Received April 12, 2020, accepted May 3, 2020, date of publication May 12, 2020, date of current version May 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994069

The Application Analysis of Neural Network Techniques on Lexical Tone Rehabilitation of Mandarin-Speaking Patients With Post-Stroke Dysarthria

ZHIWEI MOU[1], WUJIAN YE[2,*], CHIN-CHEN CHANG[3,5], AND YITAO MAO[4]

[1]Department of Rehabilitation, The First Affiliated Hospital of Jinan University, Guangzhou 510630, China
[2]School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China
[3]Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan
[4]Department of Radiology, Xiangya Hospital, Central South University, Changsha 410008, China
[5]School of Computer Science and Technology, Hangzho Dianzi University, Hangzhou 310018, China

Corresponding author: Yitao Mao (maoyt@csu.edu.cn)

*Wujian Ye is co-first author.

This work was supported in part by the National Natural Science Foundation of China under Grant 81701847, in part by the Natural Science Foundation of Hunan Province of China under Grant 2017JJ3497, in part by the Guangdong Medical Science and Technology Research Foundation of China under Grant A2019120, A2020362, in part by the Special Project of Chinese Government for Science and Technology of Guangdong Province under Grant 2019SKJ003, in part by the Guangdong Education Department, Guangdong University of Technology under Grant 220413548, and in part by the Science and Technology Plan Project of Guangzhou City under Grant 201904010371.

ABSTRACT The Objectives of this study are (1) to evaluate tone production in Mandarin-speaking patients with post-stroke dysarthria (PSD) using an artificial neural network (ANN), (2) to investigate the efficacy of recognition performance of the ANN model contrast to the human listeners and the convolutional neural network (CNN) model, and (3) to explore rehabilitation application of the artificial intelligence recognition for lexical tone production disorder with PSD. The subjects include two groups of native Mandarin speaking adults: 31 patients with PSD and 42 normal-speaking adults (NA) in a similar age range as controls. Each subject was recorded producing a list of 7 Mandarin monosyllables with 4 tones (i.e., a total of 28 tokens). The fundamental frequency (F0) of each monosyllable was extracted using auto-correlation algorithm. The ANN was trained with F0 data of the tone tokens from the NA, to generate the final model. The recognition rates of the human ears, ANN model, and CNN model were 87.78% ± 8.96% (mean ± SD), 89.11% ±11.80%, 65.91% ± 8.79% respectively for tone production of NA group; 70.28% ± 17.61%, 63.35% ± 17.40%, 34.71% ± 6.92% respectively for tone production of PSD group. For PSD group, there was significant correlation between the performance of the ANN model and human listeners ($r = 0.826$, $P < 0.001$). However, the performance of CNN model was not correlated with that of the human ears ($r = -0.108$, $P = 0.562$). Thus, the experiments show that ANN is more objective and efficient, which could replace human listeners in the assessment of lexical tone production disorder in Mandarin-speaking patients with PSD. Furthermore, using ANN may reduce the heterogeneity of rehabilitation evaluation among different speech therapists and may give the feedback for achievement of rehabilitation treatment more accurately.

INDEX TERMS Lexical tone rehabilitation, post-stroke dysarthria (PSD), ANN, human listeners, application analysis.

I. INTRODUCTION

Mandarin Chinese, which is spoken by the largest population in the world, is a tonal language different from English or other alphabetic languages. Mandarin tones convey lexical meanings based on the pitch variation patterns, which means one syllable could have different meanings when it has been spoken out with different tones. There are four tones in Mandarin, which are Tone 1, 2, 3, and 4. Mandarin tone patterns are determined by the fundamental frequency (F0) variation of a syllable. Tone 1 has a flat and high F0 contour. Tone 2 has a rising F0 contour. The F0 contour of Tone 3 is like V-shape, falling at the beginning then followed by a rising, with a dip in the middle. Tone 4 has a high

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Elhoseny.

falling F0 contour [1]. In Mandarin Chinese, syllable-level F0 contour is critical for tone recognition. Syllable duration and intensity may carry limited identifiable information of tone in spontaneous speech [2]. However, the duration or intensity in speech is unstable and might be easily affected by a bunch of factors [2], [3]. Comparing to the duration and intensity, F0 is regarded as a relatively stable acoustic feature. Therefore, F0 contour is used in the present study to recognize tone patterns.

Dysarthria is a common type of speech disorders. In clinics, this pathological status is most commonly seen in the population with a post-stroke status. The injured cortex causes dysfunction of muscles which control the articulation movement [4], [5], and the abnormal articulation movement usually results in irregular phonation and deviated amplitude [6]. Therefore, dysarthria leads to compromised speech signals and reduced intelligibility of speech [7], [8]. Although a variety of acoustic studies have been conducted to capture vowel or consonant production deficits in dysarthria for a few languages, such as English, French, German, Swedish and Japanese [9]–[12], there is limited evidence to elucidate the possible pronunciation deficits for Mandarin Chinese tone in patients with post-stroke dysarthria (PSD).

A previous research [13] in dysarthria patients with cerebral palsy (CP) who were native Cantonese (a tonal language primarily spoken in Southern China) speakers showed abnormal F0 patterns in them, including (1) excessive variability, (2) excessively falling frequencies, (3) lowering of the high-level tone, (4) rising tones, and (5) abnormal contour patterns. The results of the acoustic analysis supported previous findings of perceptual difficulty in tone level contrasts for Cantonese speakers with dysarthria. There were evidences showing that Cantonese-speaking dysarthria patients with CP had tone production errors either by human listener evaluation or by acoustic analysis of F0 contours [13]–[15]. Thus these evidences indicated compromised intelligibility for those patients. According to literature, there were little studies focusing on the Mandarin Chinese tone production of dysarthria patients. In the present study, therefore, the Mandarin tone production characteristics have been investigated in native Mandarin-speaking patients with PSD.

