Published in final edited form as:

Ear Hear. 2008 June ; 29(3): 326–335. doi:10.1097/AUD.0b013e3181662c42.

Lexical Tone Recognition with an Artificial Neural Network

Ning Zhou¹, Wenle Zhang², Chao-Yang Lee¹, and Li Xu¹

¹*School of Hearing, Speech and Language Sciences, Ohio University, Athens, Ohio*

²*School of Electrical Engineering and Computer Science, Ohio University, Athens, Ohio*

Abstract

Objectives—Tone production is particularly important for communicating in tone languages such as Mandarin Chinese. In the present study, an artificial neural network was used to recognize tones produced by adult native speakers. The purposes of the study were (1) to test the sensitivity of the neural network to speaker variation typically in adult speaker groups, (2) to evaluate two normalization procedures to overcome the effects of speaker variation, and (3) to compare tone recognition performance of the neural network with that of the human listeners.

Design—A feedforward multilayer neural network was used. Twenty-nine adult native Mandarin Chinese speakers were recruited to record tone samples. The F0 contours of the vowel part of the 1044 monosyllabic words recorded were extracted using an autocorrelation method. Samples from the F0 contours were used as inputs to the neural network. The efficacy of the neural network was first tested by varying the number of inputs and the number of neurons in the hidden layer from 1 to 16. The sensitivity of the neural network to speaker variation was tested by (1) using the raw F0 data from speech tokens of a number of randomly drawn speakers that varied from 1 to 29, (2) using the raw F0 data from speech tokens of either male-only or female-only speakers, and (3) using two sets of normalized F0 data (i.e., tone 1-based normalization and first-order derivative) from speech tokens from a number of randomly drawn speakers that varied from 1 to 29. The recognition performance of the neural network under several experimental conditions was compared with the corresponding recognition performance of 10 normal-hearing, native Mandarin Chinese speaking adult listeners.

Results—Three inputs and four hidden neurons were found to be sufficient for the neural network to perform at about 85% correct using speech samples without normalization. The performance of the neural network was affected by variation across speakers particularly between genders. Using the tone 1-based normalization procedure, the performance of the neural network improved significantly. The recognition accuracy of the neural network as a whole or for each tone was comparable with that of the human listeners.

Conclusions—The neural network can be used to evaluate the tone production of Mandarin Chinese speaking adults with human listener-like recognition accuracy. The tone 1-based normalization procedure improves the performance of the neural network to human listener-like accuracy. The success of our neural network in recognizing tones from multiple speakers supports its utility for evaluating tone production. Further testing of the neural network with hearing-impaired speakers might reveal its potential use for clinical evaluation of tone production.

Introduction

Mandarin Chinese uses tones to provide semantic information carried by words. Tone is considered functionally equivalent to segmental structure, the variation of which produces

semantically independent lexical items. This explains how the four distinctive lexical tones contribute to a substantial increase in the number of words in Mandarin Chinese. These tones can be defined in terms of fundamental frequencies (F0) in the vowel portion of a word (Pickett, 1999). Figure 1 illustrates the narrowband spectrograms of tones 1 through 4 of the syllable *shi*. The F0 contours of the four tones show (1) high and flat, (2) rising, (3) low and dipping, and (4) high falling tone patterns. Lexical tones are characterized by the pattern of variation in F0 (F0 contour) over time and height (i.e., absolute frequency value) of the contours. Secondary acoustic features of tones such as duration and amplitude contour also assist tone recognition when F0 information is compromised (Blicher, et al., 1990; Fu & Zeng, 2000; Whalen & Xu, 1992; Xu, et al., 2002).

An artificial neural network is an interconnected group of artificial neurons that uses a mathematical or computational model to process information. The original inspiration for the technique was from examination of bioelectrical networks in the brain formed by neurons and their synapses. In a neural-network model, simple nodes, also known as neurons or units are interconnected to form a network. The nodes operate on a principle similar to biological neurons. The incoming synaptic strength of a biological neuron is modeled with a weight of the node. Each node also has an activation function, also known as a transfer function, which dictates when the node will fire. Not only is the structure of neural networks inspired by the biological nervous system, but functions unique to the brain, such as learning, have also been simulated to a certain extent with neural networks. A neural network adapts its parameters through a continuing process of stimulation. The learning process represents the modification of connection strength between the artificial neurons. Supervised learning, used in the present study, involves the modification of synaptic weights. Essential to the supervised learning procedure is the modification of synaptic weights for the purpose of minimizing differences between the desired responses and actual responses of the network. The training is repeated until there are no significant changes in the synaptic weights, which indicates the completion of an input–output mapping (Bishop, 1995; Haykin, 1994).

