informa
healthcare

ORIGINAL ARTICLE

Recognition of lexical tone production of children with an artificial neural network

LI XU[1], XIUWU CHEN[2], NING ZHOU[1], YONGXIN LI[2], XIAOYAN ZHAO[2] & DEMIN HAN[2]

[1]*School of Hearing, Speech and Language Sciences, Ohio University, Athens, OH, USA and* [2]*Department of Otolaryngology – Head & Neck Surgery, Beijing Institute of Otorhinolaryngology, Beijing Tongren Hospital, Beijing, PR China*

Abstract

Conclusion. This study demonstrated that the artificial neural network can successfully classify Mandarin Chinese tone patterns produced by multiple children. The neural network can be used as an objective way of evaluating tone production of children. *Objectives.* Traditionally, tone production is evaluated subjectively using human listeners. The aim of the present study was to investigate the efficacy of using an artificial neural network in evaluating tone production of Mandarin-speaking children. *Subjects and methods.* Speech materials were recorded from 61 normal-hearing children. The fundamental frequency (F0) of each monosyllabic word was extracted and then used as inputs to a feed-forward backpropagation artificial neural network. The number of inputs was set at 12, whereas the number of hidden neurons was set at 16 in the neural network. The output layer consisted of four neurons representing the four Mandarin tone patterns. The tone recognition performance of the neural network was further compared with that of native Mandarin-speaking adult listeners. *Results.* The neural network successfully classified the tone patterns of the 61 child speakers with an accuracy of about 85% correct. This high accuracy exceeded the tone recognition performance by the adult listeners. Individual child speakers showed varied tone production accuracy as recognized by the adult listeners or by the neural network.

Keywords: *Tone language, tone recognition, Chinese tone patterns, pattern recognition*

Introduction

Tones bear special linguistic functions in tone languages such as Mandarin Chinese. They are used to distinguish meanings of words. Tone patterns of Mandarin Chinese are defined by the fundamental frequency (F0) variation over time of the voiced part of a syllable. These patterns include (1) flat and high, (2) rising, (3) low and dipping, and (4) falling. The spectrograms in Figure 1 show four tones of a syllable produced by a 5-year-old girl. The F0 contours in the spectrograms show the typical four patterns of the tones. Tone acquisition co-develops with other segmentals in tone languages. The timeframe for Mandarin Chinese tone development in children was reported to vary from an early stage of life to more than 3 years of age [1,2]. The advent of multichannel cochlear implants has helped to overcome the obstacles in language development in many children with profound hearing loss. However, due to the lack of pitch information in the electrical stimulation in current cochlear implant technology (see Moore [3] for review), deficit in tone perception [4–7] and production [8,9] in children with a cochlear implant whose native language is a tone language has been documented in previous studies. Children with hearing impairment have various levels of deficit in tone development [10]. The present study aimed to apply an artificial neural network to assess tone production in normal-hearing children, as one step toward the goal of using this neural network to assess tone production in children with cochlear implants.

An artificial neural network is an interconnected group of artificial neurons that uses a mathematical

Correspondence: Li Xu, MD, PhD, School of Hearing, Speech and Language Sciences, Ohio University, Athens, OH 45701, USA. Tel: +1 740 593 0310. Fax: +1 740 593 0287. E-mail: XuL@ohio.edu

(Received 7 September 2006; accepted 14 September 2006)

ISSN 0001-6489 print/ISSN 1651-2551 online © 2007 Taylor & Francis
DOI: 10.1080/00016480601011477

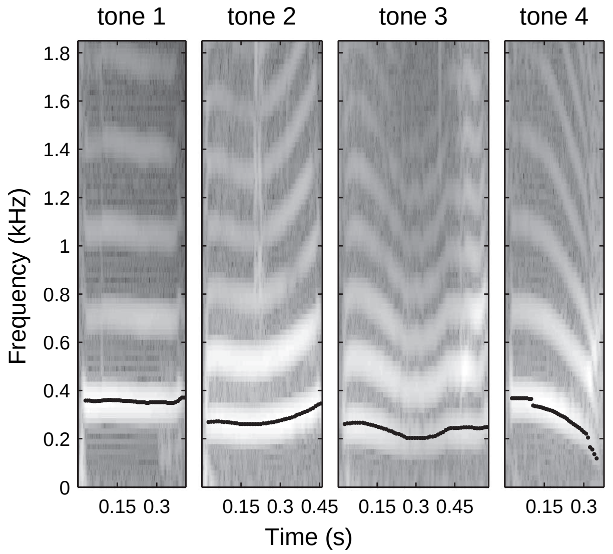

Figure 1. Spectrograms and F0 contours of the four tone patterns of *shi* spoken by a 5- year-old girl (subject F11). The spectrograms are plotted in the narrowband format with the gray scale indicating energy associated with time (abscissa) and frequency (ordinate). The F0 contours extracted with the autocorrelation method are plotted with the black symbols.

