Temporal and spectral cues in Mandarin tone recognition[a)]

Ying-Yee Kong[b)]
Hearing and Speech Research Laboratory, Department of Cognitive Sciences, University of California-Irvine, Irvine, CA 92697

Fan-Gang Zeng
Hearing and Speech Research Laboratory, Departments of Anatomy and Neurobiology, Biomedical Engineering, Cognitive Sciences, and Otolaryngology-Head and Neck Surgery, University of California-Irvine, Irvine, CA 92697

(Received 19 October 2005; revised 24 July 2006; accepted 6 August 2006)

This study evaluates the relative contributions of envelope and fine structure cues in both temporal and spectral domains to Mandarin tone recognition in quiet and in noise. Four sets of stimuli were created. Noise-excited vocoder speech was used to evaluate the temporal envelope. Frequency modulation was then added to evaluate the temporal fine structure. Whispered speech was used to evaluate the spectral envelope. Finally, equal-amplitude harmonics were used to evaluate the spectral fine structure. Results showed that normal-hearing listeners achieved nearly perfect tone recognition with either spectral or temporal fine structure in quiet, but only 70%–80% correct with the envelope cues. With the temporal envelope, 32 spectral bands were needed to achieve performance similar to that obtained with the original stimuli, but only four bands were necessary with the additional temporal fine structure. Envelope cues were more susceptible to noise than fine structure cues, with the envelope cues producing significantly lower performance in noise. These findings suggest that tonal pattern recognition is a robust process that can make use of both spectral and temporal cues. Unlike speech recognition, the fine structure is more important than the envelope for tone recognition in both temporal and spectral domains, particularly in noise. [DOI: 10.1121/1.2346009]

PACS number(s): 43.66.Ts, 43.71.Ky [AJO] Pages: 2830–2840

I. INTRODUCTION

Acoustic characteristics and their importance for tone recognition have been demonstrated in many tonal languages, including Cantonese (Gandour, 1984), Mandarin (Liang, 1963; Lin, 1988; Whalen and Xu, 1992), and Thai (Abramson, 1978). Results have shown that lexical tones in these languages are distinguished primarily by the change in fundamental frequency (F0) during phonation, although other acoustic properties, such as syllable duration and amplitude contour, also convey tonal information (Whalen and Xu, 1992; Fu *et al.*, 1998; Fu and Zeng, 2000).

Liang (1963) systematically studied the role of spectral cues in Mandarin tone recognition by low- and high-pass filtering the speech stimuli with different cutoff frequencies. He reported that perfect tone recognition can be achieved either with the presence of F0 information with low-pass filtering at 300 Hz or with the harmonic structure with high-pass filtering at 300 Hz. However, in another experimental condition where he only preserved the first formant frequency (300–1200 Hz), performance was still maintained at the 83% level, suggesting the use of formant information in voice pitch perception. Because F0 could still be extracted from harmonics in the first formant frequency range, Liang also used whispered speech, which contained neither F0 nor harmonic fine structure, to evaluate the contribution of formant information to voice pitch perception. He found that Mandarin tone recognition could be achieved with 60%–80% accuracy (see also Jensen, 1958). Other studies (e.g., Miller, 1961; Howie, 1976), however, showed that tonal contrast was not well preserved with whispered speech. Miller (1961) reported only approximately 40% correct tone recognition in Vietnamese.

In addition to the F0, other information such as temporal envelope was shown to contribute to tone recognition in Mandarin (Fu *et al.*, 1998; Fu and Zeng, 2000; Xu *et al.*, 2002; Whalen and Xu, 1992). Most studies on temporal cues for Mandarin tone recognition concentrated mainly on the envelope cues (e.g., Lin, 1988; Whalen and Xu, 1992; and Fu *et al.*, 1998; Fu and Zeng, 2000; Xu *et al.*, 2002). For Mandarin speech, the amplitude contour cues refer to the slow changes in amplitude and such cues were shown to co-vary with the change of F0 over time in Mandarin. The periodicity cue refers to the fluctuation at a rate, which is directly related to the change in fundamental frequency. Systematically manipulating the amount of the temporal envelope cues, Fu and Zeng (2000) showed that while the duration cues play a minor role in identifying tones, the amplitude contour and periodicity cues contributed significantly to Mandarin tone recognition. Using the noise vocoder processing algorithm described in (Shannon *et al.* 1995), Fu and co-workers (1998) reported that 60%–80% correct tone

[a)]Portions of this work were presented at the 147th Meeting of the Acoustical Society of America, New York, 2004.
[b)]Author to whom correspondence should be addressed. Electronic mail: yingyeekong@gmail.com

recognition in quiet can be achieved, independent of the number of frequency bands. With the amplitude contour cue alone, tone recognition was at 65% correct in the 1-band condition. The additional periodicity cue, however, improved the tone recognition by 10–20 percentage points. The reliance on periodicity cues for tone recognition was also demonstrated by Xu *et al.* (2002), where they reported a trade-off between the amount of spectral information (number of frequency bands) and the low-pass cutoff frequency of temporal envelope extraction. As the low-pass cutoff frequency increased, the number of frequency bands required to achieve the similar tone recognition performance decreased. Xu *et al.* (2002) also created frequency-modulated (FM) pulse trains with F0 contours that mimicked the Mandarin tone contours while eliminating the temporal envelope cue. They reported that without the temporal envelope and duration cues, recognition of vocoder processed FM patterns was poor.

A recent study by Xu and Pfingst (2003) investigated the relative contribution of temporal envelope and fine structure (Hilbert definition) to Mandarin tone recognition. Using the chimearic processing algorithm developed by Smith *et al.* (2002), Xu and co-workers further supported the differential contribution of temporal envelope and fine structure to speech and pitch perception (Smith *et al.*, 2002; Nie *et al.*, 2005). Smith *et al.* found that the temporal envelope cues are sufficient to support speech recognition in quiet, but temporal fine structure is needed for pitch perception. Xu *et al.* (2003) claimed that pitch perception is the common basis for both lexical tone perception and melody recognition by showing that while Mandarin word recognition was determined by the temporal envelope cues, tone recognition was consistent with the temporal fine structure of the chimearized stimuli, with an average of 90.8%, 89.5%, and 84.5% correct for the 4-, 8-, and 16-band conditions, respectively.

The main goal of this study was to systematically investigate the relative contribution of envelope and fine structure cues to voice pitch perception. First, we extended previous studies by examining the envelope and fine structure cues necessary for tone recognition in both temporal and spectral domains. Second, we evaluated the four acoustic cues for Mandarin tone recognition in both quiet and in noise conditions. To our knowledge, no studies have been done on whether reliable Mandarin tone recognition can be achieved in the presence of noise. There are reasons to believe that the effect of noise on tone recognition differs with the type of cue. The terms temporal envelope and fine structure have been defined loosely in the literature. From the speech perception point of view, Rosen (1992) divided temporal envelope cues into three categories, depending on the rate of amplitude fluctuation: (1) envelope (2–50 Hz); (2) periodicity (50–500 Hz), and (3) fine structure (500–10 000 Hz). However, the formal mathematical definition, based on the Hilbert transform (Hilbert, 1912), the temporal fine structure is defined as the instantaneous phase information of the signal. In the present study, the temporal envelope cues were defined as follows: duration and amplitude contours (amplitude fluctuation between 2–50 Hz) and periodicity cues (50–500 Hz). The temporal fine structure was defined as the instantaneous phase information in the signals provided by the Hilbert transform (Hilbert, 1912). As for spectral cues, spectral envelope was defined as the general shape of the spectrum, a smooth curve that passes though the peaks of the spectrum (Hartmann, 1997). For speech stimuli, the spectral envelope represents the filtering properties (or the resonance) in the vocal tract, such as the formant structure of a speech signal. Spectral fine structure, which was also considered as the spectral details, was defined as the detailed frequency components in the power spectrum. The term "spectral fine structure" has been widely used to refer to the spectral details in the power spectrum and we will continue to use this term in this paper. For speech signal, the spectral fine structure represents the source information, which has a harmonic structure for a voiced source and noise structure for an unvoiced source. It should be noted that there are differences between temporal fine structure and spectral fine structure. Consider a simple case of a sinusoidally amplitude-modulated stimulus with a 1000 Hz carrier modulated at a 40 Hz rate; the temporal fine structure is the 1000 Hz carrier of the amplitude-modulated signal, but the power spectrum is composed of three components at 960, 1000, and 1040 Hz.

