LETTER

# Effects of the Temporal Fine Structure in Different Frequency Bands on Mandarin Tone Perception

**Lin YANG**[†a)], **Jianping ZHANG**[†], *Nonmembers*, **Jian SHAO**[†], *Student Member, and* **Yonghong YAN**[†], *Nonmember*

**SUMMARY** This letter evaluates the relative contributions of temporal fine structure cues in various frequency bands to Mandarin tone perception using novel "auditory chimaeras". Our results confirm the importance of temporal fine structure cues to lexical tone perception and the dominant region of lexical tone perception is found, namely the second to fifth harmonics can contribute no less than the fundamental frequency itself.

***key words:*** *tone perception, auditory chimaeras, fine structure cues*

## 1. Introduction

Mandarin Chinese, as a kind of lexically tonal language, is perceived differently from English. Perception studies have proven the importance of tones to the intelligibility of tonal languages [1]. Research on the lexical tone perception can help to determine what feature should be delivered to auditory prostheses and therefore might guide the design of the future auditory prostheses for the patients speaking tonal languages.

As we all know pitch perception is the common basis of lexical tone and melody perception, but there are some significant differences between the pitch perception of harmonic or complex stimuli and the tone perception of voices, which can be attributed to the complicated envelop structure associated with the speech signals [2]. So the matter of lexical tone perception is more sophisticated than pitch perception, and there is no perfect model for tone perception just like that for pitch perception [3].

Some prior studies have clarified the effects of various acoustic features on tone perception. Although the change in fundamental frequency (F0) during phonation plays an important role in distinguishing the lexical tones, many other factors, such as harmonic structure, formant frequency, syllable duration, amplitude contour etc., also have some effects on tone perception [4].

In addition, extensive researches have been performed to clarify the relative contributions of the envelop and fine structure to lexical tone perception in the temporal and spectral domains. Fu [5] proved that temporal envelop cues can contribute to Mandarin tone recognition by normal-hearing listeners using the similar noise vocoder processing algorithm to Shannon [6] in cochlear implant simulation. In the similar way, more detailed experiments were conducted by Xu et al. [7], which demonstrated a trade-off between the spectral resolution (the number of simulated channels) and temporal features (low-pass cutoff frequencies used in envelop extractors). Moreover, Xu and Pfingst [8] applied "auditory chimaeras" algorithm, developed by Smith [9], to investigate the relative importance of the temporal envelop and fine structure to Mandarin monosyllables' tone perception. They claimed that tone recognition was consistent with the temporal fine structure of the chimearic stimuli, with an average of 90.8%, 89.5% and 84.5% percent scores for the 4-, 8-, and 16-band conditions respectively. Except for the temporal cues, a recent study by Kong and Zeng [10] systemically evaluated the effects of spectral cues on Mandarin tone recognition in quiet and in noise. They testified the relative significance of spectral fine structure (presented by harmonic stimuli) to spectral envelop (presented by whispered speech).

Although the importance of the fine structure to tone perception has been justified, one issue which has not been fully addressed in previous studies is which part of the fine structure cues could make the most contribution to lexical tone recognition. In this paper we adopted a modified "auditory chimaeras" method, in which the temporal fine structure from six different frequency bands of one sound was modulated respectively by the envelop of the other sound. By tone-recognition tests of six normal-hearing listeners the relative roles of the temporal envelop and different parts of temporal fine structure to Mandarin tone perception were investigated systematically.

## 2. Methods

### 2.1 Subjects

Six native Mandarin listeners (three females and three males) with normal hearing, aged from 23–28 years old, participated in the experiment. They were paid for their services. All listeners had pure-tone thresholds better than 15 dB HL at octave frequencies from 125 to 8000 Hz in both ears.

