The aphasic patients' *only-words* dataset was considered in this task. The scores were further evaluated by two native Mandarin speakers recruited from the university. Scholars received monetary remuneration of 15 CNY per hour for their services. For each of the three fundamental features, means of the scores of the three scholars and the two native speakers were

calculated. For each of the 402 samples of the speech data of PWA, we ended up having a score ranging from 0 to 5 for each of the three fundamental features: articulation, fluency, and tone, which were used as the ground truth for the *only-words* speech data of PWA. The ground truth scores for each of the 402 speech samples were mapped to its corresponding true-class output activation (TCOA) produced by the ResNet-34 CNN model to form the speech lucidity assessment method, as discussed in next Subsection.

B. Assessment of Mandarin Speech Lucidity

As shown in Table III, healthy participants' *only-words* dataset has been classified with accuracy greater than 99% using the ResNet-34 CNN model. Thus, a word speech sample of a healthy individual, when fed to the CNN model, would be correctly classified with a high probability. On the other hand, a word speech sample of an aphasic patient, when fed to the CNN model, would be correctly classified with a lower probability or can be misclassified because the speech signature of an aphasic speech differs from the healthy speech due to the associated speech disability. In other words, a high lucidity speech data sample would be predicted correctly by the model with a high probability; however, a low lucidity PWA speech data sample would be predicted correctly by the model with a lower probability or even can get misclassified. Therefore, we hypothesized that the ability of the CNN model trained on healthy participants' speech data to classify PWA speech samples would correlate with the lucidity of the PWA speech samples and thereby can be used as a tool for the PWA speech assessment.

Therefore, the ResNet-34 CNN model trained on the healthy participants' *only-words* dataset was utilized to assess the lucidity of speech samples of the PWA. The word speech data of PWA were converted into TF images using the HTD method. The images were resized to $224 \times 224 \times 3$ pixels and were fed to the trained ResNet-34 model. As the PWA words speech dataset contained 20 classes, the ResNet-34 model produced 20 activations (one for each class) at the output. The output activation at the true-class node (TCOA) was noted as depicted in Fig. 2 (in green), and the procedure was repeated for the whole PWA words speech dataset. The raw activations values were normalized in the range $[0, 1]$ (named normalized TCOA onwards) for further processing.

Kolmogorov-Smirnov test ($\alpha = 0.05$) was used to confirm the normal distribution of ground truth articulation, tone, and fluency scores as well as of normalized TCOA before further analysis. The null-hypothesis of Kolmogorov-Smirnov assumes a normal-distribution, while the alternative hypothesis denies that. Skewness and Kurtosis z-value were used as an additional measure to confirm the normality. Except for the normalized TCOA, the null hypothesis of normal distribution was rejected for the three ground truth scores. The *p*-values were found to be less than 0.05, and additionally, skewness and kurtosis z-values were outside the ± 1.96 range. As a result, a non-parametric measure of correlation, Spearman rank correlation coefficient was used to evaluate the correlation between the normalized TCOA and the three ground truth scores; the corresponding *p* values were calculated using two-tailed student's t-distribution. The threshold for significance was set *a priori* at $\alpha = 0.05$. All statistical analysis was conducted using IBM SPSS Statistics 26. We hypothesized that a strong positive correlation would exist between the normalized TCOA and the three ground truth scores.

VI. Results

A machine learning algorithm that can reduce the challenges being faced in the automatic recognition of impaired speech is crucial. The challenges are abnormal speech patterns [54], speaker variability [55], and data scarcity [56]. Comparative results in Table III show that the CNN classification model with hyperbolic T-distribution based TF images as input outperforms the classical ASR approach, which utilizes MFCCs, formant frequencies, and signal energy-based features. Results in Table III show the high accuracy, precision, and recall metrics of the CNN based speech model when classifying *only-words* and *vowels + words* datasets. Therefore, the CNN based aphasic speech model was selected to investigate its suitability for automatic Mandarin aphasic speech lucidity assessment.