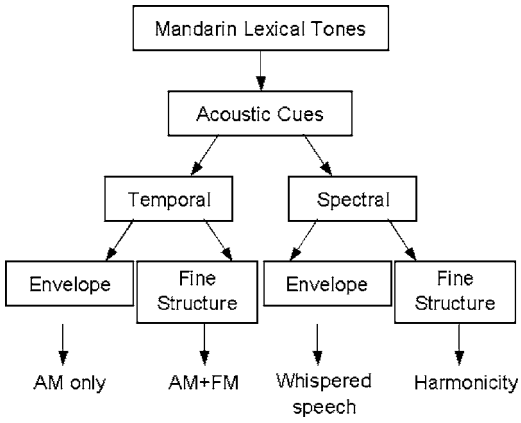

FIG. 1. A block diagram depicts the acoustic cues in the spectral and temporal domains which were evaluated for Mandarin tone recognition.

Figure 1 outlines the four types of acoustic cues that were examined in the study and the stimuli that were used to deliver this information for each experimental condition. We divided the acoustic information into envelope and fine structure in both the temporal and spectral domain. First, for temporal envelope cues, we used the noise vocoder type of processing (Shannon *et al.*, 1995) to manipulate the relative distribution of temporal and spectral information in the speech stimuli. The amount of the temporal envelope information was manipulated by changing the cutoff frequency of the envelope extraction filters, i.e., 50 or 500 Hz in the present study. Previous studies demonstrated that temporal envelope cues contributed significantly to tone recognition in quiet. We hypothesized that the envelope information alone would not be sufficient for tone recognition in noise, due to the nonsaliency of amplitude modulation for pitch perception and also the reduced detectability of amplitude modulation in noise. Second, speech stimuli processed with a novel algorithm developed by Nie *et al.* (2005) were used to assess the importance of temporal fine structure for Mandarin tone recognition. Unlike the noise vocoder processing, which pro-

vided only the amplitude modulation, Nie *et al.* derived additional slowly-varying frequency modulations around the center frequency of each analysis filter and amplitude modulated them with the extracted temporal envelope. We hypothesized that the additional frequency modulation in the speech signal would produce better Mandarin tone recognition. Third, to present only the spectral envelope information in the absence of harmonicity cues, two sets of stimuli were created: naturally recorded whispered speech and synthesized whispered speech using a Linear Predictive Coding (LPC) technique. Lastly, the contribution of spectral fine structure to Mandarin tone recognition was evaluated using stimuli that were the residue of a 14-order LPC processing. While the spectral envelope was presumably flat in this set of stimuli, harmonicity cues were preserved. Based on previous psychophysical findings, we hypothesized that robust voice pitch perception can be determined by the harmonicity cues alone and that spectral envelope cues without the harmonic structure only convey limited voice pitch information.

II. MANDARIN TONE RECOGNITION WITH TEMPORAL ENVELOPE CUES

A. Methods

1. Subjects

Five native Mandarin speakers (S1–S5) participated in the experiment. All of them were females, with ages ranging from 22 to 33. All subjects had normal hearing sensitivity below 20 dB HL at octave frequencies between 125 and 8000 Hz bilaterally.

2. Stimuli

The test material included 100 Mandarin words consisting of 25 syllables with four lexical tones for each syllable (Wei *et al.*, 2004). The original stimuli were recorded from one adult male and one adult female Beijing Mandarin speaker, resulting in a total of 200 words for each test condition. Recording was carried out in a double-walled sound-treated booth and the stimuli were digitally recorded through a preamplifier to the computer at a sampling rate of 22 050 Hz using COOL EDIT PRO (2.0) software. All stimulus waveforms were then adjusted to the same root-mean-squared (rms) amplitude.

Speech-shaped noise was used for the noise conditions. Two samples of speech-shaped noise were created, one for the male speaker and one for the female speaker, by summing the speech tokens (100 words) from either the male or female speaker and then processing them using the 10-order LPC to extract the long-term spectrum of the summed stimuli. While the rms amplitude of speech waveforms was fixed, the rms of the noise waveform varied to achieve the desired signal-to-noise ratios (−10, −5, 0, +5, and +10 dB) tested in this experiment.

The original speech stimuli, combined with the original speech-shaped noise in the noise conditions (male noise combined with the male speech tokens, and female noise with the female speech tokens), were subjected to further processing to contain either temporal envelope cues alone or additional coarse spectral cues via a cochlear implant simulation program similar to that used in Shannon *et al.* (1995). The original broadband (80–8800 Hz) stimuli were first pre-emphasized by a first-order high-pass Butterworth filter at 1200 Hz. The pre-emphasized stimuli were then divided into 1 or 8 frequency bands using sixth-order elliptical bandpass filters. The cutoff frequencies of each band were determined by approximating equal cochlear distance for each band according to the Greenwood map (1990). The corner frequencies for the 1-band processing condition were 80 and 8800 Hz and for the 8-band processing condition were 80, 220, 440, 780, 1300, 2100, 3500, 5500, and 8800 Hz. The temporal envelope was then extracted from each analysis band using two different methods: (1) Hilbert transformation followed by low-pass (LP) filtering (elliptical IIR filters) at 50 or 500 Hz (−6 dB/octave) to smooth out the envelope, or (2) full-wave rectification followed by low-pass filtering (elliptical IIR filters) with cutoff frequency at 500 Hz (−6 dB/octave). While the periodicity cues were preserved in the temporal envelope with the 500 Hz cutoff frequency, it was reduced in the 50 Hz cutoff condition. This envelope signal was then used to modulate white noise, which was then bandpass filtered with the same analysis filters used on the original signal. Stimuli were resynthesized by summing these envelope-modulated narrow-band signals. Note that the temporal envelope fine structure (Rosen's definition) could be preserved in the lower frequency bands with the full-wave rectified envelope but not in the Hilbert envelope.

3. Procedure

All tests were performed in a double-walled sound-treated booth and the stimuli were presented to the listeners monaurally through TDH-49 headphones. The target original and processed speech were presented at a fixed level of approximately 65 dBA. All subjects were first presented with a block of original stimuli followed by four blocks of stimuli processed with the Hilbert envelope (2 spectral conditions ×2 LP cutoffs). Stimuli with higher numbers of frequency bands and higher LP cutoffs were tested first. The most difficult condition (i.e., one-band with 50 Hz LP cutoff) was tested last. Each block of stimuli consisted of one quiet and five different noise conditions. Within each block, the quiet condition was presented first followed by the noise conditions from the highest SNR (10 dB) to the lowest SNR (−10 dB). Therefore, the better performance in the higher SNR conditions compared to the lower SNRs cannot be due to the learning effect. To investigate the effect of the temporal envelope extraction techniques on Mandarin tone recognition, subjects were then further tested with two extra sets of processed stimuli (1-band and 8-band) which contained the envelopes that were extracted using the full-wave rectification and low-pass filtering technique.