or computational model for information processing based on a connectionist approach to computation. In most cases an artificial neural network is an adaptive system that changes its structure based on external or internal information that flows through the network. In more practical terms neural networks are nonlinear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data.

A number of studies have used artificial neural networks to recognize Mandarin Chinese tones with different models of stimulation [11–13]. The most commonly used type was the feed-forward multilayer perceptron, arguably the simplest type of neural network devised. None of the previous studies have used neural networks to test tone recognition of speech produced by children. In the present study, we tested the neural network's sensitivity to speaker variation with multiple children as speakers. We also examined the effects of speaker age and gender on the recognition performance. Finally, we compared tone recognition performance by the neural network with that by a group of human listeners.

Subjects and methods

Sixty-one normal-hearing native Mandarin Chinese-speaking children with ages ranging from 3 to 9 years (mean = 6.2, SD = 1.7) were recruited from kindergartens and elementary schools in Beijing, China for speech material recording. There were 28 boys and 33 girls. The elicited production of Chinese monosyllables was digitally recorded at a sampling rate of 44.1 kHz with a 16-bit resolution in quiet rooms. The monosyllables used to elicit production of the four tones were the following: *ai, bao, bi, can, chi, du, duo, fa, fu, ge, hu, ji, jie, ke, la, ma, na, pao, pi, qi, qie, shi, tu, tuo, wan, wen, wu, xian, xu, ya, yan, yang, yao, yi, ying, you, yu, yuan, zan, zhi.* In total, 9760 tone tokens were recorded (40 syllables × 4 tones × 61 speakers). This database was used in the testing for both the neural network and human listeners. F0 contours that define the tones were extracted for all the tone tokens in the speech database with an autocorrelation method. Errors arising from F0 extraction, usually in forms of frequency doubling or halving, were manually corrected based on the narrowband spectrograms of the syllables as described in Xu et al. [8]. An example of extracted F0 contour is illustrated in Figure 1.

A feed-forward backpropagation multilayer perceptron was implemented in MATLAB with the Neural Network Toolbox (MathWorks, Natick, MA, USA). In our previous studies, the efficacy of a multilayer perceptron was tested [14]. Here, we constructed the architecture of the neural network based on the results of the early work. Figure 2 depicts the architecture of the multilayer perceptron. Each F0 contour was segmented into 12 evenly

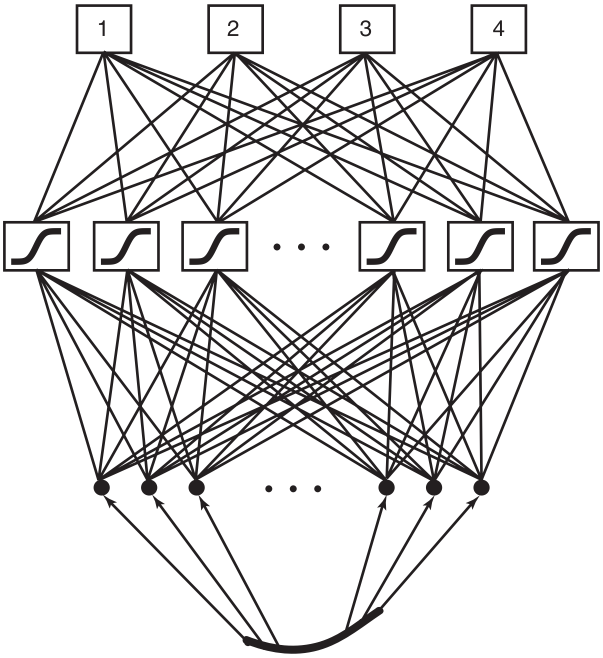

Figure 2. Architecture of the artificial neural network. In this example, a pitch contour of tone 3 of a Mandarin Chinese word was divided into 12 evenly spaced segments. The average frequency values of each of the segments were used as inputs to the neural network. There were 16 hidden neurons each with a nonlinear transfer function. The four output neurons corresponded to tones 1, 2, 3, and 4, respectively.

spaced parts and the 12 mean values of the 12 parts were used as inputs to the neural network. The hidden layer contained 16 nonlinear neurons. The training of the neural network was set to complete when the sum of squared errors became <0.01 or the number of epochs reached 200. For each of the experimental conditions that are detailed below, the neural network was tested 10 times, each time with a different randomization of inputs.