Fig. 4 shows the scatter plots between the normalized TCOA and the mean ground truth articulation, fluency, and tone scores. The plots show that recruited patients have a wide range of speech severity. The purpose of these plots is to investigate the efficiency of the proposed speech assessment method in assessing PWA's speech at the fundamental level of the lucidity. The results indicate that the relationship between the proposed model's normalized TCOA and the ground truth scores follow a linear model of the form

$$L_s = \mu M_p + \beta \tag{4}$$

where L_s is the score given by the proposed assessment method to a PWA's speech, M_p is the CNN model's normalized TCOA, μ is the slope, and β is the y-intercept with the Likert scale axis. Fig. 4 also shows the correlation coefficients (CC), *p*-values, and the 99% confidence intervals (CI) for the linear association between the normalized TCOA values and the articulation, fluency, and tone scores.

From Fig. 4(a), the linear model for the articulation feature is given by

$$L_s = 3.357\, M_p + 1.826 \tag{5}$$

This formula shows that the least articulation score that the articulation feature assessment model exhibits is 1.826. A statistically significant correlation was observed between the CNN model's normalized TCOA values and the ground truth articulation feature scores, $(r_s(400) = 0.71,\ p < 0.001,\ 99\%\ \text{CI}\ [0.64\ 0.77],\ R^2_{adj} = 0.48)$.

From Fig. 4(b), the linear model for the fluency feature is given by

$$L_s = 2.251\, M_p + 2.671 \tag{6}$$

This formula shows that the least fluency score that the fluency feature assessment model exhibits is 2.671. A