Although little attention was paid to the tone pronunciation of PSD patients, there were studies [16]–[20] which explored the tonal deficits in another special population, i.e., the pediatric cochlear implant (CI) users. The results showed that a majority of these children did not produce Mandarin tones very well, and there were evidences indicating that those CI children who spoke tonal language had remarkable deficits in tone perception as well [21]–[24]. The lack of F0 information in the CI devices might be responsible for the perception difficulty [25]. It was inferred that tone production performance in prelingually deafened CI users was dependent on accurate tone perception [19]. However, the auditory perception deficits of PSD patients are not comparable with those of CI users, as well as other factors such as the maturity degree, the type of pathology, and the degree of language acquisition making it dubious for the tone production ability of PSD patients.

Artificial neural network (ANN) is an assembling of mutually-connected artificial neurons with special mathematical relationships among themselves. The structure and function of ANN are somewhat similar to neural networks in biological brain. In most cases, ANN is a self-adaptive system which changes its parameters based on internal and external information that flows through the network. More practically, ANN could be regarded as a nonlinear statistical data processing and modeling tool. It was commonly used to mimic complex relationships between inputs and outputs or to find invisible patterns in data. ANNs were broadly used in the recognition of image or voice for the past decades [20], [26], [27]. Several previous studies also showed that ANNs were able to classify tones well in Mandarin Chinese [28]–[31]. Convolutional neural network (CNN) with machine learning algorithm was also adopted in recent years to classify Mandarin tones [32]–[34]. These newly developed neural networks performed even better comparing to the conventional ANN, with the reported accuracies of around 95% correctness. However, based on the current evidence [30], [31], the accuracy of normal speakers' tone production evaluated by human listeners could seldom achieve as high as 95% correctness, being more comparable with the performance of traditional ANNs which utilize F0 information of the tones.

In summary, there was few literatures which explored the Mandarin tone pronunciation of dysarthria to the best of our knowledge. Yet there has been no relevant study focusing on the Mandarin tone production of dysarthria with post-stoke status up to now. Given that human listeners' assessment is an important reference and since previous evidences showed that ANN's performance was strongly correlated with the performance of human ears [19]. In the present study, we hereby chose to use ANN to recognize produced tones of the Mandarin-speaking PSD patients. Normal adults (NA) were also recruited as control. We also compared the ANN's performance of tone recognition with which was evaluated by normal human listeners and CNN model, and did correlational analyses between these artificial intelligences and the human ears for PSD patients.

The aim of the present study was to explore the efficacy of ANN in assessing tone production of PSD patients and the feasibility of its potential application in speech rehabilitation assessment process for these patients. Our work not only covered the shortage in this field, but also validated a previously proposed ANN method on the tone production assessment in this post-stroke dysarthria population.

II. METHOD

Fig. 1 shows a sketched flux diagram which clarifies the whole methodology in the current study. The process of our study consists of three parts, including participants recruiting, speech materials and data collection, and acoustic analysis

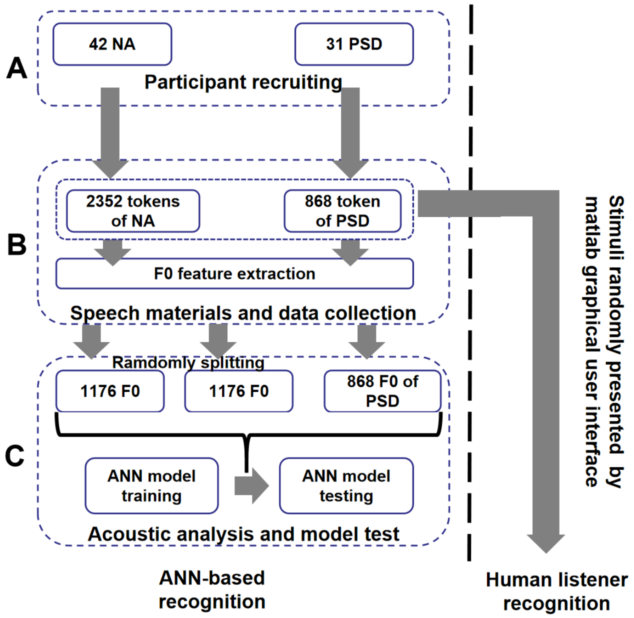


FIGURE 1. A flux diagram of the whole methodology in the study.

and tones production test. All these three parts are detailed in the following sections.

A. PARTICIPANTS RECRUITING

The subjects included 31 native Mandarin-speaking patients with post-stroke dysarthria and 42 normal adults in a similar age range. The participants resided and were recruited in Guangdong province, China, for speech material recording. The ages of the native Mandarin-speaking normal adult group ranged from 21 to 76 years old (mean ± SD: 45.88 ± 13.24 years), including 20 males and 22 females; the ages of PSD group ranged from 25 to 83 years old (mean ± SD: 56.74 ± 16.40 years), including 19 males and 12 females. All of the participants had a pure-tone average threshold at three frequencies (500, 1000, and 2000 Hz) of $<$ 25 dB HL in at least one ear. The PSD speakers met a set of four selection criteria. (1) All patients had a primary speech diagnosis of post-stroke dysarthria by scale evaluation, auxiliary examinations (such as brain CT, brain MRI, laryngoscope) and specialist diagnosis. (2) The current subjects were able to communicate fluently in Mandarin before the disorder. (3) The subjects had no alexia, visual impairment, or severe auditory comprehension impairment. (4) The subjects had no difficulty to articulate, such as only saying '/a/' or '/y/'.

