*Original Article*

# Mandarin Chinese Tone Recognition with an Artificial Neural Network

XU Li,[1] ZHANG Wenle,[2] ZHOU Ning,[1] LEE Chaoyang,[1]
LI Yongxin,[3] CHEN Xiuwu,[3] ZHAO Xiaoyan[3]

*1. School of Hearing, Speech and Language Sciences, Ohio University, Athens, OH, USA*
*2. School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, USA*
*3. Beijing Institute of Otorhinolaryngology, Beijing, China*

**Abstract** Mandarin Chinese tone patterns vary in one of the four ways, i.e, (1) high level; (2) rising; (3) low falling and rising; and (4) high falling. The present study is to examine the efficacy of an artificial neural network in recognizing these tone patterns. Speech data were recorded from 12 children (3-6 years of age) and 15 adults. All subjects were native Mandarin Chinese speakers. The fundamental frequencies (F0) of each monosyllabic word of the speech data were extracted with an autocorrelation method. The pitch data(i.e., the F0 contours) were the inputs to a feed-forward backpropagation artificial neural network. The number of inputs to the neural network varied from 1 to 16 and the hidden layer of the network contained neurons that varied from 1 to 16 in number. The output of the network consisted of four neurons representing the four tone patterns of Mandarin Chinese. After being trained with the Levenberg-Marquardt optimization, the neural network was able to successfully classify the tone patterns with an accuracy of about 90% correct for speech samples from both adults and children. The artificial neural network may provide an objective and effective way of assessing tone production in prelingually-deafened children who have received cochlear implants.

**Key words** tone recognition; artificial neural network; tone production; Chinese

## Introduction

Tone languages, such as Mandarin Chinese, use tone patterns to convey lexical meaning. Mandarin Chinese tone patterns vary in one of the four ways, i.e., (1) high level, (2) rising, (3) low falling and rising, and (4) high falling. Pitch information is not explicitly presented in the cochlear implant stimulation with current cochlear implant technology. Therefore, pitch perception of complex acoustic stimuli, such as speech and music, is unlikely to be strong and is probably lacking in current cochlear implant systems. Several studies have indicated that tone perception with cochlear implants is limited to a level that is around chance performance to about 80% correct (e.g., Wei et al, 2000; Lee et al., 2002; Ciocca et al, 2002; Huang et al, 1995, 1996; Sun et al, 1998; Wei et al, 2004). We are concerned about tone production in the tone-language-speaking children who have received cochlear implants because of the limited tone perception performance in these patients. Previous studies have shown that there was a deficiency in tone production in prelingually deaf children who have received cochlear implants(Peng et al, 2004; Xu et al, 2004). However, an objective assessment of the tone production is not available. In the present study, we evaluate artificial neural network as a potential tool for tone production assessment.

Artificial neural network is a computer algorithm in which simple elements are interconnected with each other. The network can be adjusted based on a comparison of the output and the target until the output matches the target. The artificial neural network is widely used for pattern recognition, identification, or classification. A few previous studies have explored the use of artificial neural networks to recognize tone patterns of Mandarin Chinese(e.g., Chang et al, 1990; Lan et al, 2004; Wang and Chen, 1994). In the present study, a simple artificial neural network was developed and the number of inputs and hidden units was adjusted so as

**Corresponding author:** Dr. Xu Li, School of Hearing, Speech and Language Sciences Grover Center Ohio University Athens, OH 45701 USA. E-mail: XuL@ohio.edu

to achieve the most efficient recognition of Mandarin Chinese tone patterns. Speech samples recorded from groups of adults and children were used to evaluate the recognition performance of the neural network.