The experiment was conducted using a four-alternative, forced-choice procedure, in which four Chinese characters with the same syllable and the tonal marking for each character ("–" for Tone 1, "/" for Tone 2, "∨" for Tone 3, and "\" for Tone 4) were displayed on a computer screen for each trial. The 200 words in the test battery were presented randomly in each experimental condition and subjects were

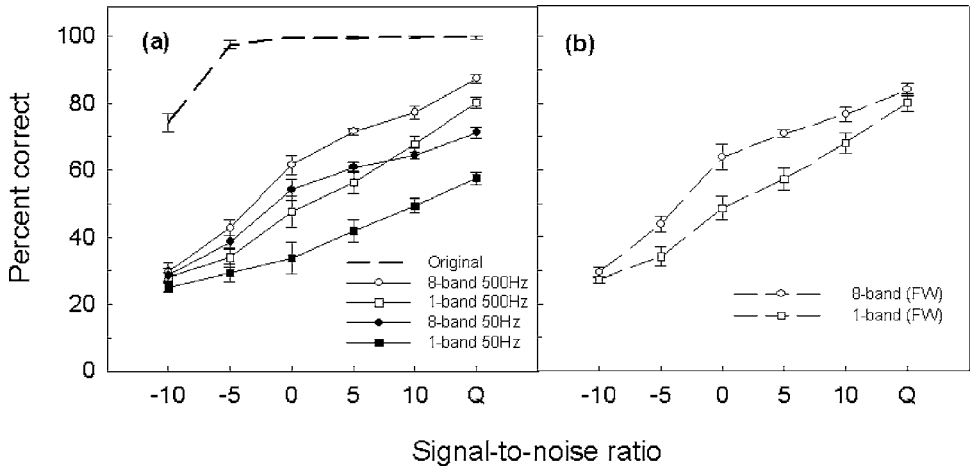

FIG. 2. Average tone recognition scores as a function of signal to noise ratio. Panel (a) shows the tone recognition performance with the Hilbert envelope. Dashed line (- - -) indicates performance with original stimuli. Square symbols represent the 1-band conditions and the circles represent the 8-band conditions. Open symbols indicate the 500-Hz LP conditions and the closed symbols indicate the 50-Hz LP conditions. Panel (b) shows the tone recognition performance with the full-wave rectified envelope in the 8-band (circles) and 1-band (squares) conditions. The vertical bars represent the standard error of the mean.

asked to choose the Chinese characters on the screen that corresponded to the tone they heard. Visual feedback was given after each response.

B. Results

Figure 2(a) shows the average percent correct tone recognition scores as a function of SNR for the original stimuli (dashed line) and for the four processed conditions with the Hilbert envelope (squares: 1-band conditions; circles: 8-band conditions; solid symbols: 50-Hz LP cutoff; open symbols: 500-Hz LP cutoff). On average, tone recognition with original stimuli was maintained at 99%–100% correct from quiet to −5 dB SNR. Performance decreased by 23 percentage points when the SNR was further reduced to −10 dB. Tone recognition was significantly poorer in all of the processed conditions compared to the original stimuli in quiet and at all noise levels ($p<0.001$).

A three-way repeated-measures ANOVA was performed to examine the following three factors: noise, spectral details (1-vs 8-bands), and temporal details (50- vs 500-Hz LP cutoff). Unlike the original condition in which performance plateaued at −5 dB SNR, tone recognition performance for the four processed conditions increased monotonically as a function of SNR ($p<0.001$) from nearly chance at −10 dB SNR for all conditions to 86%, 81%, 71%, and 58% correct in quiet for the 8-band 500-Hz LP, 1-band 500-Hz LP, 8-band 50-Hz LP, and 1-band 5-Hz LP conditions, respectively. Significant interactions were observed between the temporal and spectral details [$F(5,20)=15.0, p<0.01$], the temporal details and SNRs [$F(5,20)=13.7, p<0.01$], and spectral details and SNRs [$F(5,20)=8.5, p<0.01$], indicating that the relative importance of temporal and spectral cues for tone recognition differed in quiet and in noise. In general, tone recognition performance improved with 8 frequency bands compared to a single band [50-Hz LP: $F(1,4)=129.1, p<0.001$; 500-Hz LP: $F(1,4)=43.1, p<0.01$]. The presence of periodicity cues also contributed significantly to tone recognition independent of the amount of spectral details [1-band: $F(1,4)=151.8, p<0.001$; 8-band: $F(1,4)=583.1, p<0.001$]. When both the periodicity and spectral cues were available (8-band 500-Hz LP), tone recognition performance was superior to the other three conditions in both quiet and noise (except for the −10 dB SNR), but was the worst when both cues were absent (1-band 50-Hz LP). With either the periodicity (1-band 500-Hz LP) or spectral (8-band 50-Hz) cue, the pattern of performance differed in quiet and in noise. In quiet, the 1-band 500-Hz LP condition produced significantly better tone recognition than the 8-band 50-Hz LP condition [$F(1,4)=47.6, p<0.01$], but was worse in the 0 dB [$F(1,4)=14.4, p<0.05$], and −5 dB [$F(1,4)=8.3, p<0.05$] SNR conditions.

Figure 2(b) depicts the average tone recognition with full-wave rectified envelopes. Compared to the results in Fig. 2(a) with the 500-Hz LP conditions, we can see that tone recognition performance with Hilbert envelopes and full-wave rectified envelopes was essentially identical in both quiet and noise [1-band: $F(1,4)=0.2, p>0.05$; 8-band: $F(1,4)=1.0, p>0.05$], suggesting that listeners were unable to use the envelope fine-structure to perceive voice pitch. In general, the difference between Hilbert and full-wave rectified envelopes was less than 4 percentage points in quiet and at all SNRs for both the 1-band and the 8-band conditions.

III. MANDARIN TONE RECOGNITION WITH SLOWLY-VARYING AMPLITUDE AND FREQUENCY MODULATIONS

A. Methods

1. Subjects

Four additional normal-hearing native Mandarin speakers (S6–S9) participated in this experiment. They were three females and one male, with age ranging from 26 to 41 years. All subjects passed the hearing screening at 20 dB HL at frequencies between 125 and 8000 Hz bilaterally.

2. Stimuli

The recorded Mandarin speech tokens from both the male and female speakers in Experiment 1 were used in this

experiment. The original stimuli were processed in two different ways to extract the slowly-varying temporal envelope (amplitude modulation) alone or the amplitude modulation with additional slowly-varying frequency modulation information.

To produce stimuli with amplitude modulation (AM) cues alone, the signal processing scheme in Experiment 1 was used to extract the Hilbert envelope from 1 to 32 frequency bands in one-octave steps, which produced 1, 2, 4, 8, 16, and 32 subbands. The extracted temporal envelopes were then low-pass filtered at 500 Hz to preserve the periodicity cues in the original speech stimuli.