B. SPEECH MATERIALS AND DATA COLLECTION

The monosyllables used to elicit production of 1-4 tones were the following: /ba/, /bi/, /bo/, /du/, /ge/, /yu/, and /a/. For PSD group, the recording was implemented once, and for NA group, the recording process was repeated one more time on another day. Therefore, a total of 2352 tokens (7 monosyllabic words × 4 tones × 42 speakers × 2 recordings) for NA group and 868 tokens (7 monosyllabic words × 4 tones × 31 speakers) for PSD group were obtained. All the produced tokens from both NA and PSD group were digitally recorded and saved to a computer hard disk as audio formats (.wav) at a sampling rate of 44.1 kHz and a resolution of 16-bit. The recording process was conducted in a soundproof room ($<$ 35 dB SPL). One half of randomly selected samples from NA group were used to train the ANN, and the other half from NA as well as all samples from PSD group were used to test the developed ANN. All of the recorded samples from both groups were judged by normal human listeners.

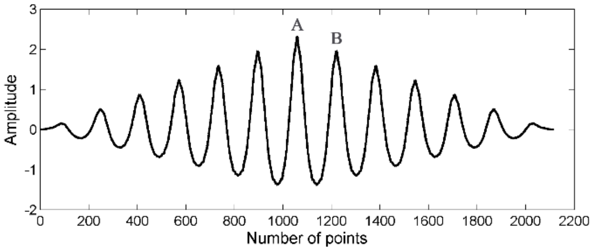


FIGURE 2. A typical auto-correlation sequence for a signal frame of the produced token "du" with Tone 2 by a normal adult. Point A is the highest peak and point B is the second highest peak. The distance (in terms of time unit) between these two peaks is the basic period of the signal frame. And its reciprocal is the F0 of this signal frame.

C. ACOUSTIC ANALYSIS AND TONES PRODUCTION TEST

1) F0 FEATURE EXTRACTION

The F0 contours of the vowels were extracted from all the recorded tokens produced by both groups using an auto-correlation method. The F0 extraction process went as follows: First, the token signal was read in by MATLAB and the signal duration was calculated according to the length of the signal and the sample frequency (duration equals to length of the signal divided by sample frequency). Second, the signal was segmented to a bunch of frames with a time window length of 24 ms and an overlap rate of 2/3. Third, the MATLAB built-in function xcorr was used to obtain the auto-correlation sequence of each of the signal frames. Fig. 2 shows a typical auto-correlation sequence for a frame of the produced tone signal by a normal adult. In Fig. 2, the waveform (i.e., the auto-correlation sequence) is symmetrical and periodic. Peak A and peak B are the highest and the second highest peak, respectively. The distance (in terms of time unit) between these two peaks was calculated by dividing the total number of points between them by sample frequency. This time distance was the basic period of the signal frame. Therefore, the F0 of this signal frame was obtained using Equation (1):

$$F0 = \frac{1}{(Bx - Ax)/fs} \tag{1}$$

Here, Bx represents the x-coordinate of peak B, Ax represents the x-coordinate of peak A, fs stands for sample frequency.

The F0 contour was then obtained by concatenation of the F0s of each signal frame [20], [30], [31],. The extracted F0 contour may occasionally have doubling or halving errors, which were then corrected manually on the narrow band spectrograms of the syllables, as described in the previous studies [20], [30], [31].

2) ARCHITECTURE OF ANN

The ANN used to classify the produced tones was a feed-forward backpropagation multilayer perceptron (MLP) [30], [31], provided by the Neural Network Toolbox in MATLAB. The following is Formula (2) of *newff* function implementing the MLP network [35]:

$$Net = newff\,(PR, [S1, \ldots, Sn], \{TF1, \ldots, TFn\}, BTF, BLF, PF) \tag{2}$$

where PR is an R × 2 matrix with R rows input, and the minimum and maximum values in each input vectors constitute the 1st and 2nd column of the matrix. Si is the vector length of the ith layer with a total of n layers; TFi is the transfer function of the ith layer; BTF is the back-propagation network training function; BLF is the back-propagation learning function of weight/bias; and PF is the performance function.

Here, based on Zhou's previous study on the influence of varied characteristics of the ANN on tone recognition performance [31], we selected a combination of numbers of input neurons and hidden neurons which would guarantee a plateau performance. Xu's study showed that when the number of hidden neurons $>$ 6 and the number of input neurons $>$ 4, the recognition performance reached plateau. And occasionally, there was a handful of token samples which had an F0 contour no more than 8 points in our F0 dataset. If the number of inputs was set to be greater than 8, these short token samples would need to be excluded from the test. Therefore, we chose 7 as the number of hidden neurons and 8 as the number of input neurons in the current study. The MLP was then structured to have 8 input neurons, 7 hidden neurons, and 4 output neurons (represented 1-4 tones, respectively). The architecture of the MLP model was briefly sketched in Fig. 3. The inputs of the ANN were averaged F0 values of the eight evenly spaced segments from the F0 contour. The training procedure of the network was set to be terminated when the number of training iterations reached 200 or the sum of squared errors (SSE) became less than 0.01. The training iterations' number of 200 and SSE of 0.01were used in previous studies focusing on the similar tone recognition task and were justified as appropriate [29]–[31]. Then, one half of the tone tokens which were randomly selected from the whole F0 dataset of NA group (i.e., 1176 tone tokens) was used to train the ANN, the other half from NA group and the whole dataset from PSD group were used to test it. The procedure was repeated 10 times, and the averaged percent-correct scores were obtained for both NA group and PSD group.

Additionally, since Mel-frequency cepstral coefficients (MFCC) are used to model the way that the human ear perceives sound, a CNN based model with MFCC as input feature was also constructed (the technical details were described in Chen's study [33]), in order to make comparison between ANN and CNN in this tone classification task.