### 2.2 Stimuli and Signal Processing

The original speech test data consisted of six Mandarin Chinese single-vowel syllables (/a/, /o/, /e/, /i/, /u/, /ü/), and each vowel had four tone patterns (tone1 "–": flat tone; tone 2 "/": rising tone; tone 3 "√": falling and rising tone; tone

Manuscript received August 6, 2007.
Manuscript revised October 13, 2007.
[†]The authors are with ThinkIT Speech Lab, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, China.
a) E-mail: yanglin@hccl.ioa.ac.cn
DOI: 10.1093/ietisy/e91–d.2.371

4 "\": falling tone), resulting in a total of 24 tokens. A male and a female native speaker recorded the materials in a double-walled sound-treated booth. All speech test materials were digitized at a sampling rate of 44.1 kHz and stored in a 16-bit format. Tokens with nearly equal duration (about 0.6 s) were selected and were normalized to 0.6 s with Praat software to eliminate the influence of syllable duration, and they are down-sampled to 16 kHz and adjusted to a fixed root-mean-squared (rms) energy before sent to the "auditory chimaeras" synthesizer.

The chimearic stimuli were generated using the method somewhat different from Xu's [8]. As shown in Fig. 1, two sounds with different tone patterns, termed as Sound1 and Sound2, were used as inputs. Because the test data were all simple single-vowels, the pairs with the same and different phonemes were all permitted. Each sound was divided into several bands by band-pass filters according to the Greenwood map [11], which reflects the relationship between the central frequencies and positions in real cochlea. The temporal envelops of Sound1 and the fine structure waveforms of Sound2 in each frequency channel were extracted by Hilbert transform. The temporal envelops is the amplitude information provided by the Hilbert transform, while the temporal fine structure is the instantaneous phase information [9]. In the traditional "auditory chimaeras", the envelop of one sound modulated the temporal fine structure of the other sound for each frequency channel. Whereas in our study only one frequency channel of temporal fine structure of Sound2 remained, the others were replaced by band-limited white noises to smear the harmonics of the corresponding frequency band. The band-limited white noise was spectrally limited by the bandpass filters in the same frequency channel and its rms energy was equalized to that of the replaced fine structure waveform in order to keep balance the energy ratio among various bands. Finally the output of the noise synthesizer (i.e. the white noise modulated by the envelop) and the output of the fine structure synthesizer (i.e. the fine structure modulated by the envelop) were summarized together to get a synthesized chimaera sound.

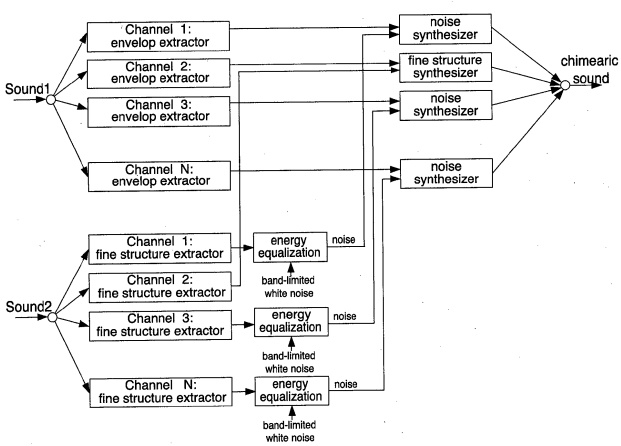


**Fig. 1** One example of the "auditory chimaeras", with the fine structure coming from the second band of Sound2.

In our study, the number of frequency bands was six in order to avoid the artifact of filter ringing induced by a large number of bands, as described by Zeng [12]. The overall frequency range was from 80 to 6000 Hz. The band-pass filters were fourth-order Butterworth filters and the cutoff frequencies were 80, 245, 533, 1032, 1897, 3397 and 6000 Hz.

### 2.3 Procedures

The synthesized stimuli were presented to the listeners at a comfortable and approximatively equal level to each other via TDH 39 headphones in a double-walled sound-treated booth. A graphical user interface (GUI) was used to present the stimuli and to collect the responses. Listeners were asked to identify the tone patterns which he or she heard and press the corresponding buttons. No feedback was given to subjects.