In addition to the amplitude modulation (AM) information, frequency modulation (FM) was extracted using the algorithm (Frequency-Amplitude-Modulation-Encoding Algorithm) developed by Nie *et al.* (2005) (see also Vongphoe and Zeng, 2005; Stickney *et al.*, 2005 for the description of this algorithm). In this algorithm, the slowly-varying AM was extracted in a similar manner as in the noise vocoder (Shannon *et al.*, 1995) described above. The slowly-varying FM was extracted by first removing each analysis band's center frequency through phase-orthogonal demodulators as used in phase vocoders (Flanagan and Golden, 1966). Two low-pass filters were then used to restrict the FM bandwidth and the FM rate. The bandwidth of the FM was restricted to 500 Hz or the bandwidth of the analysis bands whichever was less, and the FM rate was fixed at 400 Hz. To resynthesize, the slowly-varying FM signal from each subband was used to frequency modulate a sinusoid with a frequency equal to the center frequency of the subband and then the AM signal was used to amplitude modulate the frequency modulated sinusoid. The resynthesized AM and FM signals from each band were summed to recover the original signal. Unlike the fine-structure mathematically defined by the Hilbert transform, in which the obtained instantaneous frequency varies rapidly and over a broad range, the FM encoded using this technique is slowly-varying up to 400 Hz and band-limited to 500-Hz or to the critical bandwidth of the analysis bands.

3. Procedure

Presentation of the stimuli and the experimental setup were identical to those described in Experiment 1. All subjects were presented with a block of original stimuli first, followed by six blocks of AM+FM stimuli, and finally were tested with six blocks of AM-only stimuli. The six blocks within each processed condition (AM-only or AM+FM) corresponded to the six different number of frequency bands conditions from 1 to 32 bands and the blocks were presented in order from highest (32 bands) to lowest number of frequency bands (1 band). Similar to Experiment 1, each block of stimuli consisted of the quiet and different noise conditions and the subjects were presented with the quiet stimuli first followed by the noise conditions from the highest SNR to the lowest SNR. Five noise SNR conditions (−10, −5, 0, +5, +10 dB) were tested for AM-only processed stimuli whereas an additional SNR (−15 dB) was tested for both the original and the AM+FM processed stimuli.

B. Results

Figure 3 shows an average tone recognition performance with the original stimuli [the dashed line in top-left panel of Fig. 3(a)], the AM+FM stimuli (open triangles), and the AM stimuli (filled circles). Similar to the results revealed in Experiment 1, the average tone recognition for the four subjects in this experiment with the original stimuli increased from the chance level performance at −15 dB SNR to 73% correct at −10 dB SNR and then plateaued at 99% correct at −5 dB SNR. Figure 3(a) shows the tone recognition performance as a function of SNR with each panel representing a different number of frequency bands. First note that in the 1-band AM-only condition, tone recognition performance decreased monotonically from 79% correct in quiet to 51% at 0 dB SNR and approached the chance level at −5 dB SNR. While tone recognition was essentially identical between the 1-band and 2-band conditions, tone recognition performance improved significantly as the number of frequency bands further increased in quiet [$F(5,15)=25.6$, $p<0.005$] and at all noise levels [10 dB: $F(5,15)=35.5$, $p<0.005$; 5 dB: $F(5,15)=20.2$, $p<0.05$; 0 dB: $F(5,15)=38.0$, $p<0.005$; −5 dB: $F(5,15)=31.3$, $p<0.005$; −10 dB: $F(5,15)=20.8$, $p<0.05$].This finding is consistent with that reported in Experiment 1. Detailed spectral information, as much as 32 frequency bands, was required to produce tone recognition performance close to the original stimuli. Tone recognition with 32 bands decreased from 99% correct in quiet to 62% at −5 dB SNR and finally to 38% at−10 dB SNR. However, the performance difference between the original and the 32-band AM-only stimuli was significant from −10 to 5 dB SNRs [−10 dB: $F(1,3)=376.4, p<0.001$; −5 dB: $F(1,3)=189.6, p<0.005$; 0 dB: $F(1,3)=177.0, p<0.005$; 5 dB: $F(1,3)=49.0, p<0.01$]. The difference was 5, 15, 38, and 36 percentage points for 5, 0, −5, and −10 dB SNRs, respectively. In contrast, with additional FM, only 4 bands were necessary to achieve performance similarly to the original stimuli for both quiet and noise conditions. Significant performance difference between the 4-band AM+FM and the original stimuli was only 2 percentage points at 0 dB SNR [$F(1,3)=49.0, p<0.01$], 13 percentage points at −5 dB SNR [$F(1,3)=35.4, p=0.01$], and 18 percentage points at −10 dB SNR [$F(1,3)=161.5, p=0.001$].

The difference between the AM+FM and the AM-only conditions was also significant ($p<0.05$) in quiet and at all SNRs independent of the number of frequency bands. For a better visual comparison, the same set of data is replotted in Fig. 3(b) in which average scores are shown as a function of number of frequency bands with each panel representing the different SNR conditions. In general, averaged across different numbers of frequency bands, the differences between AM-only and AM+FM conditions increased from 11 to 36 percentage points as the amount of noise increased from the quiet condition to the −5 dB SNR condition. It is noted that performance with 32 frequency bands in the AM-only condition resembles that observed with two bands in the AM+FM condition, suggesting the contribution of FM cues.