Artificial neural networks have been used for automatic recognition of Mandarin Chinese tones. One study that used a simple feedforward multilayer neural network (i.e., perceptron) to recognize Mandarin Chinese tones was conducted by Chang et al. (Reference Note 1) using isolated syllables. The extracted features included the duration and energies of the voiced part of the syllable, and the means and slopes of the F0 contour from each of the three segmented parts. More recent research has begun to create neural networks that recognize tones in running speech (Wang & Chen, 1994a,b). The F0 contour of a syllable in continuous speech is subjected to various modifications because of the influence or anticipation of adjacent syllables (Xu, 1994, 1997). Various models were developed to compensate for the effects of coarticulation in connected speech. For example, incorporating direct contextual features as additional input features was proposed by Wang and Chen (Reference Note 2). Another factor that interferes with the categorization of tones is the declination of F0 in declarative sentential utterances. Classification of a tone becomes dependent on the syllable's location in a sentence. A two-layer network was proposed accordingly, where the intonation of a sentence was included as the target for modeling in the network (Wang & Chen, 1994b). Recent research has also addressed prosodic suprasegmental issues such as stress and timing structure in continuous tone recognition. A recurrent neural network was proposed to infer prosodic states from acoustic features (Wang & Chen, 1994a). Running speech creates more complicated scenarios for recognition modeling because of possible modifications of F0 contours and an increased

¹**Reference Note:** Chang, P. C., Sun, S. W., & Chen, S. H. (1990). Mandarin tone recognition by multi-layer perceptron. *Proceedings of 1990 IEEE Conf Acoust Speech Signal Process*, 517–520.

²**Reference Note:** Wang, Y. R., & Chen, S. H. (1993). Tone recognition of continuous Mandarin speech based on neural networks. *Int Symp Artif Neural Netw*, F01–F10.

variability both within and across speakers. The present study, however, was geared for a more practical purpose than modeling an optimal tone recognizer. The purpose was to develop and study the efficacy of a neural network for its potential use in clinical assessment of tone production.

Contemporary cochlear implant systems do not deliver the explicit pitch information necessary for pitch perception of complex acoustic stimuli, such as speech and music (Geurts & Wouters, 2001, 2004; Green, et al., 2002, 2004; Kong, et al., 2004; Moore, 2003; Xu & Pfingst, 2003; Xu, et al., 2002), which hinders both tone perception and production for prelingually deafened recipients (Ciocca, et al., 2002; Han, et al., 2007; Huang, et al., 1996; Lee, et al., 2002; Peng, et al., 2004; Sun, et al., 1998; Wei, et al., 2000; Xu, et al., 2004). Traditionally, tone production is elicited by repetition or picture naming and judged by human listeners in a subjective manner (e.g., Khouw & Ciocca, 2006; Peng, et al., 2004; Wong, et al., 2005). Our goal was to use an artificial neural network to assess tone production of hearing-impaired patients. In our previous studies, we have trialed a simple feedforward multilayer neural network to recognize isolated tones of 61 normal-hearing children (Xu, et al., 2007) and successfully used it to evaluate the tone production of 14 cochlear implant children (Zhou & Xu, 2008). The overall recognition accuracy of the neural network was comparable with that of the human listeners and the recognition scores for individuals were highly correlated with the individual scores given by the human listeners. The confusion matrix analysis revealed that the error patterns of the neural network and the human listeners also matched closely (Zhou & Xu, 2008).

In the aforementioned studies (Xu, et al., 2007; Zhou & Xu, 2008), the impact of speaker variation on the performance of the neural network seems to be limited because of the homogeneity of children's F0 height. In the present study, we extended our investigation to adult speakers. Assessing tone production of adult speakers is accompanied by an increase in across speaker variation, partly because of the F0 differences between genders. One of the specific aims of the present study was to study how the neural network might be affected by the variations that an adult speaker data set introduces. Furthermore, two normalization procedures were used in an attempt to overcome such variation for the neural network to achieve human listener-like recognition. One procedure followed what Gauthier et al. (2007) introduced as the first-order derivative of the F0 contour that was used as input to a neural network to overcome the effects of both across and within speaker variations. We also developed our own normalization procedure, which put more emphasis on reducing across speaker variation. We then compared the perceptual results from human listeners with the recognition patterns of the neural work using the raw F0 data and two sets of normalized F0 data. Findings provided some insight as to how the auditory system in human listeners organizes and normalizes lexical tones spoken by different speakers.

Materials and Methods

Speech Database and Feature Extraction

Twenty-nine, normal-hearing, native Mandarin Chinese speaking adults (13 males and 16 females, with ages ranging from 22 to 35 years) were recruited from the Ohio University's student and staff population for the recording of the test speech materials. The speech database consisted of spontaneous production of the four Mandarin tone patterns of the following 9 syllables: *sa*, *sha*, *xia*, *si*, *shi*, *xi*, *su*, *shu*, and *xu*. Therefore, a total of 1044 (9 syllables × 4 tones × 29 speakers) Mandarin Chinese monosyllabic words were collected. We used isolated words in this study because judging tone intelligibility on sentence levels is not practical in clinics. The speech recordings were performed in an IAC sound booth with a sampling rate of 44,100 Hz and a 16-bit resolution. The database was also established for an on-going study of fricative consonants. The three vowels used are representative of Mandarin Chinese vowels in terms of their articulatory properties. Note that certain combinations of tones and syllables

produce nonwords in Chinese. Among the 36 syllables, there are four nonwords [i.e., sa (tone 2), xia (tone 3), si (tone 2), and su (tone 3)].