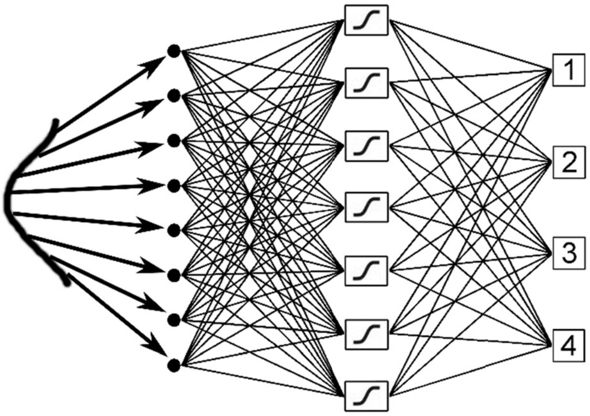


FIGURE 3. The ANN-based architecture model. In this example, a F0 contour of Tone 3 of a Mandarin Chinese word was divided into 8 evenly spaced segments. The average frequency values of each of the segments were used as inputs to the neural network. There were 7 hidden neurons each with a nonlinear transfer function. The four output neurons represented 1-4 tones, respectively.

3) PERCEPTUAL ANALYSIS

The human evaluator for the tone intelligibility evaluation included 5 females and 5 males. We adopted the following inclusion criteria for the human listeners: they need to (1) be between 18 and 40 years old, (2) be native speakers of Mandarin Chinese, (3) have a pure-tone average threshold at four frequencies (500, 1000, 3000 and 4000 Hz) of $<$ 20 dB HL in both ears, (4) had no more than incidental experience in listening to dysarthric speech (The ultimate goal of rehabilitation for the PSD patients is to communicate vocally with ordinary people, that's the reason why it was required to have no experience in listening to dysarthric speech), and (5) fully understand our assessment requirements. The listeners were asked to confirm which tone they heard from the four possible choices in a 4-alternative forced-choice paradigm. The stimuli consisted of all the 3220 tokens (2352 from NA group and 868 from PSD group). The stimuli were presented in a totally randomized sequence based on a MATLAB graphical user interface (Fig. 4) and at a comfortable loudness level via a circumaural headphone (Sennheiser, HD 265). This tone perception test was administrated in a soundproof room ($<$ 35 dB SPL).

After all the tones were presented and the responses of the listeners were collected, the total averaged tone recognition scores for all the 42 NA and the 31 PSD subjects were obtained. Then, the results were compared to those obtained by the ANN-based model.

III. RESULTS

A. THE GENERAL RECOGNITION PERFORMANCES

Tone intelligibility scores of the NA (n = 42) and PSD groups (n = 31) were shown in Fig. 5. The result of human listeners

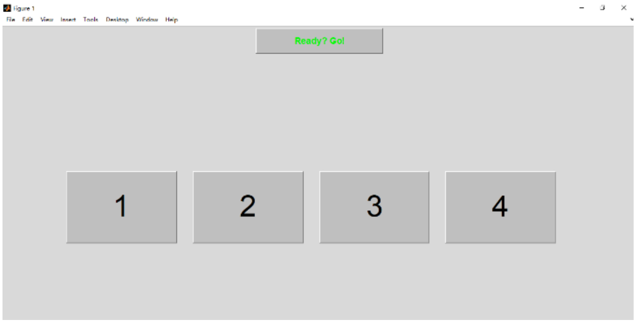
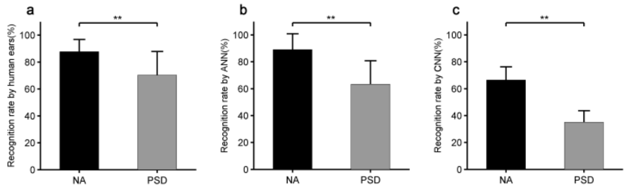


FIGURE 4. The Matlab based graphical user interface for evaluating the tone category (The top button is the start button, and the middle four buttons are the confirmation buttons of Tone 1 to Tone 4, respectively).



(a) human ears (b) ANN-based model (c) CNN-based model

FIGURE 5. The tone recognition rates for the NA and PSD groups. a, b, and c represents the recognition rates judged by human ears, ANN, and CNN, respectively (from left to right). The heights of each bar represent the recognition rates and the error bars represent the SDs. ** means a statistical significance of $P < 0.01$.

was presented in Fig. 5(a). The ten human evaluators assessed the 2352 tokens from NA group and 868 tokens from PSD group. There was good consistency among the human evaluators for the tone classification task. Cronbach's Alpha is 0.916. The judgment of the human ears revealed a recognition rate for NA group of 87.78% ± 8.96% (mean ± SD), ranging from 84.52% to 96.43% correctness. While the recognition rate of PSD group was only 70.28% ± 17.61%, ranging from 39.29% to 98.81% correctness. There was significant difference between these two groups with a two-sample t test ($t = 9.61$, $P < 0.01$).

Fig. 5(b) showed the tone recognition accuracies for both NA and PSD groups judged by ANN. The recognition rate was 89.11% ± 11.80% for NA group vs. 63.35% ± 17.40% for PSD group. The range for NA group was from 73.81% to 98.08% correctness, and for PSD group, was from 37.04% to 96.43% correctness. Again, the group difference was statistically significant ($t = 7.36$, $P < 0.01$). While for Fig. 5(c), it showed the tone recognition accuracies for both NA and PSD groups judged by CNN. The recognition rate was 65.91% ± 8.79% for NA group vs. 34.71% ± 6.92% for PSD group. The range for NA group was from 52.80% to 86.57% correctness, and for PSD group, was from 20.88% to 46.18% correctness. Again, the group difference was statistically significant ($t = 9.57$, $P < 0.01$). And the performances of CNN for both NA and PSD group were significantly lower than those of ANN ($t = 10.14$, $P < 0.01$ for NA group; $t = 13.27$, $P < 0.01$ for PSD group).

It was also indicated from the results that the variability of tone production intelligibility of PSD patients, either judged by ANN or by human ears, was visibly larger than that of the normal adults, but the case was not the same when judged by CNN.

B. THE ANALYSIS OF TONE ERROR

To explore the tone error patterns, the tone confusion matrices were obtained based on the recognition data of ANN for both NA group and PSD group. As shown in Fig. 6. The longitudinal axis represents the target tone (the tone being supposed to be produced), the horizontal axis stands for the judged tone of the ANN.

