

Published in final edited form as:

J Acoust Soc Am. 2003 December ; 114(6 Pt 1): 3024–3027.

Relative importance of temporal envelope and fine structure in lexical-tone perception (L)

Li Xu^{a)}

School of Hearing, Speech and Language Sciences, Ohio University, Grover Center, Athens, Ohio 45701, and Kresge Hearing Research Institute, Department of Otolaryngology, University of Michigan, Ann Arbor, Michigan 48109

Bryan E. Pfingst

Kresge Hearing Research Institute, Department of Otolaryngology, University of Michigan, Ann Arbor, Michigan 48109

Abstract

The relative importance of temporal envelope and fine structure in speech and music perception was investigated by Smith *et al.* [Nature (London) 416, 87–90 (2002)] using "auditory chimera" in which the envelope from one sound was paired with the fine structure of another. Smith *et al.* found that, when 4 to 16 frequency bands were used, recognition of English speech was dominated by the envelope, whereas recognition of melody was dominated by the fine structure. In the present study, Mandarin Chinese monosyllables were divided into 4, 8, or 16 frequency bands and the fine structure and envelope of one tone pattern were exchanged with those of another tone pattern of the same monosyllable. Five normal-hearing native Mandarin Chinese speakers completed a four-alternative forced-choice tone-identification task. In the vast majority of trials, subjects based their identification of the monosyllables on the fine structure rather than the envelope. Thus, the relative importance of envelope and fine structure for lexical-tone perception resembled that for melody recognition rather than that for English speech recognition. Delivering fine-structure information in cochlear implant stimulation could be particularly beneficial for lexical-tone perception.

I. INTRODUCTION

The time waveform of signals can be described mathematically as slowly varying envelope (modulation) and rapidly varying fine-time structure (carrier). In a recent report, Smith *et al.* (2002) constructed a set of acoustic stimuli, each having the envelope of one sound and the fine structure of another. They termed the hybrid sounds "auditory chimera." The operation was accomplished with a Hilbert transform that allowed the investigators to extract the envelope and the fine structure of the sounds. For example, a speech–speech chimera can be synthesized so that it contains speech information from one sentence in the envelope and the speech information from another sentence in the fine structure. The auditory chimeras provide a way to study the relative importance of the envelope and the fine structure in speech perception and pitch perception. Knowledge of the relative importance of the envelope and the fine structure might help in determining what stimulus features should be delivered to auditory prostheses and therefore might guide the design of future auditory prostheses.

Smith *et al.* (2002) found that in speech–speech chimeras with 4 to 16 frequency bands, the words represented in the envelope were identified correctly much more frequently than the words represented in the fine structure. This result was consistent with the fact that English-

^{a)}Electronic mail: xul@ohio.edu

speaking cochlear-implant users have achieved high speech recognition scores using the continuous interleaved sampling (CIS) stimulation strategy (Wilson *et al.*, 1991), in which the speech envelope is delivered to a limited number of frequency channels and the fine structure is absent. Also, Shannon *et al.* (1995) and others have shown that normal-hearing subjects can achieve good word recognition (.85% correct) using a four-band noise vocoder simulating the cochlear implant stimulation.

In contrast to the results with speech–speech chimeras, Smith *et al.* (2002) found that when melody–melody chimeras were synthesized to exchange the envelope and fine structure of two different melodies, subjects almost always reported hearing the melody represented by the fine structure, with up to 16 frequency bands. Data for number of frequency bands .16, should be interpreted with caution due to the artifacts caused by ringing of the very narrow digital filters, as pointed out by Zeng *et al.* (2003).

The question that we address in the present study is the relative importance of the envelope and the fine structure for lexical-tone perception. In tone languages, such as Mandarin Chinese, Cantonese, Thai, and Vietnamese, the tonality of syllables conveys lexical meaning. The number and the pattern of tones vary among different tone languages. In Mandarin Chinese, a language spoken by more people than any other language, there are four tone patterns (tone 1 through tone 4) defined by the contour of the fundamental frequency: (1) flat and high; (2) rising; (3) falling and then rising; and (4) falling (see Fig. 1 in Xu *et al.*, 2002). Typographically, these tone patterns are usually represented as “[unk],” “/,” “V,” and “\,” respectively.