## Materials and methods

Speech samples were obtained from 15 normal-hearing native Mandarin Chinese speaking adults (22-35 years of age) and 12 normal-hearing native Mandarin Chinese speaking children(3-6 years of age). The adult speakers were recruited from the Ohio University student population. The speech samples from the adult speakers were recorded in a sound-treated booth. The children speakers were recruited from kindergarten classes and elementary schools in Beijing and the recordings were carried out in quiet office rooms. All recordings were conducted with a sampling frequency of 44.1 kHz and a 16-bit resolution. For the adult subjects, the speech samples were recordings of spontaneous productions of the four Mandarin tone patterns of sa, sha, xia, si, shi, xi, su, shu, and xu, resulting in 540 tokens (15 subjects ×9 syllables ×4 tones). For the children, the speech was elicited by asking the subjects to repeat the four tone patterns of 40 sets of Mandarin syllables*(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)*, after an adult native Mandarin speaker resulting in 1920 tokens(12 subjects ×40 syllables ×4 tones).

Preprocessing of the speech materials focused on extraction of the fundamental frequencies (F0) using an autocorrelation method(Kent, 2002). This was performed with a custom program in MATLAB environment. The autocorrelation method sometimes produced errors in F0 estimation, commonly in forms of halving or doubling of the frequencies. These errors were corrected manually based on narrowband spectrograms of the speech samples. The pitch data(i.e., the F0 contours) were then stored in a computer for further neural-network analysis.

A feed-forward backpropagation artificial neural network was developed with the MATLAB Neural Network Toolbox. The architecture of the neural network is shown in Fig.1. The input to the neural network was the F0 contours of the Mandarin Chinese monosyllables. The number of inputs for the neural network varied from 1 to 16. If the number of inputs is N, then the F0 contour was divided evenly into N segments and the frequency values of the middle points of all segments were used as inputs to the neural network. The hidden layer of the network contained neurons that varied from 1 to 16 in number. The output of the network consisted of four neurons representing the four tone patterns of Mandarin Chinese. The network was trained wih the Levenberg-Marquardt optimization. Two thirds of the speech data(360 tokens from the adults or 1280 tokens from the children) were used in training and the remaining one third (180 tokens from the adults or 640 tokens from the children) that were previously unseen by the neural network were used in testing. The training was stopped when the number of epochs of training reached 200 or the sum of squared errors became < 0.01, whichever came first. The neural network was run 10 times and the group mean data were reported below.

## Results

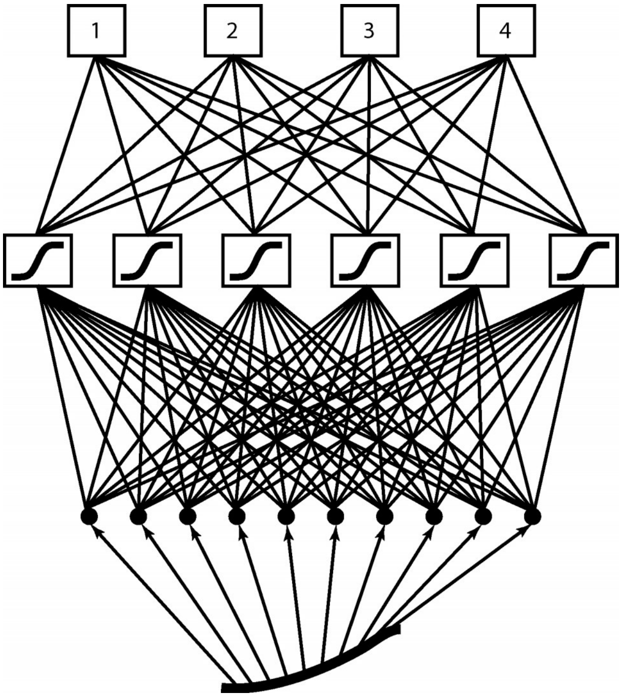

**Figure 1**. Architecture of the artificial neural network. In this example, a pitch contour of tone 2 of a Mandarin Chinese word is divided into 10 evenly spaced segments. The frequency values of the middle points of all segments are used as inputs to the neural net work. There are six hidden neurons each with a nonlinear(sigmoid) transfer function. The four output neurons corres ponded to tones 1, 2, 3, and 4, respectively.

Figure 2 plots the narrowband spectrograms and F0 contours of tones 1 through 4 of Mandarin Chinese monosyllable shu spoken by a normal-hearing female adult. The F0 contours extracted using the autocorrelation method(black symbols) matched the F0 contours in the spectrogram and showed the typical (1) high level, (2) rising, (3) low falling and rising, and(4) high falling tone patterns. F0 extraction of the speech samples from the normal-hearing children also showed the typical tone patterns as seen in the adults.