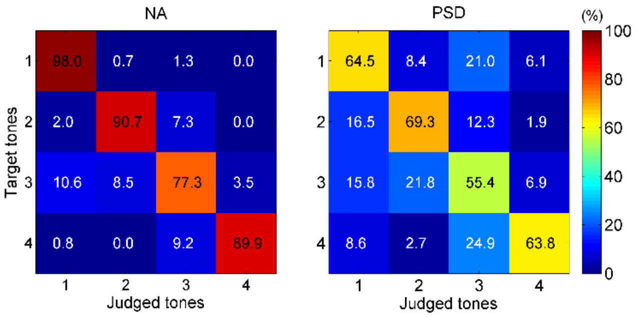


FIGURE 6. Tone confusion matrices judged by ANN for NA group (left panel) and PSD group (right panel). The rows represent the target tones and the columns represent the judged tones. The value in each cell represents the probability of a target tone being recognized as Tone 1, 2, 3, or 4, respectively. For example, in NA group, if the target tone was Tone 1, then this target tone would be recognized as Tone 1 with a probability of 98.0%, as Tone 2 with 0.7%, as Tone 3 with 1.3%, and as Tone 4 with 0.0%. The colors of the cells also reflect the values, with the scaled color bar shown on the right.

The results showed that Tone 1 in NA group reached an accuracy of 98%, while Tone 3 had an accuracy of only 77.3%, implying that Tone 3 was the most unstable tone and most likely to be pronounced mistakenly. We also found that Tone 3 was most commonly pronounced as Tone 1, other than itself. For PSD patients, similarly, Tone 3 had the lowest production accuracy, which was only 55.4% correct. However, Tone 1 had the most dramatic decrease in recognition rate among the four tones in PSD group comparing to NA group. The decreased percentage point for Tone 1 was 33.5% correct, followed by Tone 4 of 26.1% correctness. We also observed that Tone 1 and Tone 4 in PSD group were most likely confused with Tone 3. Based on the results above, the tone production error patterns were not consistent between these two groups.

C. THE CORRELATION BETWEEN HUMAN EAR JUDGEMENT AND ANN- OR CNN-BASED RECOGNITION

For the purpose of validating the accuracy of ANN or CNN, or in other words, to explore whether a patient with high production recognition by human listeners would be judged as high recognition rate by ANN or CNN either, correlation analysis was implemented. Fig. 7 shows the scatter plot of a Pearson's correlation between the recognition rates of ANN or CNN and human listeners for PSD patients. There was a statistically significantly positive correlation between the

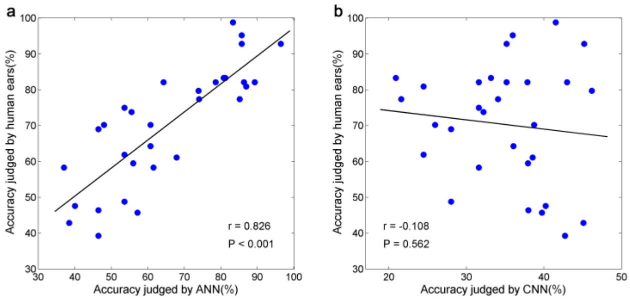


FIGURE 7. The scatter plot of the PSD patients' tone production correct scores (%). a. ANN vs. human listeners; b. CNN vs. Human listeners. Each circle symbol represents the data of each of the thirty one PSD patients. The solid line represents the linear fit of the data, with the correlation coefficient *r* and the *P* value displayed on the bottom corner.

two variables for ANN, with the correlation coefficient r of 0.826, and $P < 0.001$. However, the CNN's recognition rate did not show significant correlation with that of the human ears ($r = -0.108$, and $P = 0.562$). Therefore, the judgement of ANN was regarded as consistent with the judgement of the human ears.

IV. DISCUSSION

Using ANN in the automatic recognition of Mandarin tones is not new though, a series of works have been done about the ANNs' application in automatically recognizing Mandarin tones [20], [29]–[31]. However, besides the normal controls, these studies were mainly focused on the pediatric cochlear implant users. The ANN's recognition performances for normal control were in the range from 85% to around 90% correctness in those studies [20], [29]–[31]. Those recognition accuracies were comparable with the performance of ANN in the current study for NA group. The error pattern in tone production for normal controls was also similar with that displayed in the present study (Fig.6, the confusion matrix). We noted that the PSD patients' tone production accuracy was higher, as a whole, than the pediatric implant users in those previous studies, either being judged by ANN (63.4% vs. 42.3%; 63.4% vs. 58.8%) or by human listeners (70.3% vs. 46.8%) [19], [20]. Given that these two populations were totally different in a variety of aspects, such as age, hearing status, pathological status, and so on, there was no need to discuss the underlying reasons of the disparate production performances of these two populations. Few other researches focused on the automatic recognition of Mandarin tones. We noted a previous study which related to Cantonese (a dialect mostly used in the southern part of China) recognition, using a so-called supratone model [36]. That model had to be used in continuous Cantonese speech and achieved a recognition accuracy of 74.7%. While the focus of the present study was individual tone production, that model was not suitable in our case. More recently, newly developed artificial intelligence methods were used to classify Mandarin tones [32], [34]. Chen *et al.* used a CNN model based on MFCC to classify tones that produced by normal-hearing children. The classification accuracy rate achieved as high as 95.5% [33]. In their further study using spectrograms instead of MFCC as the inputs, the CNN model performed even better than the one based on MFCC [32]. These methods circumvented the manual correction of F0s and could be more efficient than conventional ANNs which based on F0 information. We have trained a simple CNN model in the present study based on MFCC feature and the performance was not satisfactory. It was worth to note that the speech material used in Chen's study [33] was from children from 3 to 10 years old, while ours was from adults with a much larger age span (from 21 to 76 years old). Most importantly, the training dataset in Chen's study was almost four times as ours (4500 tokens vs. 1176 tokens). We thus inferred that ANN may be more robust with relatively small training dataset when compared with CNN. Additionally, the rehabilitation goal of dysarthria was to recover the vocal communication ability with people, while human listeners' perceptual evaluation was still the important reference for tone production rehabilitation. Our results showed that the recognition performance of CNN model based on MFCC did not present positive correlation with the judgement of human listeners. And there were no evidence so far which showed that the performance of these newly developed neural networks (such as CNN) was consistent with that of human listeners. Therefore, their application on the clinical rehabilitation practice might still need to be justified.