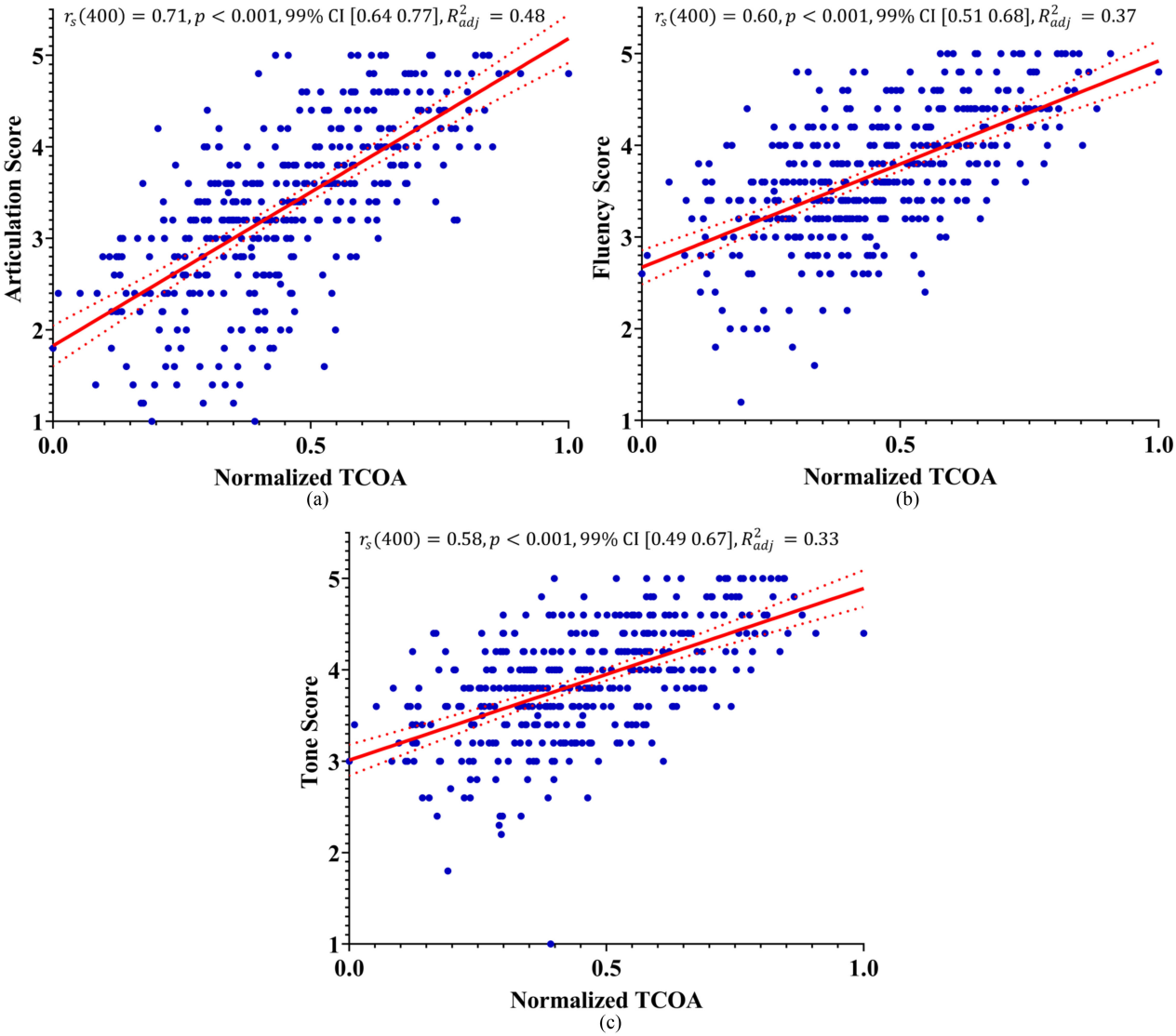

Fig. 4. Scatter plots between the normalized true-class output activations (TCOA) and (a) articulation score, (b) fluency score and (c) tone score. The solid red line in the three figures represents the best fit line using simple linear regression and the dotted red line represents the 99% confidence interval (CI). The values of Spearman rank correlation coefficient (r_s), p-value, 99% CI and adjusted R^2 are also shown on the top of the plots.

statistically significant correlation was also observed between the CNN model's normalized TCOA values and the ground truth fluency feature scores, $(r_s(400) = 0.60,\ p < 0.001,\ 99\%\ \text{CI}\ [0.51\ 0.68],\ R^2_{adj} = 0.37)$.

From Fig. 4(c), the linear model for the tone feature is given by

$$L_s = 1.878\, M_p + 3.012 \tag{7}$$

This formula shows that the least tone score that the tone feature assessment model exhibits is 3.012. A statistically significant correlation was also observed between the CNN model's normalized TCOA values and the ground truth tone feature scores,$(r_s(400) = 0.58,\ p < 0.001,\ 99\%\ \text{CI}\ [0.49\ 0.67],\ R^2_{adj} = 0.33)$.

The three linear relationships (5, 6 and 7) of the assessment method can be used to score new aphasic patients' speech lucidity by feeding their spoken words through the trained CNN model to obtain the normalized TCOA, M_p. The normalized TCOA value will be substituted in the three formulas to estimate the scores, L_s, for the three speech lucidity features. In this paper, the uncertainty of the lucidity scores predicted by the proposed linear models can be reduced by averaging the assessment scores of multiple words. The results show that the articulation feature has a higher correlation than the other two speech lucidity features. Moreover, a linear proportional relationship has been observed between the proposed aphasic speech assessment models' output and patients' severity levels, $L_s \propto M_p$. The dynamic range of L_s for the three features roughly lies between 1 to 5. Also, the linear model for the articulation feature is steeper compared to other features.

The linearity of the proposed aphasic speech assessment method and its significant correlation with the manual human scoring shows the efficacy of this method in evaluating the severity of aphasic patients' speech. Unlike other automatic

aphasic speech assessment methods [24]–[26], [48], the proposed method used the advantage of CNN's image classification abilities and the high-resolution TF images produced by the HTD to assess the patients' speech severity level.