The neural network's sensitivity to speaker variation was examined in two aspects, i.e. the effects of the number of speakers and the effects of speaker gender. The number of speakers used for testing was varied from 1 to 61 in a step size of 2. If the number of speakers is N, N speakers are randomly drawn without replacement from the speech database pool. Half of the tokens from the chosen speakers were used for training and the other half for testing. The effects of gender were examined by comparing the recognition accuracy for tone tokens sorted by gender with that for tone tokens of mixed genders. To obtain the tone recognition accuracy of gender-sorted tone tokens, training and testing were done within either gender group separately. Thus, a recognition score for each gender group was obtained. The tone recognition score using mixed genders was available from the previous test of the effects of number of speakers, where speakers were drawn at random.

A cross-validation approach was used to obtain the neural network's recognition score on each of the 61 speakers. In the cross-validation test, tone tokens from individual speakers were separately recognized after the neural network was trained by the tone tokens from the other 60 speakers. In other words, the neural network was always trained by the tokens from 60 speakers and tested with the tokens from the one speaker whose speech tokens were not included in the training.

Seven normal-hearing adult native Mandarin Chinese speakers were recruited for the tone perception tests. A custom graphical user interface was developed in MATLAB to present the tone tokens from all the 61 child speakers. The tone perception test was done in a double-walled sound-treated booth. The total 9760 speech tokens were randomized and presented to the adult listeners. The listeners were instructed to use a computer mouse to click on a button labeled 1, 2, 3, or 4 to indicate the tone that they had heard. After 9760 presentations, the average tone recognition score for all the 61 children obtained by the adult human listeners was compared to that obtained by the neural network. The tone recognition scores on each of the children by the adult listeners were compared with the results from the cross-validation test of the neural network.

Results

Tone recognition scores by the neural network were averaged across 10 trials. As the number of speakers mixed in the inputs to the neural network increased, the performance decreased from 99.6% correct with one speaker to 85.6% correct with 61 speakers (Figure 3). Tone recognition performance by human listeners was 79.5% correct, which is significantly lower than that by the neural network ($t = 7.53$, $p < 0.05$).

The average recognition score on all the girls ($n = 33$) was 88.3% correct. The corresponding gender-mixed recognition score of the same number of speakers was 86.8% correct. The average recognition score on all the boys ($n = 28$) was 84.9% correct. The corresponding gender-mixed recognition score of the same number of speakers was 86.5% correct (Figure 3). Therefore, the gender-sorted recognition scores differed by only a couple of percentage points from the gender-mixed recognition scores.

The results of the cross-validation test of the neural network provided recognition scores for individual children. The scores by the neural network ranged from 56.2% correct to 98.8% correct (mean = 84.7% correct, SD = 9.1% correct), indicating a fairly high variability of these children's tone