Feature extraction involved only F0, the most dominant and relevant acoustic correlate for lexical tone recognition (Ho, 1976). F0 contours of the vowel part of the 1044 recorded syllables were extracted with an autocorrelation method (Kent & Read, 2002). Note that the F0 contours included F0 information only from the vowel part of a syllable without the initial frication. The few errors that arose from F0 extraction, usually in forms of frequency doubling or halving, were manually corrected based on the narrowband spectrograms of the syllables. An example of the extracted F0 contour is illustrated in Figure 1, where the black lines represent the F0 contours extracted using the autocorrelation method. The update rate of the F0 extraction was 8 msec with a frame size of 24 msec. In this example, a close match of the extracted F0 and the F0 contour plotted with the spectrograms was achieved.

Efficacy Test of the Neural Network

A feedforward backpropagation multilayer perceptron was developed in MATLAB with the Neural Network Toolbox (MathWorks, Natick, MA). The network consists of multiple layers of artificial neurons, interconnected in a feedforward way. Each neuron in one layer has direct connections to the neurons of the neighboring layer. An example of the multilayer perceptron architecture is shown in Figure 2. The inputs to the neural network were sampled from the F0 contours. The output layer of the neural network consists of four neurons corresponding to the four Mandarin Chinese tones.

For the efficacy test, F0 data from speech tokens of all the 29 speakers were used. The numbers of inputs and hidden neurons varied from 1 to 16 to determine the minimal number of samples of an F0 contour and hidden neurons that could provide satisfactory tone recognition performance. If the number of inputs was N , the F0 contour of each speech token was evenly segmented to N parts. The even segmentation was equivalent to normalizing duration of the tokens. The averaged F0 values of each part made up an N -element vector that was the input to the neural network. Occasionally, if an F0 contour contained fewer than N samples, that tone token was excluded from the test. Systematically varying the numbers of inputs and hidden neurons created 256 (16×16) experimental conditions. Each condition was tested 10 times with randomized inputs. One half of the speech data was used to train the neural network using the Levenberg-Marquardt optimization, and the other half was used to test it. The training was stopped when the sum of squared errors became less than 0.01 or the number of training iterations reached 200.

Speaker Variation Sensitivity Tests

The speaker variation sensitivity tests included tests of the neural network using (1) the raw F0 data from speech tokens of various numbers of randomly drawn speakers, (2) the raw F0 data from speech tokens of either male- or female-only speakers, and (3) two sets of normalized F0 data from speech tokens of various numbers of randomly drawn speakers.

The first speaker variation sensitivity test evaluated the sensitivity of the neural network to the variability across speakers. While we fixed both the number of inputs and the number of hidden neurons at 16, the number of speakers, whose tone tokens were included for testing, varied from 1 to 29. For each condition, speakers were randomly drawn with replacement from the pool of 29 speakers. The random selection of speakers and the testing for the neural network were repeated 10 times. For the condition of 29 speakers, tone tokens from all 29 speakers were included as in the efficacy test described earlier.

A large amount of variability across speakers can be attributed to the F0 height difference between female and male speakers. Therefore, the second speaker variation sensitivity test examined the effects of F0 height difference between genders on tone recognition performance of the neural network. The tone tokens were sorted according to the gender of the speakers. Tone tokens from the 13 male and 16 female speakers were then separately used to train and test the neural network. The recognition accuracy for tone tokens from either male or female speakers was compared with the recognition accuracy for tone tokens randomly drawn from gender-mixed speakers, described earlier. The potential difference in the scores of the 13 male speakers and the 13 randomly drawn gender-mixed speakers, or difference in the scores of the 16 female speakers and 16 randomly drawn gender-mixed speakers, would indicate the effects of gender differences in F0 height on the tone recognition performance of the neural network.

Two normalization procedures were then used to reduce speaker variation. One followed the method introduced by Gauthier et al. (2007). The first-order derivative of each tone token's F0 contour was obtained using $D1 = 0.5 [F_0(t + 1) - F_0(t - 1)]$. Thus, it was the velocity of the F0 contour that was used as input to the neural network. This procedure completely removed the F0 height differences across tokens, thus the height differences across tones and across speakers were both eliminated. The second normalization procedure was based on the F0 height of tone 1. Since Mandarin Chinese tone 1 is a flat high frequency tone (see Fig. 1), we used the mean F0 of tone 1 from each subject to normalize all tone tokens from that subject. That is, the mean F0 value for tone 1 for a given speaker was subtracted from all tokens for that speaker. This procedure of normalization was repeated for all the speakers. Thus, the normalization forced tone 1 of all the speakers to the same height and at the same time kept the relative height and contour patterns among tones. The neural network was then tested using the two sets of normalized F0 data. The number of speakers whose tone tokens were included again varied from 1 to 29 following the same test procedure used for the speaker variation sensitivity test.