Does the relative importance of the envelope and the fine structure for lexical-tone perception resemble that for English speech recognition because tone perception is an aspect of speech perception? Or, does it resemble that for music recognition because tone patterns and music both show variations in the fundamental frequency and harmonics? The hypothesis that we set forth for the present study was that the relative importance of the envelope and the fine structure for tone recognition resembles that for music recognition. We used 4-, 8-, and 16-frequency-band auditory chimera (Smith *et al.*, 2002) to test this hypothesis. In the present experiments, chimeric Mandarin syllables that had one tone pattern represented in the envelope and a different tone pattern represented in the fine structure were presented to native Mandarin-speaking subjects for identification. The range of 4 to 16 frequency bands was chosen (1) to provide a situation that represents the number of frequency channels that a current multichannel cochlear prosthesis can provide; (2) to avoid the digital filter ringing which can occur when a large number of frequency bands (e.g., >16) with very narrow filter bandwidth are used (Zeng *et al.*, 2003); and (3) to overcome the natural recovery of narrow-band envelopes by cochlear filtering when the number of bands is small (e.g., 1 or 2 bands) (Ghitza, 2001). The problem with using a small number of bands was shown in a counterintuitive result in signal processing of speech signals by Ghitza (2001). He first bandpass filtered a speech signal by a critical-band filter and then low-pass filtered the envelope at 16 Hz. When the speech signal with a smoothed envelope was passed through a critical-band filter again, he found that the output had a nonsmoothed envelope that was similar to that of the original signal (Ghitza, 2001). This model illustrates a potential problem that would result if one or two wide bands were used.

II. METHODS

A. Subjects

Five normal-hearing native-Mandarin-Chinese speakers (three females and two males) were recruited from the student and staff population at the University of Michigan. Subjects ranged from 36 to 41 years of age (37.662.1, mean and s.d.). All subjects had received at least a college-level education in China. The pure-tone air-conduction thresholds for all subjects were <20 dB HL at octave frequencies between 250 and 8000 Hz. The use of human subjects in this study

was reviewed and approved by the University of Michigan Medical School Institutional Review Board.

B. Original and chimeric speech materials

The original speech test materials consisted of ten Mandarin monosyllables. The *pinyin* (i.e., the phonemic spelling system for Mandarin Chinese) of these syllables were "fu," "ji," "ma," "qi," "wan," "xi," "xian," "yan," "yang," and "yi," and each syllable had four tone patterns. The resulting 40 combinations of consonant, vowel, and tone were real words in Chinese. A male and a female speaker were used to record the speech materials. Tokens in which the durations of the four tone patterns of each syllable were equal were selected (Xu *et al.*, 2002). All speech test materials were digitized at a sampling rate of 22 050 Hz and stored in a 16-bit format. These equal-duration speech materials were used in a previous tonal-speech perception study (Xu *et al.*, 2002) and were used in the present study to create chimeric speech materials as described below.

The chimeric stimuli were generated using the methods described by Smith *et al.* (2002). Briefly, two sounds, in this case a single monosyllable with two different tone patterns (e.g., tone 1 and tone 2 of "ma"), were divided into a number of frequency bands (4, 8, or 16) whose center frequencies were chosen to reflect equal length along the basilar membrane (Greenwood, 1990). The overall frequency range for chimera synthesis was from 80 to 8820 Hz. The bandpass filters used in the present study were identical to the ones in Smith *et al.* (2002). In the case of 8 bands, the cutoff frequencies were 80, 205, 405, 724, 1236, 2055, 3366, 5463, and 8820 Hz, and the range of frequencies common to adjacent bands was about 30 Hz. Hilbert transforms were then used to extract the envelope and the fine structure in each band. Finally, the envelopes and fine structures of the two syllables with the same phonemes (e.g., "ma 1" and "ma 2") were exchanged to create two chimeric stimuli (e.g., one with envelope of "ma 1" and fine structure of "ma 2" and the other with envelope of "ma 2" and fine structure of "ma 1"). In a similar fashion, a total of 12 chimeric stimuli were generated for the four tone patterns of each syllable. All signal processing was performed in MATLAB (Mathworks, Natick, MA).

C. Procedures

A custom graphical user interface (GUI) was created, using MATLAB, to present the stimuli and to collect the responses. The GUI displayed on an LCD monitor the typographical representations of the four tone patterns of a syllable (i.e., "[unk]," "ˊ," "ˋ," "ˊˋ") and the four associated Chinese characters. The program presented to the loudspeaker one of the 12 chimeric stimuli for that particular syllable. The subject's task was to select, using a computer mouse, the response button associated with the tone pattern of the sound that he or she heard. After each response, the GUI would refresh the screen, display a new set of four Chinese characters with four tone patterns, and present the next acoustic stimulus. The test order of the number of bands (4, 8, or 16) was randomized. Each test contained 240 randomized stimulus presentations (2 voices 310 syllables 312 chimeras) and was administered five times. A total of 3600 responses (3 numbers of bands 3240 presentations 35 times) was collected from each subject.