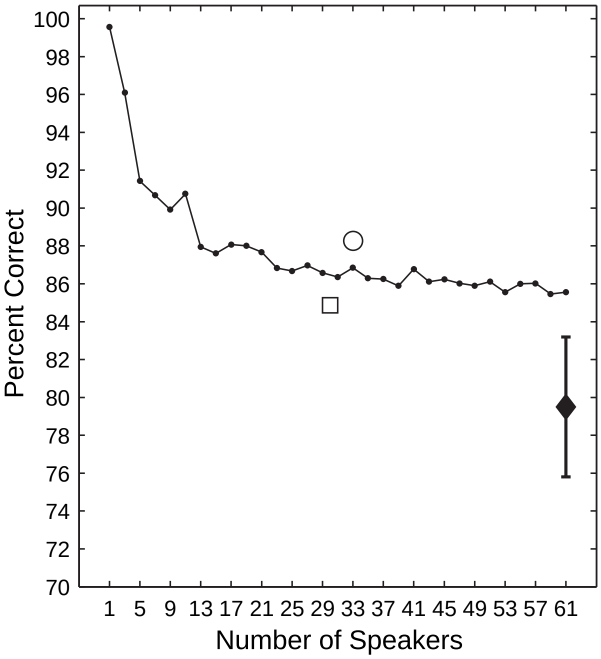

Figure 3. Tone recognition performance by the neural network and human listeners as a function of number of speakers. Mean scores of the artificial neural network are connected with a solid line as the number of speakers varied from 1 to 61 with a step size of 2. The mean scores by the neural network on the 33 girls (circle) and the 28 boys (square) are plotted at the corresponding numbers of speakers. The mean score by the human listeners is plotted with a filled diamond at the number of speakers of 61. Error bars represent standard deviation.

production intelligibility. Similarly, a fairly high variability in the children's tone production intelligibility was revealed in the tone recognition by the adult listeners. The scores by the human listeners ranged from 57.5% correct to 96.9% correct (mean = 79.4% correct, SD = 7.7% correct). A significant positive correlation was found between the recognition scores of individual children by the neural network and those by the human listeners (r = 0.66, z test, $p < 0.05$) (Figure 4).

To examine the potential trend of an increase in tone production intelligibility with the children's age from 3 to 9 years old, the individual recognition scores of all the children were sorted by age group at intervals of 6 months. The age-sorted recognition scores did not demonstrate a trend of improvement in tone intelligibility toward older children. However, we caution that the number of children in each age group was relatively small (i.e. from 1 to 9).

Discussion

As the number of speakers increased in the input, the neural network's performance decreased (Figure 3). A quick decrease in the performance from 99.6% correct with one speaker to 87.9% correct with 13 speakers was observed, followed by a fairly consistent performance of around 86% correct with more speakers. A small number of speakers seemed to have provided adequate sampling in this particular group of children for a consistent classification performance. In a recent study, we tested the effects of number of speakers using speech samples from a group of 29 adult speakers (unpublished observations). We found that the neural network produced a linear decrease in performance from 100% correct with one speaker to 85.3% correct with 29 speakers without obvious saturation. The early saturation in the recognition performance that we observed in the present study may be due to the acoustic homogeneity of the children's tone production. Relatively fewer tone tokens were needed to represent a larger group. The acoustic homogeneity was reflected by the less variation in the acoustic features of the same tone produced by children speakers. One manifestation of this acoustic homogeneity is further discussed below in terms of the lack of effects of speaker gender on recognition accuracy.

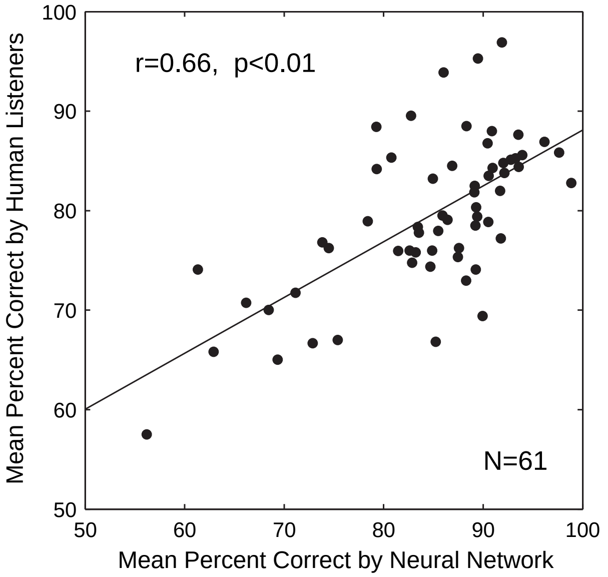

Figure 4. Tone recognition performance by the neural network and human listeners on the individual child speakers. Each symbol represents one of the 61 children. The solid line represents the linear fit of the data.

The effects of gender were examined by comparing the tone recognition performance of the neural network with inputs of speech tokens from mixed gender and that with inputs of one gender at a time (Figure 3). There was a potential difference in F0 ranges between male and female speakers and we were interested in evaluating its impact on the neural network's classification accuracy that relies on F0 values (height). We hypothesized that if the F0 difference between genders was a potential source of confusion, the recognition accuracy would increase when the inputs to the neural network were separated by gender. The current speech database consisted of 33 girls and 28 boys. Interestingly, tone recognition performance with gender sorted versus that with gender mixed differed only by 1 or 2 percentage points. In a previous study on adult speakers, the effects of gender were greater than those observed for child speakers (unpublished observations). The discrepancy between the results of the two studies may be explained by the fact that adult speakers have a greater F0 range difference between genders than child speakers whose F0 ranges have yet started to contrast between genders.