The total number of combinations of Sound1 and Sound2 was 432 (24 Sound1 × 18 Sound2), and the fine structure types of Sound2 were six. So there were 5184 tokens (2 voices × 432 combinations × 6 fine structure types) for each subject. All stimuli were divided into four sections evenly and each subject listened only one section once. The order of the tokens was randomized across listeners.

## 3. Results

Figure 2 depicts the average results of the tone recognition across six subjects as a function of the temporal fine structure types of Sound2 ("FineN" means extracting the temporal fine structure from the Nth frequency channel of Sound2). All responses are classified into three types: Type I (Solid squares) represents those consistent with the fine structure of Sound2; Type II (Solid triangles) represents those consistent with the envelop of Sound1; and Type III (Solid circles) represents those recognized as other tones.

Figure 2A shows the average results of all test stimuli. First of all, ANOVA revealed that the fine structure types of Sound2 had a significant effect on the tone recognition (Responses of type I: $F(5, 30) = 273.232$, $p<0.001$; Responses of type II: $F(5, 30) = 402.242$, $p<0.001$; Responses

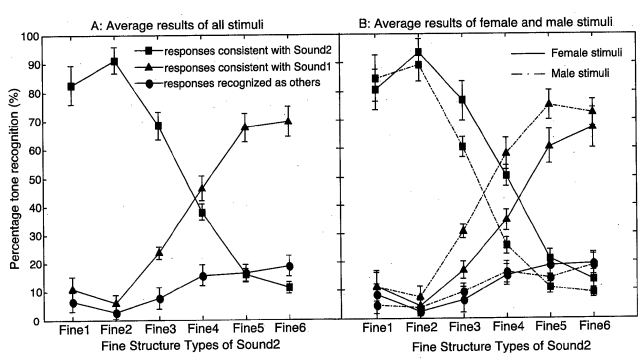


**Fig. 2** Panel A: The average results of all stimuli; Panel B: The average results of female and male stimuli presented separately (Solid lines: Female stimuli; Dashed Lines: Male stimuli).

**Table 1** The significant difference values caused by gender.

| | Significance values : p = ? (significant level:0.05) | | | | | |
|---|---|---|---|---|---|---|
| **Responses** | **Fine1** | **Fine2** | **Fine3** | **Fine4** | **Fine5** | **Fine6** |
| Type I | 0.934 | 0.133 | 0.001* | 0.001* | 0.001* | 0.156 |
| Type II | 0.387 | 0.142 | 0.001* | 0.001* | 0.001* | 0.009* |
| Type III | 0.104 | 0.359 | 0.260 | 0.623 | 0.045* | 0.804 |

* represents a significant difference.

of type III: F (5, 30) = 22.956, p<0.001). Second, from the Fig. 2A it is clear that in the case of Fine2 the tone-recognition responses belonging to Type I got the highest percentage 91.18%, and in the case of Fine1 and Fine3 the percentage remained relatively high, reaching 82.52% and 68.54% respectively, which meant that the low-frequency temporal fine structure in Sound2 dominated the tone of the chimearic sound. Whereas with the frequencies of the temporal fine structure increasing the responses of Type I decreased gradually and responses of type II increased correspondingly.