The stimuli were presented through a loudspeaker mounted 1 m away from the subject at 0° azimuth inside an Acoustic Systems (model RE2 242S) double-walled sound-attenuating booth. Stimulus level was roved randomly between 50 to 70 dB A in 5-dB steps.

III. RESULTS

There was no difference in tone recognition between male and female voices and therefore results reported below were pooled from results for both voices. Data obtained with 8 frequency bands are shown in Fig. 1. Data for 4 and 16 bands were very similar to those of 8 bands and were thus omitted from the figure. In Fig. 1, there are 12 histograms representing 12 combinations of fine structure (column-wise) and envelope (row-wise). Each histogram plots the mean and standard deviation (s.d.) of the percentages of the tone-recognition responses across the group of five subjects. Each bar represents a percentage of the responses as tone 1, 2, 3, or 4 as indicated by the labels in the lower left corner. The subjects almost always reported the tone patterns that were consistent with the fine structures of the stimuli, regardless of the envelopes.

Figure 2 summarizes the tone-recognition performance using 4, 8, and 16 frequency bands. Tone-recognition responses that were consistent with the fine structure of the stimuli are shown with the black bars using the left ordinate. Over the range of numbers of bands tested, the vast majority of the responses of the subjects was consistent with the fine structure of the stimuli. On average, 90.8%, 89.5%, and 84.5% of the responses were consistent with the fine structure for 4, 8, and 16 bands, respectively. The differences among the percentages were small but statistically significant (ANOVA, $p < 0.05$). Significant differences ($p < 0.05$) were further found to be between 4 and 16 bands and between 8 and 16 bands in *post hoc* tests with Bonferroni correction. On the other hand, only 4.3%, 5.0%, and 8.9% of the responses were consistent with the envelope of the stimuli for 4, 8, and 16 bands, respectively, as shown by the open bars using the right ordinate in Fig. 2. A small fraction of the responses was not consistent with either the envelope or the fine structure (Figs. 1 and 2). Such errors or confusions occurred at an average of 4.9%, 5.5%, and 6.6% of the responses for 4, 8, and 16 frequency bands, respectively (Fig. 2). Noticeably, the confusions between tone 2 and tone 3 occurred most often (Fig. 1).

IV. DISCUSSION AND CONCLUSIONS

Our results on lexical-tone recognition using auditory chimera were largely consistent with the results of Smith *et al.* (2002) on melody recognition for the range of numbers of frequency bands tested. Lexical-tone recognition depended on the fine structure but not on the envelope when the number of frequency bands was between 4 and 16. Lin (1988) studied tone recognition of 14 Mandarin-speaking subjects with synthesized Mandarin syllables in which the envelopes of tone patterns 1 through 4 were used to amplitude modulate the fine structure of syllables with tone 1. Those sounds could be considered equivalent to the chimeric sounds used in this study except that the number of frequency bands was fixed at 1 in that study. Lin's (1988) results showed that about 92% of the time the subjects reported that they heard tone 1, that is, the tone that contained the fine structure. Those data were comparable to our data with the number of frequency bands between 4 and 16 (Fig. 2). However, as pointed out by Zeng *et al.* (2003), when only one band is used, the ability of the cochlear filters to recover the narrow-band envelopes of the original signals (Ghitza, 2001) might hamper the interpretation of relative importance of the envelope and fine structure.

In the present study, syllables of equal duration were used to create chimeric sounds. It has been shown that syllable duration can be a strong cue for *isolated* syllables (Whalen and Xu, 1992; Xu *et al.*, 2002). However, syllable duration might not be a reliable cue for everyday speech (Xu *et al.*, 2002). Duanmu (2002) has summarized a few studies on duration and tones and has found that the differences in duration for Mandarin tones are rather small, all within 10% among different tones. The use of equal-duration syllables in the present study was a requirement of the chimera technique. In addition, it removed a potential cue for tone

recognition, thus eliminating a confounding variable (duration) and allowing us to focus on the main issue of our study, which was lexical-tone perception.

Multichannel cochlear prostheses have achieved high levels of speech perception for English in patients with profound hearing loss (Skinner *et al.*, 1994). However, music perception is still generally poor in these patients (Gfeller *et al.*, 1997, 2002). Current cochlear prostheses can deliver up to 22 frequency channels. However, research data have suggested that the patients cannot functionally use more than 8 frequency channels (Fishman *et al.*, 1997; Fu *et al.*, 1998; Friesen *et al.*, 2001). The present study tested 4 to 16 frequency bands, which should have encompassed the number of functional channels that the majority of cochlear implant patients can use. Current speech processing strategies in the cochlear prostheses emphasize the presentation of envelope information. Our results confirm that fine structure is relatively more important for pitch perception than is envelope (Smith *et al.*, 2002). Therefore, modifying cochlear implant processors to deliver fine-structure information might improve the pitch perception of implant patients, thus facilitating music perception and lexical-tone perception.