The tone recognition performance of the neural network was significantly better than that of the human listeners when the speech samples from the 61 children were used (Figure 3). The neural network seemed to have a greater capacity in tolerating individual variability than human listeners. In the cross-validation test with the neural network, tone production of each of the individual children was evaluated. The intelligibility of the tone production among the 61 children varied. This variation in tone production intelligibility was also reflected in the tone recognition scores achieved by the adult listeners. The significant positive correlation between the tone recognition scores by the neural network and human listeners on the individual speakers (Figure 4) supports the feasibility of using neural network as a tone recognizer. The cross-validation test with the neural network was also designed to examine the potential effects of age on the intelligibility of the children's tone production. A recent study reported partial acquisition of lexical tone in developing children of 3 years old [2], which suggests that tone could be acquired at various stages of language development. Although the present study did not reveal any effects of age on tone production, future studies with a larger group of children and with younger ages will be necessary to determine such effects.

In conclusion, the neural network was very successful in classifying children's tone production. The recognition performance was not affected by the age or gender of child speakers. The tone recognition performance by the neural network was positively correlated with that of human listeners. Therefore, artificial neural networks can be used in studies of tone production in prelingually deafened children, such as those who have received cochlear implants and may experience various deficits in tone development.

Acknowledgements

We are grateful to Erin Furman, Hui-Hui Li, Chessy Seebohm, Jessica Wolfanger, and Yunfang Zheng for their technical support. The study was supported in part by NIH NIDCD Grant R03-DC006161.

References

[1] Li CN, Thompson SA. The acquisition of tone in Mandarin-speaking children. J Child Lang 1997;4:185–99.

[2] Wong P, Schwartz RG, Jenkins JJ. Perception and production of lexical tones by 3-year-old, Mandarin-speaking children. J Speech Lang Hear Res 2005;48:1065–79.

[3] Moore BCJ. Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants. Otol Neurootol 2003;24:243–54.

[4] Wei WI, Wong R, Hui Y, Au DKK, Wong BYK, Ho WK, et al. Chinese tonal language rehabilitation following cochlear implantation in children. Acta Otolaryngol (Stockh) 2000; 120:218–21.

[5] Lee KYS, van Hasselt CA, Chiu SN, Cheung DMC. Cantonese tone perception ability of cochlear implant children in comparison with normal-hearing children. Int J Pediatr Otorhinolaryngol 2002;63:137–47.

[6] Liu T-C, Chen HP, Lin HC. Effects of limiting the number of active electrodes on Mandarin tone perception in young children using cochlear implants. Acta Otolaryngol (Stockh) 2004;124:1149–54.

[7] Ciocca V, Francis AL, Aisha R, Wong L. The perception of Cantonese lexical tones by early-deafened cochlear implantees. J Acoust Soc Am 2002;111:2250–6.

[8] Xu L, Li Y, Hao J, Chen X, Xue SA, Han D. Tone production in Mandarin-speaking children with cochlear implants: a preliminary study. Acta Otolaryngol (Stockh) 2004;124:363–7.

[9] Peng SC, Tomblin JB, Cheung H, Lin Y-S, Wang L-S. Perception and production of mandarin tones in prelingually deaf children with cochlear implants. Ear Hear 2004;25: 251–64.

[10] Khouw E, Ciocca V. Acoustic and perceptual study of Cantonese tones produced by profoundly hearing-impaired adolescents. Ear Hear 2006;27:243–55.

[11] Chang PC, Sun SW, Chen SH. Mandarin tone recognition by multi-layer perceptron. Proc 1990 IEEE Conf Acoust Speech Signal Process 1990:517–20.

[12] Lan N, Nie KB, Gao SK, Zeng FG. A novel speech-processing strategy incorporating tonal information for cochlear implants. IEEE Trans Biomed Eng 2004;51:752–60.

[13] Wang YR, Chen SH. Tone recognition of continuous Mandarin speech assisted with prosodic information. J Acoust Soc Am 1994;96:2637–45.

[14] Xu L, Zhang W, Zhou N, Lee C-Y, Li Y, Chen X, et al. Mandarin Chinese tone recognition with an artificial neural network. J Otol 2006;1:30–4.