Tone Recognition Test by the Human Listeners

Ten additional, normal-hearing, native Mandarin Chinese speakers (4 males and 6 females, aging from 26 to 43 years) were recruited for the tone recognition experiment. A custom graphical user interface was created in MATLAB to present the tone tokens and to collect the participants' responses. The tone recognition test was done in an IAC sound booth in two sessions. In the first session, the presentation of tone tokens was in random order for tones and for speakers. This session of the tone recognition tests corresponded to the test condition of the neural network where the number of speakers was equal to 29. In the second session, the speech tokens were presented in the order of speakers, one speaker per block. Presentation of tone tokens in the order of speakers corresponded to the test condition of the neural network where the number of speakers was equal to 1. Because the participants had already listened to all of the tone tokens in the first session, the second session was scheduled at least 2 wks after the first one to minimize potential learning effects. In both sessions, the subjects were instructed to use a computer mouse to click on a button labeled 1, 2, 3, or 4 after each presentation of the stimulus to indicate the tone of the speech token they heard. Clicking on any of the buttons activated the presentation of the next tone token. All 10 participants completed the first session, whereas five completed the second one.

Results

The tone recognition performance by the neural network was dependent on both the number of inputs and the number of hidden neurons (Fig. 3A). As the number of inputs to the neural network or the number of hidden neurons increased, the performance of the neural network improved. With as few as 2 inputs, the recognition accuracy of the neural network reached

80% correct. A performance plateau occurred when one more input was fed into the neural network (Fig. 3B). A tone confusion matrix analysis was performed to elucidate which tone (s) contributed to the performance plateau achieved with 3 inputs as opposed to 2 inputs. It can be seen that an increase in the number of inputs from 2 to 3 greatly improved the classification accuracy of tone 3 (Fig. 3C). The number of hidden neurons sufficient for the neural network to reach a performance plateau was as few as 4 (Fig. 3D). Tone recognition performance by the neural network stayed around 85% correct when the number of inputs was ≥ 3 or the number of hidden neurons was ≥ 4 .

Figure 4 shows the mean tone-recognition performance by the neural network and the human listeners. The mean performance by the neural network using the raw F0 data is plotted using a solid line against the number of speakers. The mean performance starts with a perfect score and decreases linearly as the number of speakers increases. When tone tokens from all 29 speakers were used, the performance reaches 85.3% correct (SD = 1.7%).

The mean tone-recognition performance of the neural network for the tone tokens of the male-only and female-only speakers is plotted in Figure 4 for 13 and 16 speakers, respectively. Compared with the mean performance of the neural network using the same numbers of speakers drawn at random (solid line, Fig. 4), an improvement in tone recognition of 2.0 and 6.5 percentage points was observed when tone tokens from the male-only and female-only speakers were used, respectively. The improvement was statistically significant (male: $t = 3.414$, $p = 0.003$; female: $t = 9.148$, $p < 0.001$). The tone confusion matrices describe in detail the error patterns under the mixed gender and single gender conditions (Fig. 5). The improvement of tone recognition performance using tone tokens from the 13 male speakers over 13 mixed-gender speakers arose predominantly from an increase in recognition accuracy for both tone 3 and tone 4 (Fig. 5, top row). On the other hand, the improvement of tone recognition performance using tone tokens from the 16 female speakers over 16 mixed-gender speakers arose from an increase in recognition accuracy for all four tones, although the most dramatic increase was again seen for tone 3 (Fig. 5, bottom row).

In Figure 4, the mean recognition performances of the neural network using the two normalization procedures are plotted in the dashed and dotted lines, respectively. The two functions start with a perfect score and decrease linearly as the number of speakers increases from 1 to 29. When the tone tokens of all 29 speakers were included, the performance using the first-order derivative normalization procedure ends at 83.4% correct, and the performance using the tone 1-based normalization ends at 91.5% correct. The results of three oneway analysis of variances showed that the effects of number of speakers were statistically significant for the performance using the raw F0 data ($F(28, 261) = 56.8$, $p < 0.001$) and for the performance after two normalization procedures (First-order derivative: $F(28, 261) = 48.74$, Tone 1 based: $F(28, 261) = 12.92$, both $p < 0.001$). The results of each normalization procedure were compared with that using the raw F0 data. Using the first-order derivative procedure caused the performance of the neural network to decrease from its performance using the raw F0 data ($F(1, 522) = 40.05$, $p < 0.001$). An interaction was also found to be significant ($F(28, 522) = 1.66$, $p < 0.05$), suggesting greater decrease for the first-order derivative inputs as the number of speakers increased. The tone 1-based normalization procedure significantly improved the performance of the neural network ($F(1, 522) = 248.3$, $p < 0.001$), and such improvement was greater as the number of speakers increased ($F(28, 522) = 6.29$, $p < 0.001$).