While on the other hand, conventional ANN seems to be more comparable with human listeners. Our results showed a strong positive correlation between the performance of ANN and human ears. In a previous study by Zhou *et al.* [19], the researchers also found that the rated scores judged by ANN and those assessed by human ears were strongly correlated based on their 76 pediatric cochlear implant users, with a correlation coefficient of 0.94. Given the excellent consistency between ANN and human ears, it was inferred that ANN could be used to replace the therapists in clinical practice for the assessment of tone production by pediatric cochlear implant users. Technically, ANN is more objective and efficient than human listeners and its performance was as accurate as human ears, which was the reason that we adopted ANN as the assessment tool to evaluate the tone production of PSD patients in the present study. The input information was F0s of the produced tones by the PSD patients. The extraction of F0s was based on auto-correlation algorithm and involved somewhat manual correction of F0s, thus probably bringing down its working efficiency. Nevertheless, ANN was still supposed to be helpful to reduce the inconsistency among different human evaluators or rehabilitation institutions and shorten the assessment duration in tone production evaluation.

Our results revealed that PSD patients' recognition scores were lower than the scores obtained from NA controls, regardless of either by ANN or by human-ear judgement. This is not out of our expectancy since dysarthria with post-stroke status may impair the physiological function of articulation-related organs, resulting in shitty tone productions. We also observed that for NA group, Tone 1 and 2 were the most easily recognized tones, followed by Tone 4, and Tone 3 was the worst. It is reasonable given that Tone 3 has a more complicated F0 structure (i.e., a decline at the beginning followed by a rising up) than the other tones, making it the most difficult tone to produce. While for the PSD patient group, all of the recognition scores of the four tones declined when compared with the NA group, however, the score of Tone 1 decreased more significantly than the other three tones. Considering the F0 contours of Tone 1 in PSD group varied dramatically for a majority of these subjects, it could be inferred that PSD patients may not be able to articulate a syllable stably and persistently, which could be attributed to the discoordination of the articulation organs being out of the superior control. Tone 2 had a relatively higher recognition score than the other three tones in PSD patients. As for Tone 2, the F0 curve has a low-rising pattern with a relatively low frequency onset, which makes it to be produced out easily since strained vocal cords are not required at the beginning of its production. For this case, it still could obtain satisfactory recognition for Tone 2 even with relatively shorter duration. It could be inferred from the results of the present study that the rehabilitation priority of Mandarin tones for PSD patients should be Tone 2, followed by Tone 1, then Tone 4, and finally Tone 3. This sequence was differed from a previous study focused on the Mandarin tone recognition ability of pediatric implant users by Mao *et al.* [20], in which the most easily recognized tone was Tone 1, followed by Tone 4, then Tone 2, and finally Tone 3. This discrepancy could be accounted by the potentially different characteristics and mechanisms of tonal disability with different pathologies.

Furthermore, our research may provide an aided diagnosis and guidance for clinical speech treatment to correct the tone errors in the PSD patients, based on objective acoustic features. First of all, we can use ANN to present the tone production accuracy for each patient. If the accuracy rate is high enough, then the patient could be exempted for further interventions. Second, if a patient needs further rehabilitation interventions, the tone confusion matrix would be helpful in the decision-making of rehabilitation strategies. In addition, the tone error pattern in the confusion matrix could reflect potential site of dysfunction in the articulation organs. For example, if the produced Tone 1 of a patient is most likely recognized as Tone 3, it could be inferred that the dysfunction site is more likely the muscles which control the motion of vocal cord or the relevant neural control, rather than the vocal cord itself. While if the produced Tone 1 is most likely recognized as Tone 4, then there could be an uncoordinated respiration motion for that patient either. Our further studies will focus on the impact of various pathologies on the tone articulation and the potential mechanisms.

Nowadays, China is entering the era of aging society, and the occurrence of stroke is increasing in the middle-aged population. With the expanding population of PSD, there is an urgent need for this population to evaluate the pattern and the severity of tone production disability. An objective, accurate, yet efficient assessment method could help therapists to formulate rehabilitation schemes so as to benefit the recovering of normal communication ability for PSD patients. Our results indicate that ANN is such an evaluation tool and has a promising application value in the clinical rehabilitation practice. However, our study still has some limitations which need to be solved in the future.

1) The performances of different ANN architectures applied on our task need to be researched.
2) Exploring other sound/acoustic characteristics (such as vowels, consonants, compound vowels, duration, prosody, etc.) of Mandarin-speaking PSD population is the planned research work.
3) This method involves multidisciplinary coordination, which includes acoustics, phonetics, computer science, rehabilitation medicine, therapeutics, and so on. The methodology and technique still need to be further improved.
4) Although the CNN did not show a promising result in the present study, our result does not exclude the possibility that CNN-based model could improve considerably when dealing with similar classification tasks by selecting other features, modulating its structure or parameters, modifying some preprocessing, or polishing its inner algorithm. This is another whole different research field which needs comprehensive explorations in the future.

ACKNOWLEDGMENT

(Zhiwei Mou and Wujian Ye contributed equally to this work.). Specifically, Zhiwei Mou was responsible for the research designing, Yitao Mao and Zhiwei Mou carried out the experiments and manuscript writing, Wujian Ye performed the data analysis and software implementation, Chin-chen Chang gave theoretical guidance, writing and reviewing. It has been approved by all the authors for either the content of the manuscript or the order of authorship list. At last, we are also grateful to Li Xu for his technical support.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] D. H. Whalen and Y. Xu, "Information for mandarin tones in the amplitude contour and in brief segments," *Phonetica*, vol. 49, no. 1, pp. 25–47, 1992.