The F0 contours of the Mandarin Chinese monosyllabic words were fed into the artificial neural network for tone recognition. With a single layer of hidden neurons, our neural network was able to successfully classify the Mandarin tone patterns. The mean asymptotic performance in tone recognition with the training set was 98.6% and 92.6% correct for the adult and children speech samples, respectively. For the unseen testing set, however, the neural network performance dropped a few percentage points. Figures 3 and 4 show the testing results with the unseen testing data set. For the adult speech samples, the network performance increased as the number of hidden neurons increased(Fig. 3, left) or as the number of inputs increased(Fig. 3, right). Overall, the neural network achieved more than 90% correct in recognizing the four tone patterns of Mandarin Chinese with as few as 2 inputs and 3 or 4 hidden neurons (Fig. 3).

For the children's speech samples, the performance of the neural network in recognizing the tone patterns was slightly lower than that for the adult speech materials. In both adult and children speech samples, the neural networks produced recognition performance just around the chance performance of 25% correct（ranging from approximately 18% to 36% correct）when the number of input neurons or the number of hidden neurons was equal to 1(Figs. 3 and 4). For the children speech samples, the asymptotic performance was just below 90% correct(Fig. 4) and was achieved with 3 or 4 input neurons and 4 or 5 hidden

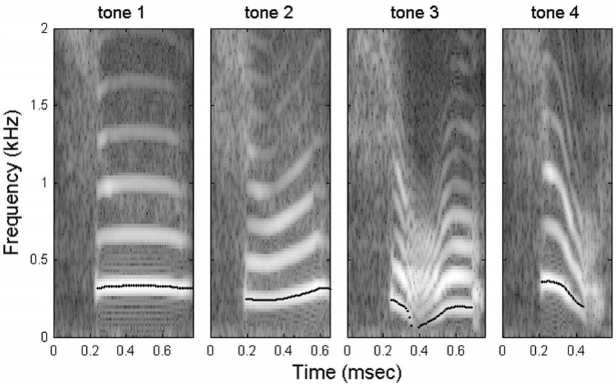

**Figure 2.** Spectrograms and F0 contours of the four tone patterns of Mandarin Chinese monosyllable shu. The speech samples were recorded from a female adult (Subject #F1). The spectrograms are plotted in the narrowband format with the grayscale indicating energy associated with time (abscissa) and frequency(ordinate). The F0 contours extracted with the autocorrelation method are plotted with the black symbols.