A comparison of confusion matrices at the maximal number of speakers showed that all tones were more accurately recognized by the neural network after the tone 1-based normalization, with the greatest benefit observed, again, for tones 3 and 4 (Fig. 6).

Tone recognition results from human listeners consisted of data from two experimental conditions: presentations of tone tokens in blocks and in random order of speakers. The mean performance for tone tokens blocked by speakers is plotted at 1 speaker (abscissa) in Figure 4. This corresponded to the recognition results of the neural network for tone tokens from only one speaker (Fig. 4). The mean performance for tone tokens presented in a random order of speakers is plotted at 29 speakers (abscissa). This corresponded to how well neural network recognized tones from all 29 speakers. The human listeners did remarkably better when the presentation of tone tokens was in blocks of speakers (mean = 97.9%, SD = 0.75%) than when the presentation was in a random order (mean = 90.6%, SD = 3.51%) ($t = 4.550$, $p < 0.001$). For the condition of 29 speakers, the performance of the human listeners was comparable with the performance of the neural network that used the tone 1-based normalization (mean = 91.5%, SD = 1.31%) ($t = -0.791$, $p = 0.440$), but was significantly better than the performance of the neural network using raw F0 data or the first-order derivative data ($p < 0.001$). Furthermore, the human listeners and the neural network using the tone 1-based normalization procedure showed similar recognition accuracy for each of the four tones from the analysis of tone confusion matrices.

The recognition of the four tones by the neural network were 97.1%, 92.6%, 77.5%, and 98.0% correct, whereas the recognition of the four tones by the human listeners were 97.7%, 95.2%, 71.1%, and 99.0% correct, respectively.

Discussion

In the present study, the input to our neural network involved only samples from the F0 contours. The F0 contours typically demonstrate (1) high and flat, (2) rising, (3) low and dipping, and (4) high falling tone patterns (Fig. 1). As our results indicated, three evenly spaced samples from any of the four F0 contours were sufficient to define its pattern (Fig. 3A, B). With two inputs the performance was slightly below the plateau level, which was due to the fact that tone 3 has a complex F0 variation pattern that requires at least three samples to define. The tone confusion matrix analysis confirmed that adding one more F0 sample as input greatly improved the classification accuracy of tone 3, whereas the performance for other tones was not noticeably changed (Fig. 3C).

We also found that four hidden neurons were sufficient for the neural network to reach a performance plateau (Fig. 3D). In a study of tone recognition of isolated Cantonese syllables (Lee, et al., 1995), they used five 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 F0 contours as in the present study. Note that the number of tone patterns in Cantonese is six as opposed to four in Mandarin Chinese.

Variation of tone production can occur both within and across speakers. Our neural network achieved perfect tone recognition when the inputs were blocked by speakers, suggesting that the neural network was highly resistant to the variability within speakers. It is worth noting, however, that tones in their isolated forms produce less within speaker variability than those in connected speech (Xu, 1997). The neural network was affected by across speaker variations to a larger degree. As the number of speakers included in the inputs became larger, the tone recognition performance of the neural network was also increasingly compromised. The decreased performance of the neural network could be the result of the across speaker variations in the speakers' F0 height. One obvious variation from an adult speaker group is the F0 height difference between genders, the difference being that the F0 of female speakers is usually

higher in frequency than that of male speakers. Consider cases where the F0 contour of a low tone by a female speaker may fall in the neighborhood of the F0 contour of a high tone spoken by a male speaker. We compared the confusion matrices for recognizing single-gender tone tokens with the confusion matrices for recognizing randomly drawn gender-mixed tone tokens. The results showed that the F0 height differences between genders contributed to recognition errors among all tones, but especially between tones 3 and 4 (Fig. 5). Tone 3, the one that has the greatest variability among the four tones, often loses the rising part of its contour (Liu & Samuel, 2004; Speer, et al., 1989). This type of variation makes tone 3 essentially a falling tone that resembles the contour of tone 4, except that the falling of tone 3 starts at a lower frequency value (Speer, et al., 1989). When the contours of tones 3 and 4 contrast only in height, it is conceivable that confusion may arise between a relatively high tone 3 produced by a female speaker and a relatively low tone 4 produced by a male speaker.