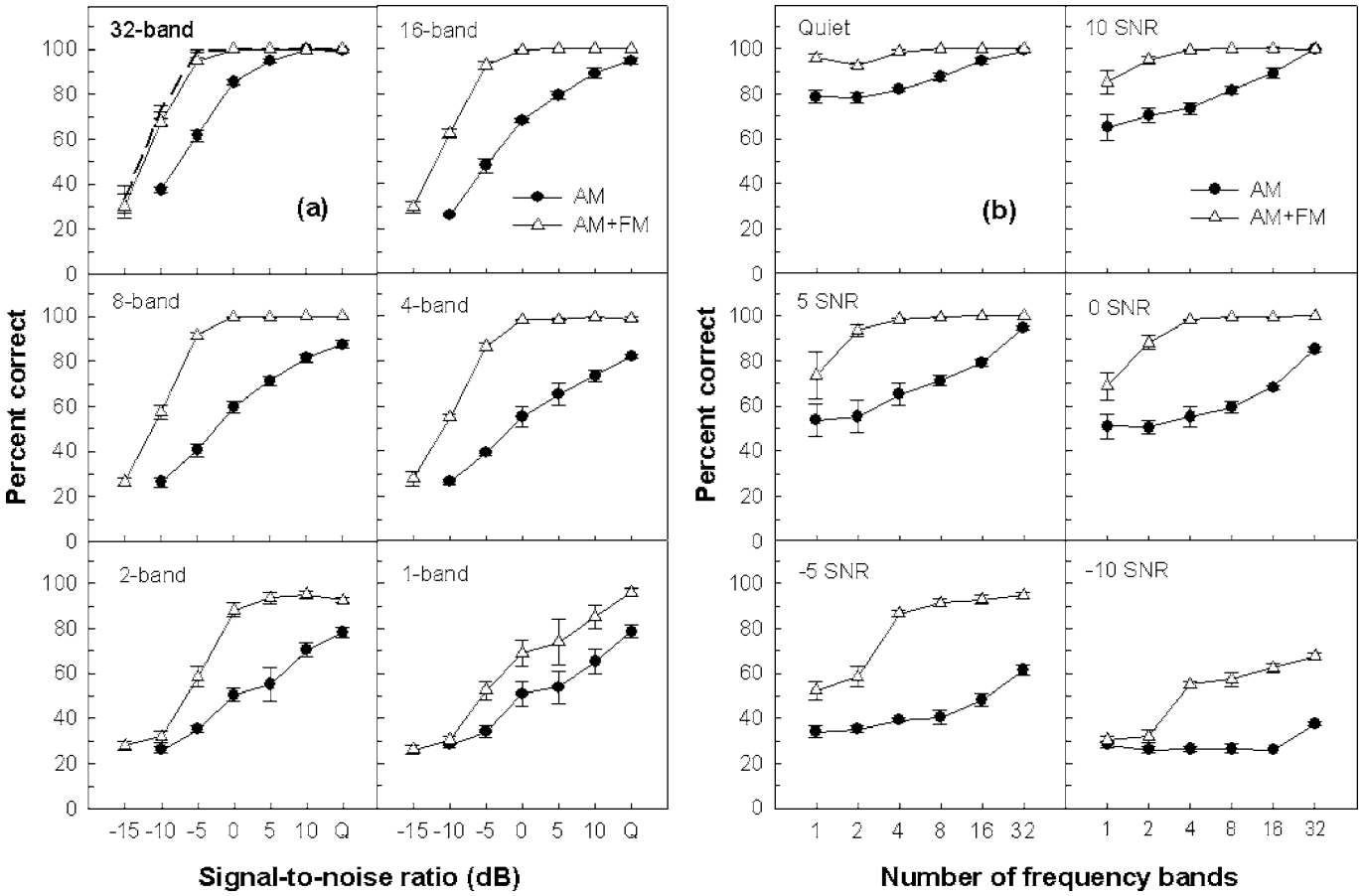

FIG. 3. Average tone recognition with original stimuli [dashed lines in panel (a)], AM alone (solid circles), and AM+FM (open triangles). Panel (a) shows recognition scores as a function of signal-to-noise ratio with individual panels indicating different number of frequency bands. Panel (b) shows the same set of data, but is plotted as a function of number of frequency bands. Individual panels represent different signal-to-noise ratio conditions. The vertical bars represent the standard error of the mean.

IV. MANDARIN TONE RECOGNITION WITH SPECTRAL ENVELOPE CUES

A. Methods

1. Subjects

The same five normal-hearing native Mandarin speakers (S1–S5) that participated in Experiment 1 also participated in this experiment.

2. Stimuli

The same Mandarin word list from Experiment 1 was used in this experiment. Three sets of stimuli were created to evaluate the contribution of spectral envelope and the harmonicity cues to Mandarin tone recognition. Two of the sets, which contained the spectral envelope cues were (1) naturally recorded whispered speech and (2) synthesized whispered speech.

Natural whispered speech was recorded from the same male and female speakers used in Experiment 1, resulting in a total of 200 whispered speech tokens. The recording procedures were identical to those described in Experiment 1. All speech waveforms were adjusted to have equal rms amplitude. Note that the amplitude contour patterns in the whispered speech were similar to those observed in their phonated counterparts. While the harmonic structure of the phonated speech is replaced by a noise source due to the opening of the vocal folds, the filtering properties (or resonance) of the vocal tract are preserved. In other words, whispered speech contained the formant structures that are similar to its phonated counterparts (von Helmholtz, 1863; Thomas, 1969; Hingashikawa and Minifie, 1999).

Synthesized whispered speech was created by extracting the spectral envelope of the original phonated speech recorded in Experiment 1, using a 14-order LPC technique. Signal processing was performed on every 10-ms time window with a 5-ms overlap. The extracted spectral envelope was then used to modulate with white noise.

To evaluate the contribution of spectral fine structure (i.e., harmonicity cues) to tone recognition, another set of stimuli which contained only the F0 and the harmonics of the original phonated speech was created. The F0 and harmonics of the original phonated speech were extracted as the by-product of the 14-order LPC processing, which was used for creating the synthesized whispered speech. With this processing, the F0, the harmonic, and the phase spectrum of the original stimuli were preserved, but the spectral envelope was flat with no identifiable formant peaks.

3. Procedure

Presentation of the stimuli and the experimental setup were identical to those described in Experiment 1. All sub-

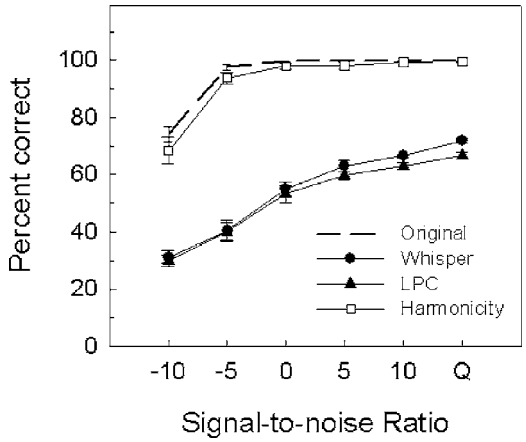

FIG. 4. Mean tone recognition scores with spectral envelope and harmonicity cues. Tone recognition scores with original stimuli (dashed line) from Experiment 1 are replotted in this figure for comparison. Solid circle and triangle symbols represent the natural whispered speech and synthesized whispered speech using the LPC technique, respectively. Open squares represent stimuli with only harmonicity cues. The vertical bars represent the standard error of the mean.

jects were presented with a block of original stimuli first, followed by three blocks of experimental stimuli (natural whispered speech, synthesized whispered speech, and speech with only harmonicity cues). Each block of stimuli consisted of the quiet and five different noise conditions and the subjects were presented with the quiet stimuli first followed by the noise conditions from the highest SNR to the lowest SNR. The noise used for the whispered speech was created by summing all the whispered speech tokens from the male speaker or the female speaker (see noise generating procedures in experiment 1). As for the other two conditions, phonated target speech underwent LPC processing before mixed with noise.

B. Results

Figure 4 shows the average tone recognition scores for the natural whispered speech (circles), synthesized whispered speech (triangles), and speech stimuli with harmonicity cues alone (squares). Mean recognition scores with original stimuli from Experiment 1 are included to facilitate comparison. First note that tone recognition performance with harmonicity cues alone was essentially identical to that with original stimuli at all SNR conditions ($p>0.05$). Second, tone recognition scores with primarily spectral envelope cues (both natural and synthesized whispered speech) were significantly poorer ($p<0.0001$) than those obtained with harmonicity cues alone in quiet and at all SNRs. Tone recognition with spectral envelope cues increased monotonically as the SNR increased from chance level performance at −10 dB SNR to 72% and 67% correct in quiet for the natural whispered and synthesized whispered speech, respectively. Third, post-hoc pairwise comparisons with Bonferroni correction revealed that subjects performed similarly ($p>0.05$) between the natural whispered speech and synthesized whispered speech, except in the quiet condition $[F(1,4)=32.7, p=0.005]$ where the natural whispered speech produced about 5 percentage points better recognition than the synthesized speech.

V. DISCUSSION

A. Comparison with previous studies

Consistent with previous studies (e.g., Fu *et al.*, 1998; Xu *et al.*, 2002), listeners in our study were able to use solely the slowly-varying temporal envelope information, i.e., the amplitude contours and durational cues, to identify the tonal patterns with accuracy of about 60% in the absence of any spectral contents (1-band 50-Hz LP) in quiet. Tone recognition was further improved with additional periodicity cues (1-band 500-Hz LP) by about 20 percentage points. The present results are also in agreement with previous studies of the effect of the number of bands on tone recognition. Both the present and previous results found that tone recognition was essentially the same between one and two frequency bands but improved as the number of frequency bands was increased to 4–8 bands. Using the more challenging stimuli of monosyllabic words with equal durations, Xu *et al.* (2002) found that tone recognition continued to improve up to 12 bands, the highest number of bands tested in their study. In our study, we found further improvement up to 32 bands, with such improvement being particularly apparent in noise.