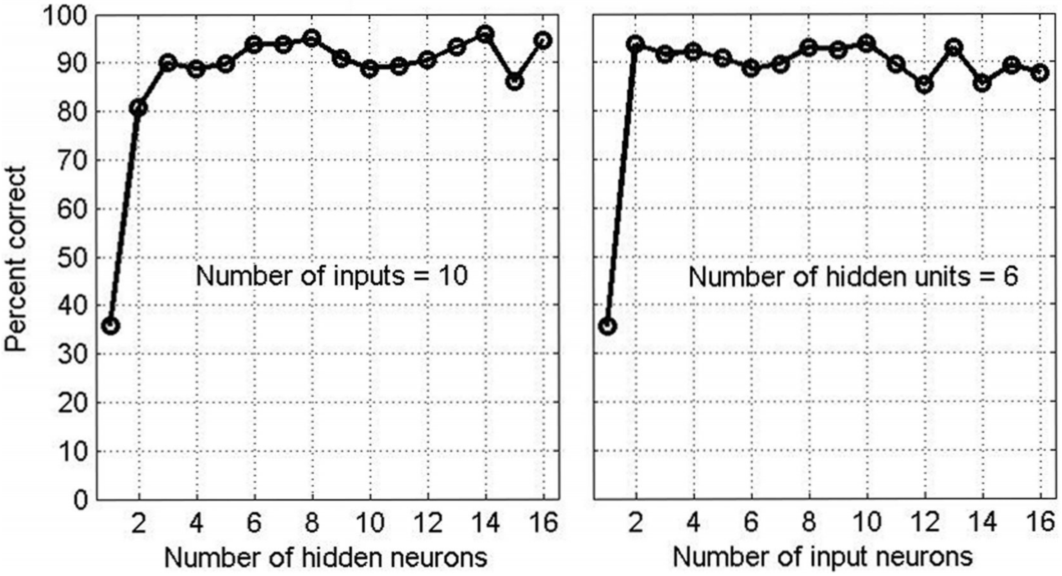

**Figure 3**. Tone recognition performance with the artificial neural network using the speech materials from the adults. Left : Performance as a function of number of hidden neurons with the number of inputs fixed at 10. Right : Performance as a function of number of inputs with the number of hidden neurons fixed at 6.

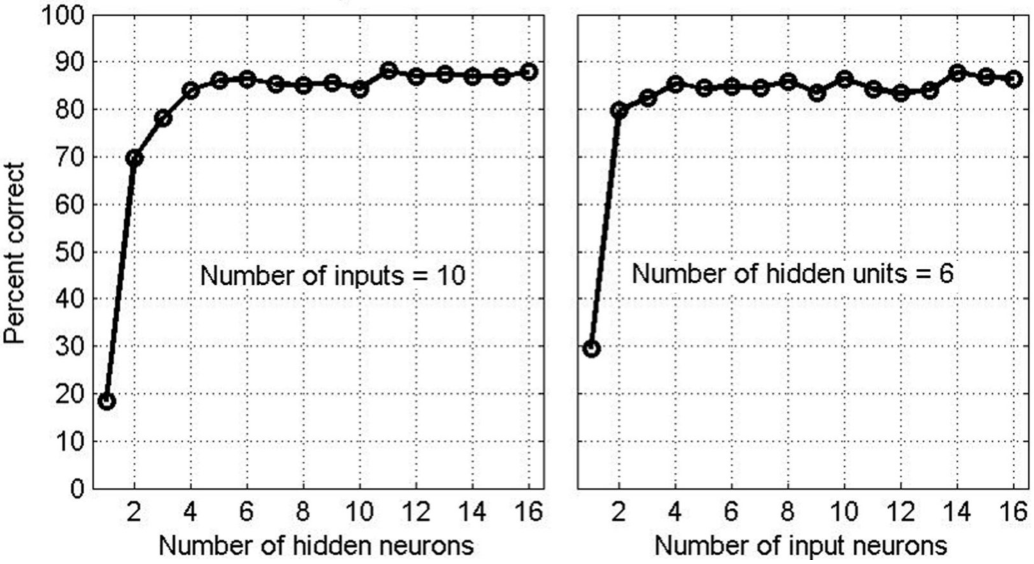

**Figure 4.** Tone recognition performance with the artificial neural network using the speech materials from the children. Conventions as in Figure 3.

neurons. Note that these numbers that were required for the neural network to achieve asymptotic performance appeared to be slightly greater than those for adult speech samples.

## Discussion

A simple artificial neural network with a single layer of hidden neurons can be used to classify tone patterns of Mandarin Chinese. Chang et al (1990) tested tone recognition of isolated Mandarin monosyllables with artificial neural networks in which the number of hidden layers varied from 1 to 3. They found that a single layer of hidden neurons was just as effective as multiple layers of hidden neurons. The hidden neurons in our neural networks had a sigmoid transfer function. In future studies, we will attempt to identify the optimal neural network for tone pattern recognition by implementing other transfer functions, such as linear or radial basis functions, in the hidden-layer neurons.

For the adult speech samples, the neural network can successfully recognize the four Mandarin Chinese tone patterns with as few as 2 inputs and 3 or 4 hidden neurons, with recognition performance slightly above 90% correct. Recognition of tone patterns of other tone languages such as Cantonese and Thai may present a challenge to the neural network because the number of patterns in those languages is greater than four. In a study of tone recognition of isolated Cantonese syllables, Lee et al (1995) used 5 suprasegmental components (including the relative pitch levels, temporal pitch variation patterns and duration of voiced portion) as inputs to a neural network. They found that as many as 25 to 35 hidden neurons were necessary to achieve good tone recognition performance. It remains to be tested whether a small number of hidden neurons will be sufficient for tone recognition of Cantonese syllables if the inputs to the neural networks are the F0 contours as in the present study.