One potential way to deliver fine-structure information might be to use analog stimulation or variable pulse rates. However, it is uncertain whether or not patients would be able to use the pitch information in the fine structure of the electrical signal. Some studies have indicated that subjects can use rate-pitch information only up to about 300 pps, or 300 Hz for sinusoids (Shannon, 1983; McKay *et al.*, 1994; Zeng, 2002), although others have reported that some subjects can discriminate rate-pitch changes up to 1000 pps (Townshend *et al.*, 1987; Pijl and Schwarz, 1995). A limited upper boundary of temporal code in electrical hearing would make it difficult for patients to utilize fine structure information. Future research is warranted to determine why some patients seem to have better rate-pitch perception than others and to identify ways to overcome the apparent 300-Hz boundary.

ACKNOWLEDGMENTS

We are grateful to Dr. Bertrand Delgutte for providing the software to create the auditory chimeras and for many useful discussions. Dr. Fan-Gang Zeng, an anonymous reviewer, and Associate Editor, Dr. Peter Assmann, provided useful comments on earlier versions of the manuscript. This research was supported by NIH Grants F32-DC00470 and RO1-DC03808.

References

1. Duanmu S. . The Phonology of Standard Chinese. Oxford University Press; Oxford: 2002.
2. Fishman KE, Shannon RV, Slattery WH. "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor". *J Speech Lang Hear Res* 1997;40:1201–1215. [PubMed: 9328890]
3. Friesen L, Shannon RV, Baskent D, Wang X. "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants". *J Acoust Soc Am* 2001;110:1150–1163. [PubMed: 11519582]
4. Fu QJ, Shannon RV, Wang X. "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing". *J Acoust Soc Am* 1998;104:3586–3596. [PubMed: 9857517]
5. Gfeller K, Witt S, Woodworth G, Mehr MA, Knutson J. "Effects of frequency, instrumental family, and cochlear implant type on timbre recognition and appraisal". *Ann Otol Rhinol Laryngol* 2002;111:349–356. [PubMed: 11991588]
6. Gfeller K, Woodworth G, Robin D, Witt S, Knutson J. "Perception of rhythmic and sequential pitch patterns by normally hearing adults and adult cochlear implant users". *Ear Hear* 1997;18:252–260. [PubMed: 9201460]
7. Ghitza O. "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception". *J Acoust Soc Am* 2001;110:1628–1640. [PubMed: 11572372]

8. Greenwood DD. "A cochlear frequency-position function for several species—29 years later". *J Acoust Soc Am* 1990;87:2592–2605. [PubMed: 2373794]
9. Lin MC. "The acoustical properties and perceptual characteristics of Mandarin tones". *Zhongguo Yuwen* 1988;3:182–193.
10. McKay CM, McDermott HJ, Clark GM. "Pitch percepts associated with amplitude-modulated current pulse trains in cochlear implantees". *J Acoust Soc Am* 1994;96:2664–2673. [PubMed: 7983272]
11. Pijl S, Schwarz DWF. "Melody recognition and musical interval perception by deaf subjects stimulated with electrical pulse trains through simple cochlear implant electrodes". *J Acoust Soc Am* 1995;98:886–895. [PubMed: 7642827]
12. Shannon RV. "Multichannel electrical stimulation of the auditory nerve in man. I. Basic psychophysics". *Hear Res* 1983;11:157–189. [PubMed: 6619003]
13. Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. "Speech recognition with primarily temporal cues". *Science* 1995;270:303–304. [PubMed: 7569981]
14. Skinner MW, Clark GM, Whitford LA, Seligman PM, Staller SJ, Shipp DB, Shallop JK, Everingham C, Menapace CM, Arndt PL, Antogenelli T, Brimacombe JA, Pijl S, Daniels P, George CR, McDermott HJ, Beirer AL. "Evaluation of a new Spectral Peak coding strategy for the Nucleus 22 Channel Cochlear Implant System". *Am J Otol* 1994;15(Suppl 2):15–27. [PubMed: 8572106]
15. Smith ZM, Delgutte B, Oxenham AJ. "Chimaeric sounds reveal dichotomies in auditory perception". *Nature (London)* 2002;416:87–90.
16. Townshend B, Cotter N, Van Compernelle D, White RL. "Pitch perception by cochlear implant subjects". *J Acoust Soc Am* 1987;82:106–115. [PubMed: 3624633]
17. Whalen DH, Xu Y. "Information for Mandarin tones in the amplitude contour and in brief segments". *Phonetica* 1992;49:25–47. [PubMed: 1603839]
18. Wilson BS, Finley CC, Lawson DT, Wolford RD, Eddington DK, Rabinowitz WM. "Better speech recognition with cochlear implants". *Nature (London)* 1991;352:236–238.
19. 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]
20. Zeng F-G. "Temporal pitch in electric hearing". *Hear Res* 2002;174:101–106. [PubMed: 12433401]
21. Zeng F-G, Nie K-B, Stickney G, Liu S, Rio ED, Kong Y-Y, Chen H-B. "Facts and artifacts in auditory chimaeras". *Assoc Res Otolaryngol Abstr* 2003;26:213.