Figure 2B gives the average recognition results of female and male voices separately (Female stimuli: solid lines; Male stimuli: dashed lines). In Fig. 2B a parallel tendency of tone-recognition results of female and male voices can be observed. They all got the highest percentage of type I responses in the case of Fine2 and similarly a relatively dominant effect of the fine structure was noticeable in the cases of Fine1 and Fine3. However there were a few differences between the two kinds of stimuli. First, a multivariate ANOVA was performed with the types of responses as the dependent variables and the gender as the within-subjects factors for each fine structure type. Table 1 lists the significant difference values caused by gender. Noticeable in the case of Fine3, Fine4 and Fine5 the differences of tone-recognition are significant. For example the Type I responses of female stimuli were 76.74%, 50.23% and 20.91% for the fine structure type of Fine2, Fine3 and Fine4 respectively, while the corresponding percentage of male stimuli were 60.34%, 25.62% and 10.73%. Second, from Fig. 2B, in the case of Fine1 the percentage of Type I responses derived from the male stimuli were higher than that from the female stimuli; and in the case of Fine2 a reversed relationship was observed. But the differences were not significant and no more than 5%.

## 4. Discussions

Consistent with the previous studies [8], [10], the importance of the temporal fine structure cues to lexical tone perception was also confirmed by the "auditory chimaeras" strategy in our experiment. When temporal fine structure in the second frequency band of Sound2 was extracted and incorporated into the envelops of Sound1, 91.18% of the responses were consistent with the tone of Sound2, similar to Xu's result. Moreover, the importance of fine structure cues to tone recognition is consistent with the previous psychophysical findings on pitch perception of pure tones or complex tones [14]. Salient pitch perception is based on the

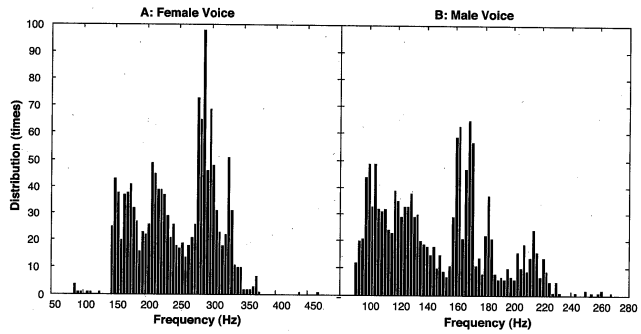


**Fig. 3** The distribution of F0 derived from the female and male voices.

fundamental frequency and resolved harmonics, which are mainly included in the fine structure cues.

Furthermore, from our experimental results it can be concluded that not all the frequency components of fine structure contributed to the Mandarin tone recognition. The fine structure in the second frequency band ranged from 245–533 Hz made the most significant contribution to the tone recognition, consistent with our previous study [13]. A trade off between the envelop and fine structure was achieved when the fine structure at the mid-frequency range 1000–2000 Hz was presented. And given the fine structure with high frequency components the envelop dominated the tone recognition, with 69.75% of responses consistent with the envelop in the case of Fine6, which was comparable to the percent correct of tone-recognition based on the tonal envelop cues alone in Fu's study [5]. Fu reported that 66.9%–80.8% of tone recognition was achieved by using four-channel CIS (Continuous Interleaved Sampler) simulation when the cutoff frequencies of envelop extractor varied from 50 to 500 Hz. Hence, the fine structure extracted from the second frequency band has the most significant effect on the Mandarin tone recognition, while the high-frequency components of fine structure contributes less to the tone perception.

The different results of female and male stimuli can be reasoned by the difference of fundamental frequencies. Figure 3 gives the histogram about the distribution of F0 extracted from all of the female and male stimuli on every 40-ms time window with 10-ms overlap by Praat software. F0 of the female voice mainly centered at the region of 150–350 Hz, while F0 of the male voice at 90–230 Hz. In the case of Fine1 the fine structure covered more information of F0 for the male voice than for the female voice, thus the male stimuli resulted in a relatively high percentage of responses consistent with the fine structure. In addition, except for the fundamental frequency cues, it was assumed that the low-frequency harmonics were also important to the tone perception. There is an analogy between the lexical tone perception and pitch perception. In Ritsma's study [14] on pitch perception, he found a so-called dominant region covered by the frequency components of the third, fourth, and fifth harmonics, and proposed that the pitch of a complex tone is determined by the dominant spectral region, where

the harmonics are obviously resolvable. For the male voice the second frequency band ranged from about 250–500 Hz was mainly covered by the second to fifth harmonics, and the percentage of responses consistent with the fine structure was even a little higher than that in the case of Fine1, where F0 was mainly included. So it can be ascertained that the low-frequency harmonics from second to fifth make a comparable contribution to the tone perception with the fundamental frequency cues, and there exists a similar dominant region for lexical tone perception to that for pitch perception.