Xu *et al.* (2002) also reported the existence of a trade-off between temporal and spectral cues for Mandarin tone recognition in quiet. They showed that subjects can achieve a similar tone recognition performance using either a small number of frequency bands combined with a higher temporal envelope cutoff frequency, or a larger number of bands, but with a low envelope cutoff frequency. Here, we observed another relationship—a complementary contribution between temporal periodicity cues and spectral cues for tone recognition between quiet and noise conditions. Note that while the 1-band 500-Hz LP condition produced better tone recognition performance than in the 8-band 50-Hz LP condition in quiet, a reverse pattern was observed in noise. This complementary contribution between periodicity cues and spectral cues suggests that periodicity cues are more important than coarse spectral cues for Mandarin tone recognition in quiet, but they are more susceptible to noise compared to spectral cues.

B. Relative contribution of envelope and fine structure cues to tone recognition

Our study demonstrated that pitch perception was dominated by fine structure rather than the envelope cues. Although tone recognition can be achieved by all four cues in quiet, fine structure cues play a more important role than the envelope cues in noise. In quiet, spectral or temporal envelope cues alone produced relatively low (70%–80% correct) tone recognition performance, but almost perfect performance was observed with spectral or temporal fine structure cues. The difference in performance between envelope and fine structure cues is even greater in noise. From −10 to 10 dB SNR, while tone recognition was still maintained at almost perfect performance level with spectral or temporal fine structure cues, it was 30–60 percentage points lower

with spectral or temporal envelope cues, suggesting that envelope cues are less salient and are more susceptible to noise in voice pitch perception. It should be noted that spectral cues are available in the lower numbers of bands conditions in our AM+FM stimuli and they could be used by our normal-hearing subjects to perceive pitch. Nevertheless, the improved tone recognition in noise with higher numbers of bands (i.e., 16- and 32-bands) is likely to be attributed to the fine timing information within as well as across bands.

The importance of fine structure cues for Mandarin tone recognition is consistent with previous psychophysical findings on pitch perception with pure tones and complex tones. Pitch of a harmonic complex can be extracted based on the place cue (Goldstein, 1973; Terhardt, 1974), or the timing cue (Licklider, 1951; Schouten *et al.*, 1962), or both (Oxenham *et al.*, 2004). Salient pitch percepts can be elicited on the basis of fundamental frequency and the lower resolved harmonics (Ritsma, 1967). Thus, it was not surprising that Mandarin tone recognition was much improved with harmonicity cues compared to spectral envelope cues alone.

In contrast to the fine structures, the saliency of the pitch provided by the temporal envelopes is rather weak, as demonstrated in previous findings on melody recognition with sinusoidally amplitude-modulated noise (Burns and Viemeister, 1981) and modulation frequency discrimination for amplitude-modulated stimuli (Formby, 1985; Grant *et al.*, 1998). The susceptibility of temporal envelope cues to noise in tone recognition could be attributed to the low salience of temporal envelope pitch and the reduced detectability of amplitude modulation in noise.

C. Contribution of formant frequencies to tone recognition

1. Processed speech with temporal envelope cues

Although we have established the role of coarse spectral cues for Mandarin tone recognition, the type of spectral information in this processed speech that enabled the enhancement of tone recognition has not yet been determined. The most important type of spectral information is the F0 and the harmonics. In our recorded speech, the F0 variation patterns were consistent with those reported previously. An acoustical analysis of the F0 of the recorded stimuli was conducted using the correlation algorithm of the TFR™ software.[1] For the male speech, the average F0 of Tone 1 was 174 Hz, the F0 of Tone 2 rose from 113 to 193 Hz, Tone 3 initially fell from 114 to 87 Hz, then rose to 147 Hz, and Tone 4 fell from 189 to 85 Hz. The female speech exhibited similar pitch contours, but with a higher F0 for all tones. In the 8-band AM-only processed speech, the F0 differences among the four tones may not be well represented, due to the fact that the low harmonics were not resolved by the low-frequency analysis filters. Therefore, it was unlikely that better tone recognition with 8-band processing was due to the better encoding of the F0 and the harmonics in the original speech stimuli.

Another type of spectral cue present in speech is the formant structures. Unlike the harmonics, formant information of speech could be coarsely represented in the 8-band processed stimuli and in turn, vowel recognition performance could be achieved with at least 80% accuracy in quiet (Friesen *et al.*, 2001). However, how is this formant information able to convey differences in Mandarin tones for the same syllable? A careful examination of the formant frequencies of the original recorded single-vowel syllables (/ma/, /p^hɔ/, /k^hɛ/, /ɕi/, /du/, /dʑy/) from the male speaker[2] measured by both the TFR™ software and by a speech synthesis and analysis program PRAAT (Boersma and Weenink, 2005) revealed that:

(1) Formant structures are different for different lexical tones of the same syllable. This finding is supported by the recent articulatory and acoustic data reported in Erickson *et al.* (2004). Using an electromagnetic articulatograph, Erickson and co-workers demonstrated that the jaw and tongue position were significantly more retracted for Tone 3 than for Tone 1 for the vowel /a/, resulting in a higher F1 frequency for Tone 3.
(2) The patterns of difference in formants among tones are not the same for all syllables. In our measurement, F1 frequency was higher for Tone 3 than for Tone 1 for the low vowel, but the pattern of difference was inverted for high and mid vowels. The difference in F1 between Tone 1 and Tone 3 was as large as 100 Hz, particularly for the high vowels. The second formant (F2), however, was consistently higher for Tone 1 than Tone 3 in all vowels, and the difference reached 200 Hz for the back vowels.
(3) Even with a single vowel, the formant frequencies changed considerably over time and the patterns of change differed for different tones. The interesting finding was that the direction of formant changes did not covary with the F0.

The above findings suggest that there may be a possible coupling between the source (vocal folds vibration) and the filter (vocal track configuration) in the articulatory system. This is different from the speech production mechanism proposed in the tradition source-filter theory (Fant, 1960). The source-filter model assumed a linear system in which there was disassociation between the source and the filter.

The considerable difference between F1s among the four lexical tones and the changing frequency of the formants over time for each tone could be represented in an 8-band processed condition. Also, one cannot rule out the possibility that the excitation patterns arising from different filters, due to the overlapping of the filters, may provide additional non-pitch cues for tone recognition. Since the F1 differences among the tones were larger for some vowels (e.g., high vowels) than for others (e.g., mid vowels), it would be of interest to see how tone recognition performance with processed stimuli differed with different syllables. Such information, however, was not recorded in our result matrix during our experiment. Further examination of the effects of vowels on tone recognition is needed.