For the children's speech samples, the performance of the neural network in recognizing the tone patterns was slightly lower than that for the adult speech materials. The asymptotic performance was just below 90% correct. The number of hidden neurons and the number of input neurons required for the neural network to achieve asymptotic performance appeared to be greater for the children speech samples than for the adult speech samples (Figs. 3 and 4). Many reasons might contribute to the differences between adult and children data. For example, there were more syllables recorded for the children than for the adults, which might make the tone recognition task of the neural network more difficult using the children speech samples. It is also possible that the younger children in our sample had not mastered tone production. This latter speculation is supported by a recent report that 3-year-old children have not developed perfect tone production (Wong et al, 2005). Our future studies will focus on developmental aspects of tone production.

The artificial neural network can be used to study the salient features in the pitch contour for tone perception. For example, tone recognition with the artificial neural network can be compared with the human perceptual data of tone recognition in situations where various parts of the pitch contours are removed(e.g., Tao et al, 2005). We are especially interested in determining whether the artificial neural network can be used to study tone production in children with cochlear implants. It has been documented that these children have deficiency in tone production in various degrees (Peng et al, 2004; Xu et al, 2004). We will train the neural network with speech samples recorded from normal-hearing, native Mandarin-speaking children . Then, speech samples recorded from the implant children will be subject to test with the neural network. We expect that the neural network will provide an objective and efficient way of assessing tone production in those children.

## Acknowledgements

We are grateful to Erin Furman, Huihui 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 Ciocca V, Francis AL, Aisha R, Wong L. The perception of Cantonese lexical tones by early-deafened cochlear im- plantees. J Acoust Soc Am, 2002, 111: 2250-2256..

2 Chang PC, Sun SW, Chen SH. Mandarin tone recognition by multi-layer perceptron. ICASSP-90, 1990, 517-20.

3 Huang TS, Wang NM, Liu SY. Tone perception of Mandarin-speaking postlingually deaf implantees using the nucleus 22-channel cochlear mini system. Ann Otol Rhinol Laryngol Suppl, 1995, 166: 294-298.

4 Huang TS, Wang NM, Liu SY. Nucleus 22-channel cochlearmini-system implantations in Mandarin-speaking patients. Am J Otol, 1996, 17: 46-52.

5 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: 752-760.

6 Kent, RD, Read, C. Acoustic analysis of speech, 2nd edn. Albany, NY: Singular, 2002.

7 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: 137-147.

8 Lee T, Ching PC, Chan LW, Cheng YH, Mak B. Tone recognition of isolated Cantonese syllables. IEEE Trans Speech Audio Proc, 1995, 3: 204-209.

9 Peng SC, Tomblin JB, Cheung H, Lin YS, Wang LS. Perception and production of Mandarin tones in prelingually deaf Children with cochlear implants. Ear Hear, 2004, 25: 251-264.

10 Sun JC, Skinner MW, Liu SY, Wang FNM, Huang TS, Lin T. Optimization of speech processor fitting strategies for Chinese-speaking cochlear implantees. Laryngoscope, 1998, 108: 560-568.

11 Tao L, Lee CY, Bond ZS. Perception of acoustically modified Mandarin tones by native and non-native listeners. International Conference on Processing Chinese and Other East Asian Languages, Hong Kong, 2005.

12 Wang YR, Chen SH. Tone recognition of continuous Mandarin speech assisted with prosodic information. J Acoust Soc Am, 1994, 96: 2637-2645.

13 Wei CG, Cao KL, Zeng FG. Mandarin tone recognition in cochlear-implant subjects. Hear Res, 2004, 197: 87-95.

14 Wei WI, Wong R, Hui Y, Au DKK, Wong BYK, Ho WK, Tsang A, Kung P, Chung E. Chinese tonal language rehabilitation following cochlear implantation in children. Acta Otolaryngol, 2000, 120: 218-221.

15 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: 1065-1079.

16 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, 2004, 124: 363-367.