From above it can be hypothesized that the application of the fundamental frequency and its low-frequency harmonics would improve the performance of auditory prostheses for the patients speaking tonal languages.

## 5. Conclusion

In this paper, Mandarin tone recognition was measured in a group of normal hearing listeners by chimearic stimuli with various types of fine structure cues. The main conclusions are:

(1) The temporal fine structure cues are important to Mandarin tone perception;

(2) The low-frequency components of the temporal fine structure make the most contribution to the lexical tone perception, while the high-frequency components show no great effect on the tone perception;

(3) The importance of the second to fifth harmonics were observed in our experiment, which mean that there exists a dominant region for the lexical tone perception similar to that of pitch perception.

## Acknowledgement

This work is partially supported by MOST (973 program2004CB318106), National Natural Science Foundation of China (10574140, 60535030), the National High Technology Research and Development Program of China (863 program, 2006AA010102, 2006AA01Z195). The authors wish to thank Dr. Junfeng Li from Japan Advanced Institute of Science and Technology for his helpful comments on the earlier version of this manuscript.

### References

[1] M.C. Lin, "The acoustic characteristics and perceptual cues of tones in standard Chinese," Chinese Yuwen, vol.204, pp.182–193, 1988.
[2] T. Green, A. Faulkner, and S. Rosen, "Spectral and temporal cues to pitch in noise-excited vocoder simulation of continuous-interleaved-sampling (CIS) cochlear implants," Speech, Hearing and Language: work in progress, vol.13, pp.23–38, 2001.
[3] R. Meddis and L. O'Mard, "A unitary model of pitch perception," J. Acoust. Soc. Am., vol.102, pp.1811–1820, 1997.
[4] D.H. Whalen and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica, vol.49, pp.25–47, 1992.
[5] Q.J. Fu, F.G. Zeng, R.V. Shannon, and S.D. Soli, "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am., vol.104, pp.505–510, 1998.
[6] R.V. Shannon, F.G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," Science, vol.270, pp.303–304, 1995.
[7] L. Xu, "Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses," J. Acoust. Soc. Am., vol.112, pp.247–258, 2002.
[8] L. Xu and B.E. Pfingst, "Relative importance of temporal envelope and fine structure in lexical-tone recognition (L)," J. Acoust. Soc. Am., vol.114, pp.3024–3027, 2003.
[9] Z.M. Smith, B. Delgutte, and A.J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," Nature, vol.416, pp.87–89, 2002.
[10] Y.Y. Kong and F.G. Zeng, "Temporal and spectral cues in Mandarin tone recogntion," J. Acoust. Soc. Am., vol.120, pp.2830–2840, 2006.
[11] D.D. Greenwood, "A cochlear frequency-position function for several species–29 years later," J. Acoust. Soc. Am., vol.87, pp.2592–2605, 1990.
[12] F.G. Zeng, K. Nie, S. Liu, G. Stickney, E.D. Rio, and Y.Y. Kong, "On the dichotomy in auditory perception between temporal envelop and fine structure cues (L)," J. Acoust. Soc. Am., vol.116, pp.1351–1354, 2004.
[13] L. Yang, J. Zhang, and Y. Yan, "Contributions of temporal fine structure cues to Chinese speech recognition in cochlear implant simulation," Interspeech, pp.386–389, 2007.
[14] R.J. Ritsma, "Frequency dominant in the perception of the pitch of complex sounds," J. Acoust. Soc. Am., vol.42, pp.191–198, 1967.