Two normalization procedures were used to reduce such across speaker variation. The first-order derivative procedure adopted from Gauthier et al. (2007) removed the F0 height contrast between tones and also across speakers. The input to the neural network represented the pattern or the velocity of the contours. Consistent with their findings, a recognition accuracy of 97% for 3 speakers, our corresponding score was 99.5% correct (Fig. 4). However, when speech samples from a larger number of speakers were used, the slope of the decrease in performance closely followed that using the raw F0 data with no evidences of improvement. Note that the multiplayer perceptron network that we used in this study was very different from the self-organizing map neural network that was used by Gauthier et al. (2007). However, the less successful tone recognition of speech samples of a larger number of speakers might be rooted in the procedure itself. The first-order derivative procedure successfully normalized the F0 height differences from one speaker to another, yet the intrinsic F0 height cue between tones was also compromised. This may affect the recognition of tones that contrast only in height. For example, when tone 3 loses its rising part as discussed earlier, it could be mistaken as tone 4 after the first-order derivative normalization. However, use of the first-order derivative information may benefit tone recognition in the context of running speech, when such intrinsic height relationships between tones of a particular speaker also become unreliable. In that case, one could have had worse performance without normalization to start with, if the input to the network were samples of connected speech. In our case of isolated tone recognition, the benefit of normalization by completely removing F0 height differences did not seem to compensate for the loss of the intrinsic F0 height cue between tones. Thus, the performance with this approach did not seem to improve from the performance without normalization. Our procedure (i.e., normalization based on mean F0 of tone 1), on the other hand, maintained the height differences among tones within speakers. With all speakers' mean tone 1 height set to zero, this procedure matched the tonal spaces of different speakers and maintained the relative height relationship between the four tones by any particular speaker. It would be interesting to test whether such a normalization procedure would also benefit tone recognition of connected speech.

The confusion matrix analysis showed that normalization based on tone 1 significantly improved the recognition accuracy of tone 3 by reducing the false recognition of tone 3 as tone 4 (Fig. 6). Similarly, the recognition of tone 4 benefited from a reduction of false recognition of tone 4 as tone 3. Such pattern of error reduction between tones 3 and 4 resembled what was observed when the tone tokens from male-only and female-only speakers were used to test the tone recognition of the neural network (Fig. 5). This observation suggests that the normalization to a large degree removed the variability in F0 height across speakers, especially between genders. This normalization procedure not only greatly improved the performance of the neural network, but also, and more importantly, produced a recognition pattern comparable with that of the human listeners.

Human listeners are known to have remarkable resistance to speaker variation. Early studies of speaker normalization showed that human listeners can correctly perceive a vowel that has different formant frequencies spoken by female, male, and children speakers (e.g., Ladefoged & Broadbent, 1957). Theories on speaker variation in vowel perception emphasize either “extrinsic specification” or “intrinsic specification” of the vowels. The intrinsic normalization model claims that acoustic content of a speech signal is by itself sufficient for identification if all the important aspects of the content are appropriately represented (Peterson, 1961); whereas the extrinsic normalization model hypothesizes that when a speech signal is acoustically ambiguous, the signal needs to be put in the “frame of reference” of the speaker to be correctly identified (Joos, 1948). To apply the normalization theory of vowels to tones, the intrinsic specification of the Mandarin Chinese tones becomes their distinctive patterns and absolute heights of the F0 contours. Ambiguity occurs when the tones lose their distinctive patterns, especially in coarticulation, and when the F0 height cue is blurred across speakers. Under such circumstances, the extrinsic model explains how human listeners normalize and perceive tones. The F0 height of a tone hints at its identity in the reference of the speaker's F0 range or tonal space. To recognize tones that contrast only in F0 height, such as those of Cantonese tones, Wong and Diehl (2003) found that the human listeners rely strongly on the contextual F0 height to make tone identification. The same ability of Mandarin speakers to normalize F0 height has also been demonstrated in laboratories (Leather, 1983; Moore & Jongman, 1997) and in everyday life (e.g., when a child is having a conversation with his or her parents). Our tone 1-based normalization procedure preserves the F0 height relationship across tones and presents this relationship in a matched tonal space (i.e., zero mean tone 1 for all speakers), which is equivalent to referencing a tone token's F0 height in the tonal space of that speaker. Hence, this procedure shows great similarity to the extrinsic model of human speaker normalization. Its success and, more importantly, its correlation with perceptual performance suggest possible links of this normalization method to the mechanisms that the auditory system uses for tone recognition.

In conclusion, with 3 inputs and 4 neurons, the feedforward multilayer neural network was able to recognize Mandarin Chinese tones with ~85% accuracy. Normalization based on the height of tone 1 effectively improved the performance of the neural network to human listener-like accuracy. The success of using our neural network to recognize tones of normal-hearing speakers suggests future use of this neural network in clinical practice. The neural network can be trained with normal speech samples and then be used to classify speech samples from the patients. The output from the neural network would indicate the tone production accuracy of the patients. Therefore, the neural network can potentially become an objective way of evaluating tone production in clinics.

Acknowledgements

The authors are grateful to Erin Furman, Hui-Hui Li, Chessy Seebohm, and Jessica Wolfanger for their technical support.

The study was supported in part by NIH NIDCD Grant R03-DC006161.

References

- Bishop, CM. Neural Networks for Pattern Recognition. Oxford: Oxford University Press; 1995.
- Blicher DL, Diehl R, Cohen LB. Effects of syllable duration on the perception of the Mandarin Tone 2/ Tone 3 distinction: evidence of auditory enhancement. *J Phon* 1990;18:37–49.
- Ciocca V, Francis AL, Aisha R, et al. The perception of cantonese lexical tones by early-deafened cochlear implantees. *J Acoust Soc Am* 2002;111:2250–2256. [PubMed: 12051445]
- Fu QJ, Zeng FG. Identification of temporal envelope cues in Chinese tone recognition. *Asia Pac J Speech Lang Hear* 2000;5:45–57.