VII. Discussion

The goal of this study was to propose a machine learning based method for speech assessment of Mandarin-speaking aphasic patients. Overall, a widely used machine learning based ASR technique and a novel method that utilizes the CNN model and a high-resolution TF imaging technique for ASR were compared. True-class output activations (TCOA) of the CNN model were used to develop a novel method for aphasic patients' speech assessment. Significantly strong correlations were observed between the TCOA of the CNN model and ground truth scores of the patients' speech, which validates the proposed aphasic speech lucidity assessment method.

For the articulation feature, the production of initial and final characters of a spoken Mandarin word not only depends on the vocal cords but also requires the articulation organs such as lips and tongues [57]. Initial and final characters that are responsible for articulation are segmental phonemes, which are mainly controlled by the function of the left hemisphere of the brain [58], [59], while tones are super-segmental phonemes, which are controlled by the right hemisphere of the brain [60], such as rhythm and prosody [53]. Most aphasia patients are affected by the damage to the left hemisphere of the brain [58], [59], which is the dominant side for language functions for most people, resulting in impaired language function. The articulation result in Fig. 4(a) shows the impact of the aphasia on the recruited patients' articulation system. The linear model exhibits a minimum articulation value of 1.826, which is lowest among the three features' minimum values. The reason for this is that the recruited patients included five patients with dysarthria (see Table I). Dysarthria is a speech disorder that results in unclear articulation of speech. Recruiting such patients in this research helped in examining the performance of the proposed speech lucidity assessment method over a wider range of speech disorder patients. Moreover, the model demonstrated its efficacy in response to dysarthria in the articulation feature, while the tone and fluency features were impacted to some extent. The tone and fluency feature assessment models show that these two features have better retention in aphasic patients in comparison to the articulation feature. Tone production is reflected in the fundamental frequency change of pronunciation, which depends on the degree of tension of the vocal cords [57].

The healthy participants' speech dataset has been utilized as a benchmark for the assessment method in this study. However, the recruited healthy participants and aphasic patients had a significant age difference. To investigate if the age difference in the recruited healthy participants and aphasic patients affects the aphasic speech assessment method, the speech assessment of the patients with significant age difference was analyzed. For the 49 years old P10 patient (see Table I), when speaking 'chuan1 yi1' Mandarin verb (see Table II), the mean articulation, fluency, and tone score were 1.2, 2.2, and 3, and the normalized TCOA was 0.35, respectively. While for the 87 years old P09 patient, when speaking the same Mandarin verb, the mean articulation, fluency, and tone score were 4.4, 4.4, and 4.6, and the normalized TCOA was 0.73, respectively. Despite being significantly older, P09 patient exhibited higher TCOA activations than P10 patient, which can be attributed to the comparatively higher recovery level of P09 patient as he had stroke 730 days before the experiment date, while P10 patient had stroke 30 days before the experiment date. A similar trend was observed in other patients also. Also, researchers in [33] showed that there is no significant difference in speech scores when considering participants of different ages and genders. Thus, it can be said that the age difference in the participants of this study has no effect on the aphasic speech lucidity assessment method.

Conventionally, MFCC based features along with classical machine learning (CML) algorithms are used for ASR. As observed in this research, the conventional technique has given a high accuracy for ASR. However, the performance of CML based approaches largely depends on how faithfully the extracted features represent the underlying characteristics of the dataset under consideration. On the other hand, turning the ASR problem to an image classification problem enabled the use of CNN, which permitted automatic detection of important features for ASR and also gave better classification results in comparison to the CML. Notably, high correlations between the normalized TCOA and the three fundamental features not only shows the ability of the CNN model to evaluate the speech quality through the HTD TF images but also provides a novel basis to assess the aphasic patients' speech disability level.