[2] Q.-J. Fu and F.-G. Zeng, "Identification of temporal envelope cues in Chinese tone recognition," *Asia Pacific J. Speech, Lang. Hearing*, vol. 5, no. 1, pp. 45–57, Mar. 2000.

[3] L. Xu, Y. Tsai, and B. E. Pfingst, "Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses," *J. Acoust. Soc. Amer.*, vol. 112, no. 1, pp. 247–258, Jul. 2002.

[4] J. R. Duffy, "Motor speech disorders: Clues to neurologic diagnosis," in *Parkinson's Disease and Movement Disorders*. Totowa, NJ, USA: Humana Press, 2000, pp. 35–53.

[5] P. Enderby, "Disorders of communication: Dysarthria," in *Handbook of Clinical Neurology*. Amsterdam, The Netherlands: Elsevier, 2013, pp. 273–281.

[6] H. Tolba and A. S. El_Torgoman, "Towards the improvement of automatic recognition of dysarthric speech," in *Proc. 2nd IEEE Int. Conf. Comput. Sci. Inf. Technol.*, vol. 8, Aug. 2009, pp. 277–281.

[7] K. C. Hustad, "Estimating the intelligibility of speakers with dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 3, pp. 217–228, 2006.

[8] S. Landa, L. Pennington, N. Miller, S. Robson, V. Thompson, and N. Steen, "Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding," *Int. J. Speech-Language Pathol.*, vol. 16, no. 4, pp. 408–416, Aug. 2014.

[9] W. Ziegler and D. Von Cramon, "Vowel distortion in traumatic dysarthria: A formant study," *Phonetica*, vol. 40, no. 1, pp. 63–78, 1983.

[10] S. Watanabe, K. Arasaki, H. Nagata, and S. Shouji, "Analysis of dysarthria in amyotrophic lateral sclerosis–MRI of the tongue and formant analysis of vowels," *Rinsho shinkeigaku = Clin. Neurol.*, vol. 34, no. 3, pp. 217–223, 1994.

[11] A. Löfqvist, B. Sahlén, and T. Ibertsson, "Vowel spaces in Swedish adolescents with cochlear implants," *J. Acoust. Soc. Amer.*, vol. 128, no. 5, pp. 3064–3069, 2010.

[12] Y. Maryn, M. de Bodt, B. Barsties, and N. Roy, "The value of the acoustic voice quality index as a measure of dysphonia severity in subjects speaking different languages," *Eur. Arch. Otorhinolaryngol.*, vol. 271, pp. 1609–1619, Oct. 2014.

[13] T. L. Whitehill and V. Ciocca, "Perceptual-phonetic predictors of single-word intelligibility: A study of Cantonese dysarthria," *J. Speech, Lang., Hearing Res.*, vol. 43, no. 6, pp. 1451–1465, Dec. 2000.

[14] T. L. Whitehill and V. Ciocca, "Speech errors in Cantonese speaking adults with cerebral palsy," *Clin. Linguistics Phonetics*, vol. 14, no. 2, pp. 111–130, 2000.

[15] V. Ciocca, T. L. Whitehill, and S. S. Ng, "Contour tone production by Cantonese speakers with cerebral palsy," *J. Med. Speech-Lang. Pathol.*, vol. 10, no. 4, pp. 243–248, 2002.

[16] L. Xu, Y. Li, J. Hao, X. Chen, S. A. Xue, and D. Han, "Tone production in mandarin-speaking children with cochlear implants: A preliminary study," *Acta Oto-Laryngologica*, vol. 124, no. 4, pp. 363–367, May 2004.

[17] D. Han, N. Zhou, Y. Li, X. Chen, X. Zhao, and L. Xu, "Tone production of mandarin Chinese speaking children with cochlear implants," *Int. J. Pediatric Otorhinolaryngol.*, vol. 71, no. 6, pp. 875–880, Jun. 2007.

[18] S.-C. Peng, J. B. Tomblin, H. Cheung, Y.-S. Lin, and L.-S. Wang, "Perception and production of mandarin tones in prelingually deaf children with cochlear implants," *Ear Hearing*, vol. 25, no. 3, pp. 251–264, Jun. 2004.

[19] N. Zhou, J. Huang, X. Chen, and L. Xu, "Relationship between tone perception and production in prelingually-deafened children with cochlear implants," *Otol. Neurotol., Off. Publication Amer. Otological Soc., Amer. Neurotol. Soc. Eur. Acad. Otol. Neurotol.*, vol. 34, no. 3, p. 499, 2013.

[20] Y. T. Mao, Z. M. Chen, and L. Xu, "The application of artificial neural network on the assessment of lexical tone production of pediatric cochlear implant users," *Chin. J. Otorhinolaryngol. Head Neck Surg.*, vol. 52, no. 8, pp. 573–579, 2017.

[21] W. I. Wei, R. Wong, Y. Hui, D. K. Au, B. Y. Wong, W. K. Ho, A. Tsang, P. Kung, and E. Chung, "Chinese tonal language rehabilitation following cochlear implantation in children," *Acta Oto-Laryngol.*, vol. 120, no. 2, pp. 218–221, 2000.

[22] V. Ciocca, A. L. Francis, R. Aisha, and L. Wong, "The perception of Cantonese lexical tones by early-deafened cochlear implantees," *J. Acoust. Soc. Amer.*, vol. 111, no. 5, pp. 2250–2256, 2002.

[23] K. Y. S. Lee, C. A. van Hasselt, S. N. Chiu, and D. M. C. Cheung, "Cantonese tone perception ability of cochlear implant children in comparison with normal-hearing children," *Int. J. Pediatric Otorhinolaryngol.*, vol. 63, no. 2, pp. 137–147, Apr. 2002.