- Gauthier B, Shi R, Xu Y. Learning phonetic categories by tracking movements. *Cognition* 2007;103:80–106. [PubMed: 16650399]
- Geurts L, Wouters J. Coding of the fundamental frequency in continuous interleaved sampling processors for cochlear implants. *J Acoust Soc Am* 2001;109:713–726. [PubMed: 11248975]
- Geurts L, Wouters J. Better place-coding of the fundamental frequency in cochlear implants. *J Acoust Soc Am* 2004;115:844–852. [PubMed: 15000196]
- Green T, Faulkner A, Rosen S. Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants. *J Acoust Soc Am* 2002;112:2155–2164. [PubMed: 12430827]
- Green T, Faulkner A, Rosen S. Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implant. *J Acoust Soc Am* 2004;116:2298–2310. [PubMed: 15532661]
- Han D, Zhou N, Li Y, et al. Tone production of Mandarin-Chinese-speaking children with cochlear implants. *Int J Pediatr Otorhinolaryngol* 2007;71:875–880. [PubMed: 17376544]
- Haykin, S. *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ: Prentice Hall; 1994.
- Ho AT. The acoustic variation of Mandarin tones. *Phonetica* 1976;33:353–367.
- Huang TS, Wang NM, Liu SY. Nucleus 22-channel cochlear mini-system implantations in Mandarin-speaking patients. *Am J Otol* 1996;17:46–52. [PubMed: 8694134]
- Joos M. Acoustic phonetics. *Language* 1948;24(suppl 2):1–136.
- Kent, RD.; Read, C. *Acoustic analysis speech*. 2nd. Albany, NY: Singular; 2002.
- Khouw E, Ciocca V. Acoustic and perceptual study of Cantonese tones produced by profoundly hearing-impaired adolescents. *Ear Hear* 2006;27:243–255. [PubMed: 16672793]
- Kong YY, Cruz R, Jones JA, et al. Music perception with temporal cues in acoustic and electric hearing. *Ear Hear* 2004;25:173–185. [PubMed: 15064662]
- Ladefoged P, Broadbent DE. Information conveyed by vowels. *J Acoust Soc Am* 1957;29:98–104.
- Leather J. Speaker normalization in perception of lexical tone. *J Phon* 1983;11:373–382.
- Lee KYS, van Hasselt CA, Chiu SN, et al. Cantonese tone perception ability of cochlear implant children in comparison with normal-hearing children. *Int J Pediatr Otorhinolaryngol* 2002;63:137–147. [PubMed: 11955605]
- Lee T, Ching PC, Chan LW, et al. Tone recognition of isolated cantonese syllables. *IEEE Trans Speech Audio Processing* 1995;3:204–209.
- Liu S, Samuel AG. Perception of Mandarin lexical tones when F0 information is neutralized. *Lang Speech* 2004;47:109–138. [PubMed: 15581188]
- Moore BCJ. Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants. *Otol Neurotol* 2003;24:243–254. [PubMed: 12621339]
- Moore CB, Jongman A. Speaker normalization in the perception of Mandarin Chinese tones. *J Acoust Soc Am* 1997;102:1864–1877. [PubMed: 9301064]
- Peng S, Tomblin JB, Cheung C, et al. Perception and production of Mandarin tones in prelingually deaf children with cochlear implants. *Ear Hear* 2004;25:251–264. [PubMed: 15179116]
- Peterson G. Parameters of vowel quality. *J Speech Hear Res* 1961;4:10–29. [PubMed: 13734834]
- Pickett, JM. *The acoustics of speech communication: fundamentals, speech perception theory, and technology*. Boston: Allyn & Bacon; 1999.
- Speer SR, Shih C, Slowiczek ML. Prosodic structure in language understanding: evidence from tone sandhi in Mandarin. *Lang Speech* 1989;32:337–354. [PubMed: 2485849]
- Sun JC, Skinner MW, Liu SY, et al. Optimization of speech processor fitting strategies for Chinese-speaking cochlear implantees. *Laryngoscope* 1998;108:560–568. [PubMed: 9546270]
- Wang YR, Chen SH. Tone recognition of continuous Mandarin speech assisted with prosodic information. *J Acoust Soc Am* 1994a;96:2637–2645. [PubMed: 7983269]
- Wang YR, Chen SH. Tone recognition of continuous Mandarin speech based on hidden Markov model. *Intern J Pattern Recognit Artif Intell* 1994b;8:233–246.
- Wei WI, Wong R, Hui Y, et al. Chinese tonal language rehabilitation following cochlear implantation in children. *Acta Otolaryngol* 2000;120:218–221. [PubMed: 11603776]