A number of standard assessment schemes such as Aachen Aphasia Test (AAT) [50], Western Aphasia Battery (WAB) assessment test [48], [49], and CRRCAE standard [30] are being used by medical practitioner around the world for the assessment of speech disability of PWA. The sophisticated assessment techniques are time-consuming and require the expertise of the assessment procedure to carry out the assessment. Moreover, a standard like CRRCAE standard [30] assesses the overall speech clarity rather than targeting specific lucidity features. In several automatic speech assessment pieces of research, these standards were embraced as a guideline to assess aphasic patients [21]–[23], [31]. Most of these studies use a classical machine learning approach based on ASR features extraction. Using feature extraction methods for the assessment of speech intelligibility features [48] makes the assessment prone to patient variability [55] effects. In [22], the automatic speech assessment for PWA relies on the manual transcription of audio speech into Chinese characters. Further, the characters were converted into Cantonese syllables using a pronunciation lexicon. The severity assessment method in their investigation followed Cantonese Aphasia Battery. CML classifiers were used to classify the transcribed speech into its correct task. However, Cantonese speech lucidity features (i.e., articulations, tones, and fluency) assessment were not considered in their study. Like Cantonese, Mandarin is a monosyllabic and tonal dialect. Each Mandarin character is spoken as a monosyllable with a specific tone, and speech lucidity features influence the meaning of the spoken utterance and the overall speech clarity [52], [53]. For the

articulation feature, patients with aphasia sometimes have poor coordination among speech organs, which affects the pronunciation of the initial and final consonants [53]. Hence, it results in a wrong meaning. Similarly, with the tone feature, a wrongly pronounced tone makes the word sound like another word [53]. Due to the importance of these features, an aphasic speech assessment method should target the speech lucidity features individually.

The proposed automatic speech lucidity assessment method has the capability of targeting individual speech lucidity features; hence it will enable SLPs to identify the weak fundamental attributes of the aphasic speech. Customized rehabilitation sessions for the impaired fundamental speech lucidity feature can improve the efficacy of the rehabilitation procedure, in terms of the rate of recovery from the speech disability. Likewise, aphasia assessment and rehabilitation for patients with a high severity level can be difficult when using words or continuous speech. Hence, Mandarin vowels can be suitable for their assessment and rehabilitation.

VIII. Conclusion and Future Work

In this paper, a standardized automatic speech lucidity assessment method for Mandarin-speaking aphasic patients using an efficient machine learning based technique has been proposed. The proposed solution envisages supporting in-home aphasia rehabilitation and the early supported discharge (ESD) scheme recommended by the World Health Organization (WHO). We compared the performance of the ResNet-34 CNN model and a widely used classical machine learning technique in automatic speech recognition (ASR) to evaluate their suitability for the development of the speech impairment assessment method. A high-resolution time-frequency imaging technique was used to improve the CNN model accuracy. In the comparative results, the CNN model outperformed the conventional machine learning based technique in automatic speech recognition.

The proposed speech lucidity assessment method linearly modeled the articulation, fluency, and tone features with significant correlations between the CNN model's normalized TCOA and the aphasic patients' severity levels. The articulation linear model was also able to represent and verify the recruited patients' data, where it has shown the low articulation score due to the dysarthria patients. The method showed that the fundamental lucidity features of Mandarin speech could be assessed and targeted individually. The generalization ability of the proposed method is evident from its observed performance despite noticeable variability in recruited patients' age, native dialects, cardinal symptoms, and days since the stroke. The proposed assessment method can provide feedback to aphasic patients about their verbal output during the rehabilitation procedure, without a speech-language pathologist intervention, thus improving the effectiveness of in-home aphasia rehabilitation therapies. In the future, the generalization of the proposed method over other Chinese dialects as well as over other international languages will be investigated. Provision of aphasic speech severity levels or ground-truth labels of speech lucidity is essential for the generalization of this method to other languages.

References

[1] Z. M. Hussain and B. Boashash, "Adaptive instantaneous frequency estimation of multicomponent FM signals using quadratic time-frequency distributions," *IEEE Trans. Signal Process.*, vol. 50, no. 8, pp. 1866–1876, Aug. 2002.

[2] National Stroke Foundation, "National Stroke Audit Rehabilitation Services Report 2012," 2012.