[24] A. O. C. Wong and L. L. N. Wong, "Tone perception of Cantonese-speaking prelingually hearing-impaired children with cochlear implants," *Otolaryngol.–Head Neck Surg.*, vol. 130, no. 6, pp. 751–758, Jun. 2004.

[25] B. C. J. Moore, "Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants," *Otol. Neurotol.*, vol. 24, no. 2, pp. 243–254, 2003.

[26] J. P. Teixeira, P. O. Fernandes, and N. Alves, "Vocal acoustic analysis–classification of dysphonic voices with artificial neural networks," *Procedia Comput. Sci.*, vol. 121, pp. 19–26, Nov. 2017.

[27] P. R. Rajarapollu, D. Adhikari, and N. V. Bansode, "Use of artificial neural network for abnormality detection in medical images," in *Optimization in Machine Learning and Applications*. Singapore: Springer, 2020, pp. 1–12.

[28] N. Lan, K. B. Nie, S. K. Gao, and F. G. Zeng, "A novel speech-processing strategy incorporating tonal information for cochlear implants," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 5, pp. 752–760, May 2004.

[29] X. Li, Z. Wenle, Z. Ning, L. Chaoyang, L. Yongxin, C. Xiuwu, and Z. Xiaoyan, "Mandarin Chinese tone recognition with an artificial neural network," *J. Otol.*, vol. 1, no. 1, pp. 30–34, Jun. 2006.

[30] L. Xu, X. Chen, N. Zhou, Y. Li, X. Zhao, and D. Han, "Recognition of lexical tone production of children with an artificial neural network," *Acta Oto-Laryngologica*, vol. 127, no. 4, pp. 365–369, Jan. 2007.

[31] N. Zhou, W. Zhang, C.-Y. Lee, and L. Xu, "Lexical tone recognition with an artificial neural network," *Ear Hearing*, vol. 29, no. 3, pp. 326–335, Jun. 2008.

[32] C. Chen, R. Bunescu, L. Xu, and C. Liu, "Mandarin tone recognition based on unsupervised feature learning from spectrograms," *J. Acoust. Soc. Amer.*, vol. 140, no. 4, p. 3394, Oct. 2016.

[33] C. Chen, R. Bunescu, L. Xu, and C. Liu, "Tone classification in mandarin Chinese using convolutional neural networks," in *Proc. Interspeech*, Sep. 2016, pp. 2150–2154.

[34] J. Tyler, H. Zou, H. Zhou, H. Su, and J. Braasch, "Automated mandarin tone classification using deep neural networks trained on a large speech dataset," *J. Acoust. Soc. Amer.*, vol. 145, no. 3, p. 1814, Mar. 2019.

[35] *Global Optimization Toolbox User's Guide*, MathWorks, Natick, MA, USA. 2016.

[36] Y. Qian, T. Lee, and F. K. Soong, "Tone recognition in continuous Cantonese speech using supratone models," *J. Acoust. Soc. Amer.*, vol. 121, no. 5, pp. 2936–2945, May 2007.

ZHIWEI MOU was born in Hubei, China, in 1978. He received the M.D. degree in neurology from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in rehabilitation from Jinan University, Guangzhou, China, in 2018. He served as a Consultant with the Department of Rehabilitation, The First Affiliated Hospital of Jinan University. From 2016 to 2018, he has been a Visiting Scholar with the School of Rehabilitation and Communication Sciences, Ohio University. He was the author of seven books and nine invention patents, and over ten articles in pathological speech field. His current research interests include the acoustic analysis, neural networks, and speech treatment.

WUJIAN YE received the B.S. degree in computer science and technology from the School of Computers, Guangdong University of Technology, Guangzhou, China, in 2010, and the M.S. and Ph.D. degrees in computer science from Dankook University, South Korea, in 2012 and 2015, respectively. Since 2016, he has been a Lecturer with the School of Information Engineering, Guangdong University of Technology, China. His research interests include deep learning and computer vision, machine learning application, computer network and security analysis, and voice recognition. He was the author of 16 invention patents, and over 20 articles.

CHIN-CHEN CHANG received the Ph.D. degree in computer science from National Tsing-Hua University, in 1982. Since then, he served as an Associate Professor with National Chiao-Tung University, a Professor with National Chung-Hsing University, the Chair and a Professor of the Computer Science Department, National Chung-Cheng University, the Director of the Automation Research Center, the Dean of the College of Engineering, Provost, and the Acting President of the National Chung-Cheng University. He also served as the Director of Advisory Office, Ministry of Education, Taiwan and was a Visiting Scholar/Researcher with Tokyo University and Kyoto University. He is currently a Chair Professor with Feng-Chia University, an Honorary Professor with National Chung-Cheng University, and holds a joint appointment with National Chiao-Tung University. He has worked on many different topics in information security, cryptography, multimedia image processing and published several hundreds of articles in international conferences and journals and over 30 books. He was cited over 27 668 times and has an h-factor of 80 according to Google Scholar. Several well-known concepts and algorithms were adopted in textbooks.

YITAO MAO received the B.S. degree in medical imaging from Sun Yat-sen University, China, in 2009, and the Ph.D. degree in otorhinolaryngology from Central South University, China, in 2014. He was as a Visiting Scholar with Ohio University, USA, from 2012 to 2015, focusing on the rehabilitation research of cochlear implant users, especially for those Mandarin-speaking users. He has authored over 30 articles published or accepted by refereed international journals [e.g., the J Otol, J Acoust Soc Am, Sci Rep, Int J Audiol, Laryngoscope, Int J Pediatr Otorhinolaryngol, Hum Brain Mapp, and J Cent South Univ (Med Sci)]. His current research interests are mainly in the area of auditory imaging and hearing rehabilitation of various pathologies.