- Whalen DH, Xu Y. Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica* 1992;49:25–47. [PubMed: 1603839]
- Wong P, Schwartz RG, Jenkins J. Perception and production of lexical tones by 3-year-old, Mandarin-speaking children. *J Speech Lang Hear Res* 2005;48:1065–1079. [PubMed: 16411796]
- Wong PCM, Diehl RL. Perceptual normalization for inter-and intratalker variation in Cantonese level tones. *J Speech Lang Hear Res* 2003;46:413–421. [PubMed: 14700382]
- Xu L, Pfingst BE. Relative importance of the temporal envelope and fine structure in tone perception. *J Acoust Soc Am* 2003;114:3024–3027. [PubMed: 14714781]
- Xu L, Chen X, Zhou N, et al. Recognition of lexical tone production of children with an artificial neural network. *Acta Otolaryngol* 2007;127:365–369. [PubMed: 17453455]
- Xu L, Li Y, Hao J, et al. Tone production in Mandarin-speaking children with cochlear implants: a preliminary study. *Acta Otolaryngol* 2004;124:363–367. [PubMed: 15224854]
- Xu L, Tsai Y, Pfingst BE. Features of stimulation affecting tonal-speech perception: implications for cochlear prostheses. *J Acoust Soc Am* 2002;112:247–258. [PubMed: 12141350]
- Xu Y. Production and perception of coarticulated tones. *J Acoust Soc Am* 1994;95:2240–2253. [PubMed: 8201120]
- Xu Y. Contextual tonal variations in Mandarin. *J Phon* 1997;25:61–83.
- Zhou N, Xu L. Development and evaluation of methods for assessing tone production skills in Mandarin-speaking children with cochlear implants. *J Acoust Soc Am*. 2008In Press

Fig. 1.

Spectrograms and F0 contours of the four tone patterns of *shi* spoken by a female subject (F1). The spectrograms are plotted in the narrowband format with the grayscale indicating energy associated with time (abscissa) and frequency (ordinate). The black lines represent the F0 contours extracted with the autocorrelation method.

Fig. 2.

Architecture of the multilayer perceptron neural network. In this example, a pitch contour of tone 4 of a Mandarin Chinese word was divided into 10 evenly spaced segments. The averaged frequency values of each of the segments were used as inputs to the neural network. There were six hidden neurons each with a nonlinear transfer function. The four output neurons corresponded to tones 1, 2, 3, and 4, respectively.

Fig 3.

Tone recognition performance by the artificial neural network. A, Tone recognition performance as a function of the number of inputs (abscissa) and the number of hidden neurons (ordinate). The gray scale in each cell indicates the performance score with reference to the color bar on the right. B, Performance of the neural network as a function of number of inputs with the number of hidden neurons fixed at 6. Error bars are standard deviations. C, Confusion matrices produced by the neural networks with 2 inputs (left) and 3 inputs (right). The number of hidden neurons was set at 4. The value in the cell of row j and column k is the percent of times when stimulus tone j was recognized as tone k ($j = 1, 2, 3, \text{ or } 4$; $k = 1, 2, 3, \text{ or } 4$). The gray scale in each cell reflects the value in it with reference to the color bar on the right. D, Performance of the neural network as a function of number of hidden neurons with the number of inputs fixed at 9. Error bars are standard deviations.

Fig. 4.

Tone recognition performance by the neural network and the human listeners. Mean scores of the artificial neural network are plotted as the numbers of speakers varied from 1 to 29. The mean scores with the raw F0 data are connected with a solid line, whereas the mean scores with the tone 1-based normalization and first-order derivative data are connected with a dotted line and dashed line, respectively. The mean scores by the neural network on the 13 male speakers (open square) and the 16 female speakers (open circle) are plotted at the corresponding numbers of speakers. The number of inputs and hidden neurons of the neural network were both fixed at 16. Mean tone recognition score by the human listeners when the presentations were in the order of speakers is plotted using a filled circle at 1 speaker. Mean tone recognition score by the human listeners when the presentations were in a random order of speakers is plotted using a filled circle at the number of speakers of 29.

Fig. 5.

Tone confusion matrices. Top row: error patterns of neural-network recognition of tone tokens from the 13 randomly drawn gender-mixed speakers versus those from the 13 male-only speakers. Bottom row: Error patterns of neural-network recognition of tone tokens from the 16 randomly drawn gender-mixed speakers vs. those from the 16 female-only speakers (bottom row). The value in the cell of row j and column k is the percent of times stimulus tone j was recognized as tone k ($j = 1, 2, 3$, or 4 ; $k = 1, 2, 3$, or 4). The gray scale in each cell reflects the value in it with reference to the color bar on the right.

Fig. 6.

Tone confusion matrices. The tone recognition error patterns of the neural network for the raw tone tokens (left) and the tone tokens normalized based on tone 1 (right). The value in the cell of row j and column k is the percent of times stimulus tone j was recognized as tone k ($j = 1, 2, 3$, or 4 ; $k = 1, 2, 3$, or 4). The gray scale in each cell reflects the value in it with reference to the color bar on the right.