2. Whispered speech

The contribution of formant frequencies to tone recognition was further supported by our findings with whispered

speech stimuli. Previous studies on whispered speech demonstrated that listeners can (1) match the pitch of whispered vowels to some standard frequency using only the formant frequency information [F1: von Helmholtz (1863); F2: Thomas (1969); F1 and F2: Hingashikawa and Minifie (1999)]; (2) identify speaker's sex with approximately 70%–80% correct performance (Lass *et al.*, 1976; Tartter, 1991); and (3) achieve 60%–80% correct tone recognition in other languages, including Norwegian, Swedish, Slovenian, and Mandarin Chinese (Jensen, 1958). In particular, our results are consistent with Jensen's findings, in which listeners can achieve about 60%–70% correct for whispered speech in quiet. Although most of the studies on whispered speech suggested the contribution of formant information to voice pitch perception, a study by Remez *et al.* (1997) on sinewave speech did not show this relationship. Remez *et al.* (1997) used three sinusoids tracking the lowest three formants of a natural speech and found that normal-hearing listeners could correctly identify the both the speech and speaker. However, they found that the formant (F1 and F2) differences between speakers and the error patterns for speaker identification were poorly correlated, suggesting that listeners did not rely on formant information to identify speakers. They concluded that listeners might be able to use other linguistic features, such as idiolect, dialect, or style differences to identify the talkers. The disagreement between sinewave speech and the whispered speech remains determined.

Acoustical analyses on whispered Mandarin speech were conducted and results showed that the differences of Fl and F2 among the four tones were similar between the whispered speech and their phonated counterparts[3]. These findings were also evidenced in a recent study by Li and Xu (2005). Similar to the phonated speech, both the durational and amplitude contour cues are also available in whispered speech. One might claim that the significant Mandarin tone recognition performance with whispered speech could be attributed to solely the contribution of the temporal envelope cues. However, we argued that the durational and amplitude contour cues alone could not account for relatively high levels of tone recognition in both quiet and noise. The 1-band 50-Hz LP condition yielded a performance level at about 58% in quiet, 14 percentage points lower than the whispered speech and the difference between whispered speech and 1-band 50-Hz LP speech was as large as 21 percentage points at 5 and 0 dB SNR (see Fig. 5). The differences in F1 and F2 among the four tones could be better detected by normal-hearing listeners in the whispered speech compared to the 8-band processed speech. Surprisingly, tone recognition performance was essentially the same between the 8-band 50-Hz LP condition and the whispered speech in both quiet and in noise (see Fig. 5). In addition to the percent correct scores, the error patterns between the 8-band processed and natural whispered speech were similar. These findings suggest that (1) listeners used the same cues to perform the task in both conditions, and (2) well-defined formant information, as in whispered speech, was not necessary for Mandarin tone recognition.

The further improvement beyond 8 frequency bands, i.e., the 16- and 32-band conditions in Experiment 2, may be attributed to the resolvability of the harmonics. Qin and Oxenham (2005) investigated the F0 discrimination limen (F0DL) for a harmonic complex processed by a noise vocoder similarly to Shannon *et al.* (1995). The F0DLs were found similar from 1 to 8 frequency bands and improved considerably in the 24- and 40-band conditions. They concluded that the better F0DLs with the higher number of frequency bands were associated with the spectral cues, and the poorer F0DLs with fewer number of frequency bands were the result of non-salient pitch associated with the envelope cues.

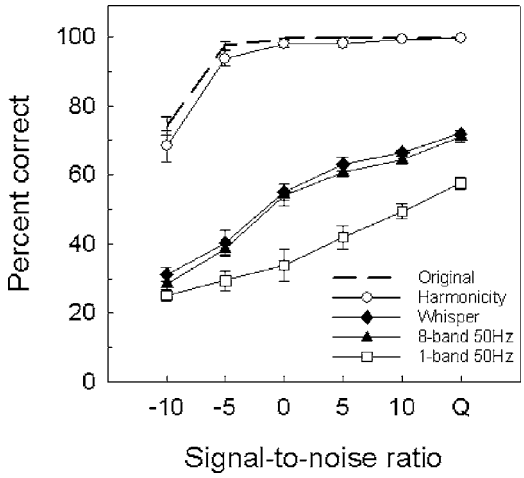

FIG. 5. Mean tone recognition scores for original (dashed line), natural whispered speech (solid diamonds), processed stimuli with 8-band 50-Hz LP (solid triangles), 1-band 50-Hz LP (open squares), and stimuli with harmonicity cues alone (open circles). The vertical bars represent the standard error of the mean.

D. Implication for cochlear-implant design

Except for the analog processing strategy, current cochlear implant processing provides only the temporal envelope information from several frequency bands. Both temporal and spectral fine structures are not explicitly encoded in these processing schemes. A study by Wei *et al.* (2004) on Mandarin tone recognition in cochlear-implant listeners measured performance as a function of number of electrodes. They revealed that implant listeners performed only at average 57% correct with a range of 25%–71% correct when listening with a 20-electrode map, considerably lower than the normal-hearing performance even at the 1-band AM-only conditions observed in the present study. Although, the appropriateness of using standard noise-band vocoders as an acoustic model for pitch perception in cochlear implants is being questioned (Laneau *et al.*, 2006), we would still expect implant listeners to be able to perform better than the 1-band AM-only condition given the potential limitation on envelope-periodicity pitch cues in the vocoder. The inability of cochlear-implant listeners to perceive lexical tones was also documented in other languages. Ciocca *et al.* (2002) studied a group of early-deafened Cantonese-speaking cochlear-implant children. They reported that very few of their cochlear implantees performed above the chance level in a two-alternative choice tone identification task.

Improving current cochlear-implant performance in pitch perception can be accomplished by better encoding the temporal and/or spectral fine structures. The present study on Mandarin tone recognition with additional slowly-varying frequency modulation showed significant improvement in Mandarin tone recognition even with a fewer number of frequency bands and also in noisy conditions. One potential problem of adding temporal fine-structure in cochlear implants is that the timing information may not be mapped at the right place along the cochlea in electric hearing, which in turn would still not elicit the salient pitch necessary for voice pitch and music perception. Oxenham *et al.* (2004), using transposed stimuli that resembled the low-frequency pure tones, showed that frequency discrimination with the transposed tones, presented at a higher-frequency place, were considerably poorer than for the pure tones.

Pitch perception in cochlear implants can also be improved with better frequency-to-electrode mapping and/or increasing the number of functional channels. Our results, together with Qin and Oxenham (2005), showed that tone recognition and F0 discrimination improved as increasing the number of frequency bands.

VI. CONCLUSIONS

Mandarin tone recognition was measured in a group of normal-hearing listeners with the goal to understand the underlying mechanisms for voice pitch perception. Four types of speech stimuli were created to evaluate the relative contributions of temporal and spectral envelope and fine structure cues to tone recognition. The main findings are:

(1) Listeners can achieve nearly perfect tone recognition performance with spectral or temporal fine structure cues alone in quiet, but only 70%–80% correct with the envelope cues alone. In noise, the performance difference between fine structure and envelope cues can range from 30 to 60 percentage points from −10 to 10 dB SNR.
(2) While envelope cues, either in the spectral or temporal domain, are sufficient for speech recognition in quiet, fine structure is critical for voice pitch perception in noise.
(3) There exists a complementary contribution of temporal and spectral envelope cues to tone recognition. While the 1-band 500-Hz LP condition produced better tone recognition than the 8-band 50-Hz LP condition in quiet, a reverse pattern was observed in noise. This suggests that temporal envelope cues are susceptible to noise, but spectral cues are more resistant to noise.
(4) Differences in formant frequencies were observed among the four tones for the same vowel in Mandarin speech, suggesting a potential coupling between F0 and formant frequencies.

ACKNOWLEDGMENTS

We are grateful to Dr. Kaibao Nie for providing the simulation programs, Jing Zhuo for assisting in data collection, and Sheng Liu for the technical support. We thank Ackland Jones for his helpful comments on the earlier version of this manuscript. We also thank Dr. Virginia Mann for her inspiration on the use of whispered speech in Experiment 3 and for her helpful discussion during the entire project. This work was supported in part by the National Institutes of Health, Department of Health and Human Services (A research supplement award to Y.Y.K. and 2 RO1-DC-02267 to F.G.Z).