[3] R. Bonita, S. Mendis, T. Truelsen, J. Bogousslavsky, J. Toole, and F. Yatsu, "The global stroke initiative," *Lancet Neurol.*, vol. 3, no. 7, pp. 391–393, Jul. 2004.

[4] X. M. Cheng *et al.*, "Stroke in China, 1986 through 1990," *Stroke*, vol. 26, no. 11, pp. 1990–1904, Nov. 1995.

[5] Q. Jia, L.-P. Liu, and Y.-J. Wang, "Stroke in China," *Clin. Exp. Pharmacology Physiol.*, vol. 37, no. 2, pp. 259–264, 2010.

[6] W. W. Zhang, S. Speare, L. Churilov, M. Thuy, G. Donnan, and J. Bernhardt, "Stroke rehabilitation in China: a systematic review and meta-analysis," *Int. J. Stroke*, vol. 9, no. 4, pp. 494–502, Jun. 2014.

[7] S. T. Engelter *et al.*, "Epidemiology of aphasia attributable to first ischemic stroke," *Stroke*, vol. 37, no. 6, pp. 1379–1384, 2006.

[8] S. A. Thomas and N. B. Lincoln, "Predictors of emotional distress after stroke," *Stroke*, vol. 39, no. 4, pp. 1240–1245, Apr. 2008.

[9] A. R. Luria, *Higher Cortical Functions in Man*. Berlin, Germay: Springer, 1966.

[10] P. Shinn and S. E. Blumstein, "Phonetic disintegration in aphasia: acoustic analysis of spectral characteristics for place of articulation," *Brain Lang.*, vol. 20, no. 1, pp. 90–114, Sep. 1983.

[11] A. Basso, *Aphasia and Its Therapy*. New York, NY, US: Oxford University Press, 2003.

[12] G. A. Davis, *Aphasiology: Disorders and Clinical Practice*, 2nd ed. London, U.K.: Pearson, 2007.

[13] N. Helm-Estabrooks, M. L. Albert, and M. Nicholas, *Manual of Aphasia and Aphasia Therapy*, 3rd ed. Austin, TX, USA: Pro-Ed, 2013.

[14] N. Simmons-Mackie, A. Raymer, E. Armstrong, A. Holland, and L. R. Cherney, "Communication partner training in aphasia: A systematic review," *Archives Phys. Med. Rehabil.*, vol. 91, no. 12, pp. 1814–1837, Dec. 2010.

[15] J. M. C. Lam and W. P. Wodchis, "The relationship of 60 disease diagnoses and 15 conditions to preference-based health-related quality of life in ontario hospital-based long-term care residents," *Med. Care*, vol. 48, no. 4, pp. 380–387, Apr. 2010.

[16] W. Qiu *et al.*, "Psychometric properties of the chinese-version stroke and aphasia quality of life scale 39-generic version (SAQOL-39g)," *Topics Stroke Rehabil.*, vol. 26, no. 2, pp. 106–112, Feb. 17, 2019.

[17] S. K. Bhogal, R. Teasell, and M. Speechley, "Intensity of aphasia therapy, impact on recovery," *Stroke*, vol. 34, no. 4, pp. 987–992, Apr. 2003.

[18] R. R. Robey, "A meta-analysis of clinical outcomes in the treatment of aphasia," *J. Speech, Lang., Hearing Res.: JSLHR*, vol. 41, no. 1, pp. 172–187, Feb. 1998.

[19] S. Wang *et al.*, "The association between post-stroke depression, aphasia, and physical independence in stroke patients at 3-month follow-up," *Front. Psychiatry*, vol. 9, Aug. 20, 2018, Art. no. 374.

[20] S.-P. Law, A. P.-H. Kong, and C. Lai, "An analysis of topics and vocabulary in Chinese oral narratives by normal speakers and speakers with fluent aphasia," *Clin. Linguistics Phonetics*, vol. 32, no. 1, pp. 88–99, 2018.

[21] D. Le, K. Licata, and E. M. Provost, "Automatic quantitative analysis of spontaneous aphasic speech," *Speech Commun.*, vol. 100, pp. 1–12, Jun. 2018.