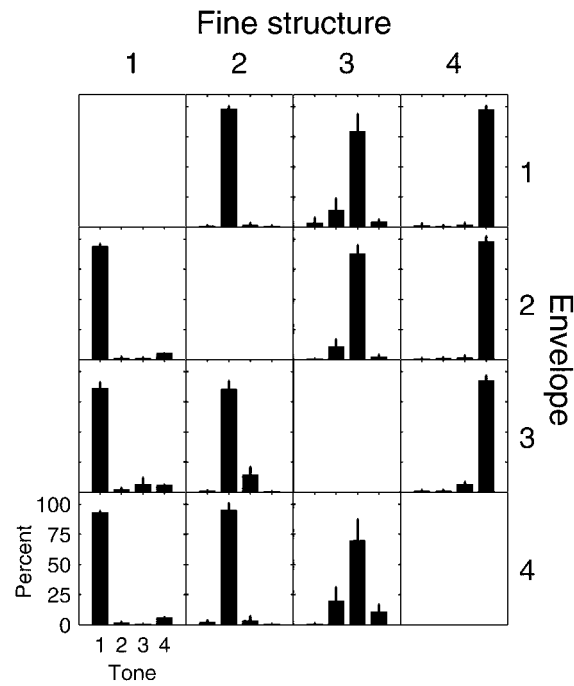


FIG. 1. Means and standard deviations (s.d.'s) of the percentages of the tone-recognition responses across the group of five subjects using 8-band auditory chimera. The 16 small panels represent 16 combinations of four fine structures (column-wise) and four envelopes (row-wise). When the fine structure and envelope were from the same tone patterns, the stimuli were the original sounds and were not tested for tone recognition, as plotted in gray. Results from the remaining 12 combinations of fine structures and envelopes are shown in a histogram format in which the means and s.d.'s of the percentages of the responses as tone 1, 2, 3, or 4 are represented by the bars and the whiskers from left to right, as labeled in the lower left panel.

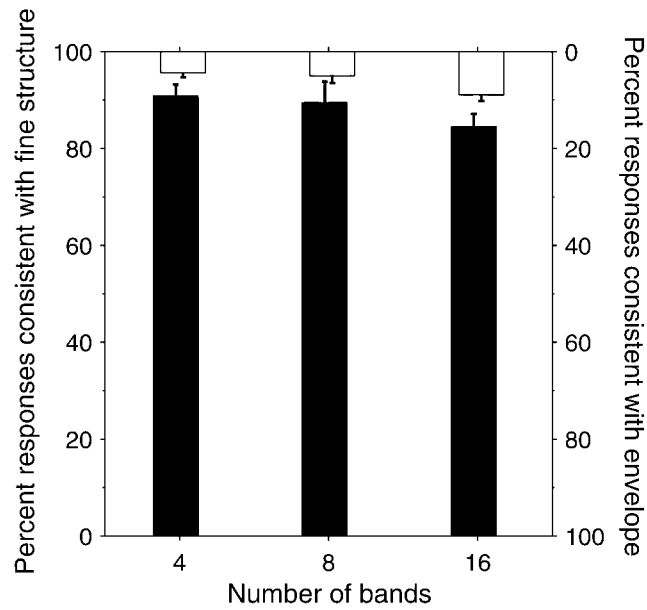


FIG. 2.. Means and s.d.'s of tone-recognition performance for 4, 8, and 16 frequency bands across the group of five subjects. The percentage of responses that were consistent with the fine structure of the chimeric stimuli are shown with black bars using the left ordinate. The percentage of responses that were consistent with the envelope of the chimeric stimuli are shown with open bars using the right ordinate. The error bars indicate the s.d.'s.