[1]TFR: Time Frequency Representation software Copyright ©1996, 2000 AVAAZ Innovations Inc.
[2]Measurements of formants for the female speaker were somewhat unreliable due to the high F0.
[3]The formants in the syllable /ɕi/ in the whispered speech could not be reliably obtained due to the excessive noise produced by the speaker.

Abramson, A. S. (**1978**). "Static and dynamic acoustic cues in distinctive tone," Lang Speech **21**, 319–325.
Boersma, P., and Weenink, D. (2005). "Praat: doing phonetics by computer (Version 4.3.14)" (Computer program). Retrieved May 26, 2005, from http://www.praat.org/.
Burns, E. M. and Viemeister, N. F. (**1981**). "Played again SAM: Further observations on the pitch of amplitude-modulated noise," J. Acoust. Soc. Am. **70**, 1655–1660.
Ciocca, V., Francis, A. L., Aisha, R., and Wong, L. (**2002**). "The perception of Cantonese lexical tones by early-deafened cochlear implantee," J. Acoust. Soc. Am. **111**, 2250–2256.
Erickson, D., Iwata, R., Endo, M., and Fujino, A. (2004). "Effect of tone height on jaw and tone articulation in Mandarin Chinese," presented at the International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, China.
Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, Hague).
Flanagan, J. L., and Golden, R. M. (**1966**). "Phase vocoder," Bell Syst. Tech. J. **45**, 1493–1509.
Formby, C. (**1985**). "Differential sensitivity to tonal frequency and to the rate of amplitude modulation of broadband noise by normally hearing listeners," J. Acoust. Soc. Am. **78**, 70–77.
Friesen, L. M., Shannon, R. V., Basken, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.
Fu, Q.-J. and Zeng, F.-G. (**1998**). "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am. **104**, 505–510.
Fu, Q.-J. and Zeng, F.-G. (2000). "Identification of temporal envelope cues in Chinese tone recognition," Asia Pacific Journal of Speech, Language, and Hearing **5**, 45–57.
Gandour, J. (**1984**). "Tone dissimilarity judgments by Chinese listeners," J. Chin. Linguist. **12**, 235–260.
Goldstein, J. L. (**1973**). "An optimum processor theory for the central formation of the pitch of complex tones," J. Acoust. Soc. Am. **54**, 1496–1516.
Grant, K. W., Summers, V., and Leek, M. R. (**1998**). "Modulation rate detection and discrimination by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **104**, 1051–1060.
Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species–29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
Hartmann, W. M. (**1997**). *Signals, Sound and Sensation (Modern Acoustics and Signal Processing Series)* (Springer-Verlag, New York).
Higashikawa, M. and Minifie, F. D. (**1999**). "Acoustical-perceptual correlates of 'whisper pitch' in synthetically generated vowels," J. Speech Lang. Hear. Res. **42**, 583–591.
Hilbert, D. (**1912**). *Grundzuge einer Allgerneinen Theorie der linearen Integralgleichungen (Foundations of the General Theory of Linear Integral Calculus)* (Teubner, Leipzig).
Howie, J. M. (**1976**). *Acoustical Studies of Mandarin Vowels and Tone* (Cambridge University Press, Cambridge, England).
Jensen, M. K. (**1958**). "Recognition of word tones in whispered speech," Word **14**, 187–196.
Laneau, J., Moonen, M., Wouters, J. (**2006**). "Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants," J. Acoust. Soc. Am. **119**, 491–506.
Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., and Bourne, V. T. (**1976**). "Speaker sex identification from voiced, whispered, and filtered isolated vowels," J. Acoust. Soc. Am. **59**, 675–678.

Li, X-L., and Xu, B.-L. (**2005**). "Formant comparison between whispered and voiced vowels in Mandarin," Acta. Acust. Acust. **91**, 1079–1085.
Liang, Z.-A. (**1963**). "Hanyu Putonghua zhong shengdiao de tingjiao bianren yiju (The auditory basis of tone recognition in Standard Chinese)," Acta Phys. Sin. **26**, 85–91.
Licklider, J. C. (1951). "A duplex theory of pitch perception," Experientia **VII**, 128–134.
Lin, M.-C. (1988). "Putonghua shengdiao de shengxue texing he zhijiao zhengzao. (The acoustic characteristics and perceptual cues of tones in Standard Chinese)," Zhongguo Yuwen (Chinese Language) **204**, 182–193.
Miller, J. D. (**1961**). "Word tone recognition in Vietnamese whispered speech," Word **17**, 11–15.
Nie, K., Stickney, G., and Zeng, F.-G. (**2005**). "Encoding frequency modulation to improve cochlear implant performance in noise," IEEE Trans. Biomed. Eng. **52**, 64–73.
Oxenham, A. J., Bernstein, J. G. W., and Penagos, H. (**2004**). "Correct tonotopic representation is necessary for complex pitch perception," Proc. Natl. Acad. Sci. U.S.A. **101**, 1421–1425.
Qin, M. K. and Oxenham, A. J. (**2005**). "Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification," Ear Hear. **26**, 451–460.
Remez, R. E., Fellowes, J. M., and Rubin, P. E. (**1997**). "Talker identification based on phonetic information," J. Exp. Psychol. Hum. Percept. Perform. **23**, 651–666.
Ritsma, R. J. (**1967**). "Frequency dominant in the perception of the pitch of complex sounds," J. Acoust. Soc. Am. **42**, 191–198.
Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.
Schouten, J. F., Ritsma, R. J., and Cardozo, B. L. (**1962**). "Pitch of the residue," J. Acoust. Soc. Am. **34**, 1418–1424.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature (London) **416**, 87–90.
Stickney, G. S., Nie, K., and Zeng, F. G. (**2005**). "Contribution of frequency modulation to speech recognition in noise," J. Acoust. Soc. Am. **118**, 2412–2420.
Tartter, V. C. (**1991**). "Identifiability of vowels and speakers from whispered syllables," Percept. Psychophys. **49**, 365–372.
Terhardt, E. (**1974**). "Pitch, consonance, and harmony," J. Acoust. Soc. Am. **55**, 1061–1069.
Thomas, I. B. (**1969**). "Perceived pitch of whispered vowels," J. Acoust. Soc. Am. **46**, 468–470.
von Helmholtz, H. L. J. (**1863**). *On the Sensation of Tone as a Physiological Basis for the Theory of Music (translated by Alexander Ellis)* (Dover, New York **1954**).
Vongphoe, M. and Zeng, F.-G. (**2005**). "Speaker recognition with temporal cues in acoustic and electric hearing," J. Acoust. Soc. Am. **118**, 1055–1061.
Wei, C.-G., Cao, K., and Zeng, F.-G. (**2004**). "Mandarin tone recognition in cochlear-implant listeners," Hear. Res. **197**, 87–95.
Whalen, D. H. and Xu, Y. (**1992**). "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica **49**, 25–47.
Xu, L., Tsai, Y., and Pfingst, B. E. (**2002**). "Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses," J. Acoust. Soc. Am. **112**, 247–258.
Xu, L. and Pfingst, B. E. (**2003**). "Relative importance of temporal envelope and fine structure in lexical-tone perception," J. Acoust. Soc. Am. **114**, 3024–3027.