[22] Y. Qin, T. Lee, S. Feng, and A. P.-H. Kong, "Automatic speech assessment for people with aphasia using TDNN-BLSTM with multi-task learning," in *Proc. Interspeech*, 2018, pp. 3418–3422.

[23] D. Le, "Towards automatic speech-language assessment for aphasia rehabilitation," Ph.D. Thesis, University of Michigan, Dept. Comput. Sci. Eng., 2017.

[24] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[25] M. A. Shahin, B. Ahmed, J. McKechnie, K. J. Ballard, and R. Gutierrez-Osuna, "A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech," in *Proc. Interspeech*, 2014, pp. 1583–1587.

[26] R. Amami and A. Smiti, "An incremental method combining density clustering and support vector machines for voice pathology detection," *Comput. Elect. Eng.*, vol. 57, pp. 257–265, Jan. 2017.

[27] H. Ding, L. M. Lin, G. N. Wu, Y. F. Liu, and H. F. Xiao, "Jisuanji fuzhu yanyu jiaozhi xitong [A computer aided speech correction system]," *Zhongguo Shengwu Yixue Gongcheng Xuebao*, vol. 14, no. 1, pp. 39–44, Mar. 1995.

[28] A. N. Li, "Shiyuzheng huanzhe yuyin xinhao de shibie yanjiu [An investigation on speech recognition of aphasia patients]," M.S. thesis, Xi'an University of Science and Technology, Xi'an, China, Dept. Meas. Technol., 2010. [Online]. Available: CNKI

[29] H. Yang *et al.*, "A Chinese version of the Language Screening Test (CLAST) for early-stage stroke patients," *PLOS ONE*, vol. 13, no. 5, 2018, Art. no. e0196646.

[30] C. R. R. Center. Chinese Rehabilitation Research Center Aphasia Examination (CRRCAE), 2019. [Online]. Available: https://wenku.baidu.com/view/e209482cbd64783e09122bb5.html

[31] S. S. Mahmoud, Q. Fang, Y. Tang, M. Alsulami, and M. Alotaibi, "Automatic mandarin vowels recognition framework for aphasic patients rehabilitation," in *Proc. 13th IEEE-EMBS Int. Summer School Symp. Med. Dev. Biosensors (MDBS2019)*, Sep. 2019.

[32] C.-H. Wu and C. Shih, "Mandarin vowels revisited: Evidence from electromagnetic articulography study," in *Proc. 35th Berkeley Linguistics Soc.*, 2009, pp. 329–340.

[33] S. L. Li *et al.*, "Hanyu biaozhun shiyuzheng jianchafa de bianzhi yu changmo [The establishment and normalization of the Chinese standard for aphasia examination]," *Zhongguo Kangfu Lilun Yu Shijian*, vol. 6, no. 4, pp. 162–164, Sep. 2000.

[34] Q. S. Zhang *et al.*, "Zhongguo kangfu yanjiu Zhongxin hanyu biaozhun shiyuzheng jiancha liangbiao de xindu yu xiaodu fenxi [The Reliability and validity analysis for the Chinese standard for aphasia scale by China Rehabilitation Research Center]," *Zhongguo Kangfu Lilun Yu Shijian*, vol. 11, no. 9, pp. 703–705, Nov. 2005.

[35] S. S. Mahmoud, Z. M. Hussain, I. Cosic, and Q. Fang, "Time-frequency analysis of normal and abnormal biological signals," *Biomed. Signal Process. Control*, vol. 1, no. 1, pp. 33–43, 2006.

[36] L. R. Rabiner and R. W. Schafer, *Theory and Application of Digital Speech Processing*. Upper Saddle River, NJ, USA: Pearson, 2010.

[37] M. Slaney, "Auditory toolbox," Interval Research Corporation, Tech. Rep. 1998-010, no. 10, 1998.

[38] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, "Highly accurate mandarin tone classification in the absence of pitch information," in *Proc. Speech Prosody*, May 2014, pp. 673–677.

[39] W. Wang, Y. Yang, X. Wang, W. Wang, and J. Li, "Development of convolutional neural network and its application in image classification: A survey," *Opt. Eng., Rev.*, vol. 58, no. 4, 2019, Art. no. 040901.

[40] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Proc. 8th Int. Conf. Qual. Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[41] G. J. Suresh Prasad Kannojia, "Effects of Varying Resolution on Performance of CNN based Image Classification: An Experimental Study," *Int. J. Comput. Sci. Eng.*, vol. 6, pp. 451–456, 2018.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[43] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 464–472.

[44] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, *arXiv:1412.6980*.

[45] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," presented at the Proc. 4th Int. Conf. Neural Inf. Process. Syst., 1991.

[46] J. Howard, FASTAI. 2018. [Online]. Available: https://github.com/fastai/fastai

[47] Z. Warraich and J. A. Kleim, "Neural plasticity: The biological substrate for neurorehabilitation," *PM R*, vol. 2, no. 12S, pp. S208–S219, 2010.

[48] D. Le, K. Licata, C. Persad, and E. M. Provost, "Automatic assessment of speech intelligibility for individuals with aphasia," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2187–2199, Nov. 2016.

[49] A. Kertesz, "Western aphasia battery-revised(WAB-R): Examiner's manual," Harcort Assessment Incorporation, San Antonio, TX, USA, 2006.

[50] W. Huber, K. Poeck, and L. Springer, Klinik und Rehabilitation der Aphasie: eine Einführung für Therapeuten, Angehörige und Betroffene. Georg Thieme Verlag, 2013. Accessed on: Dec. 15, 2019. [Online]. Available: https://books.google.co.in/books?id=UWA36D20lwkC

[51] T. M. Byun, P. F. Halpin, and D. Szeredi, "Online crowdsourcing for efficient rating of speech: A validation study," *J. Commun. Disorders*, vol. 53, pp. 70–83, Jan./Feb. 2015.

[52] X. H. Cheng, and X. L. Tian, *Xian dai han yu [Modern Chinese]*. Hong Kong, China: Sanlian Shudian (H.K.) Youxian Gongsi, 2005.

[53] Beijing Daxue Zhongwenxi Xiandai Hanyu Jiaoyanshi, *Xian dai han yu [Modern Chinese]*. Beijing, China: Shangwu Yinshuguan, 1993.

[54] K. T. Mengistu and F. Rudzicz, "Comparing humans and automatic speech recognition systems in recognizing dysarthric speech," presented at the Advances Artif. Intell., 2011.

[55] M. B. Mustafa, F. Rosdi, S. S. Salim, and M. U. Mughal, "Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker," *Expert Syst. Appl.*, vol. 42, no. 8, pp. 3924–3932, May 15, 2015.

[56] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, Portland, OR, USA, 2012, pp. 1776–1779.

[57] S. J. Fan, "Jiyu yinwei duibi de shiyuzheng yuyin ganzhi yu chansheng tezheng jiqi ganyu yanjiu [Research on the characteristics of speech perception and production of aphasia based on phoneme contrast and its intervention]," Master's thesis, East China Normal University, Shanghai, China, Discipline Educ., 2016. [Online]. Available: CNKI

[58] L. H. Tan *et al.*, "Brain activation in the processing of Chinese characters and words: A functional MRI study," *Article*, vol. 10, no. 1, pp. 16–27, 2000.

[59] L. M. Balsamo *et al.*, "A functional magnetic resonance imaging study of left hemisphere language dominance in children," *Archives Neurol., Article*, vol. 59, no. 7, pp. 1168–1174, 2002.

[60] J. Y. Huang, L. Qin, and Y. F. Li, "Youce danao banqiu yuyan gongneng [The language function of right hemisphere]," *Zhongguo Shenjing Mianyixue He Shenjingbingxue Zazhi*, vol. 14, no. 5, pp. 310–312, May 2007.