Acoustic cues to tonal contrasts in Mandarin: Implications for cochlear implants

Yu-Ching Kuo[a)]
Department of Special Education, Taipei Municipal University of Education, No. 1, Ai-Guo West Road, Taipei, 10042, Taiwan and Department of Phonetics and Linguistics, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom

Stuart Rosen and Andrew Faulkner
Department of Phonetics and Linguistics, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom

(Received 5 January 2007; revised 14 February 2008; accepted 14 February 2008)

The present study systematically manipulated three acoustic cues—fundamental frequency (f0), amplitude envelope, and duration—to investigate their contributions to tonal contrasts in Mandarin. Simplified stimuli with all possible combinations of these three cues were presented for identification to eight normal-hearing listeners, all native speakers of Mandarin from Taiwan. The f0 information was conveyed either by an f0-controlled sawtooth carrier or a modulated noise so as to compare the performance achievable by a clear indication of voice f0 and what is possible with purely temporal coding of f0. Tone recognition performance with explicit f0 was much better than that with any combination of other acoustic cues (consistently greater than 90% correct compared to 33%–65%; chance is 25%). In the absence of explicit f0, the temporal coding of f0 and amplitude envelope both contributed somewhat to tone recognition, while duration had only a marginal effect. Performance based on these secondary cues varied greatly across listeners. These results explain the relatively poor perception of tone in cochlear implant users, given that cochlear implants currently provide only weak cues to f0, so that users must rely upon the purely temporal (and secondary) features for the perception of tone. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2896755]

PACS number(s): 43.71.Es, 43.70.Fq, 43.71.Ky [AJO] Pages: 2815–2824

I. INTRODUCTION

In Mandarin Chinese, four tones are one of the phonetic determinants of lexical meaning (e.g., Chao, 1948; 1968; Yip, 1980). These four tones are mainly distinguished by their variations in fundamental frequency (e.g., Howie, 1976; Lin, 1988). The four tonal patterns are often described in terms of the pitch range of the particular speaker. Tone 1, the level tone, starts within a speaker's high pitch range and maintains approximately the same pitch to the end of the syllable. Tone 2, the rising tone, starts in a middle pitch range, then shows a rise, which is sometimes preceded by a small dip. Tone 3, the falling–rising tone, also starts in a middle pitch range, falls gradually toward a low pitch and then rises. Tone 4, the falling tone, starts at a high pitch, and drops to a low one. A syllable with different tones represents different words. For example, the syllable *ba* expressed with tones 1, 2, 3, and 4 means "eight," "to pull," "to hold," and "father," respectively.

Although tone 3 is often referred to as a falling–rising pattern, it also appears in other forms (e.g., Chao, 1948; 1968; Shih, 1988; Chen, 2000). This variation is due to phonological processes that are described by the third tone sandhi rule. A falling–rising pattern appears in isolated words and in sentence-final position. However, in nonfinal positions, a low-falling pattern occurs. In addition, if a syllable with tone 3 is followed by another syllable with the same tone, the first syllable will be pronounced as tone 2, a rising tone. Therefore, tone 3 appears in three patterns: the falling–rising pattern, the low-falling pattern, and the rising pattern. Previous studies have shown that tone 3 was sometimes misidentified as the falling tone (tone 4) or the rising tone (tone 2) (Gårding *et al.*, 1986; Shen and Lin, 1991).

The usage of the falling–rising and the low–falling patterns also varies according to dialect and/or speaking style (Shih, 1988). Northern Chinese speakers use the falling–rising pattern in sentence-final position in all speech. By contrast, southern speakers use the low-falling pattern even in the final position in casual speech, and use the falling–rising pattern in emphatic speech and in yes–no questions. Dialect may also influence Mandarin tone. Fon and Chiang (2000) reported that Taiwan Mandarin showed a narrower tonal range, lower tonal heights, and more conservative tonal contours compared to Beijing Mandarin. This could be attributed to an influence from Taiwanese, which has a lower tonal register than Mandarin (Lin and Repp, 1989).

While the four tones are mainly distinguished by fundamental frequency, other acoustic characteristics such as overall intensity and duration tend to vary systematically with tone (Howie, 1976; Zee, 1978; Blicher *et al.*, 1990; Tseng, 1990; Whalen and Xu, 1992; Fu and Zeng, 2000). Tone 4 is often the most intense one and has the shortest duration,

[a)]Electronic mail: ykuo@tmue.edu.tw

while tone 3 is the least intense with the longest duration. However, different patterns of tone 3 differ in duration as well as in pitch contour. While the falling–rising pattern shows the longest duration among the four tones, the low-falling pattern has the shortest duration (Shih, 1988). Amplitude envelope has also been found to be highly correlated with f0 contour for tones 3 and 4 (Whalen and Xu, 1992; Fu and Zeng, 2000).

Several studies have investigated tone perception from purely temporal cues using various noise stimuli modulated by envelopes derived from natural speech stimuli. Three temporal envelope cues were defined by Fu and Zeng (2000): periodicity, amplitude contour, and duration. As in Rosen (1992), "periodicity" referred to fluctuations in the overall amplitude at a rate between 50 and 500 Hz, whereas "amplitude contour" referred to fluctuations at a rate between 2 and 50 Hz. Although percepts of pitch are strongest for truly periodic sounds, they can also be elicited by temporal fluctuations in amplitude-modulated noise (Burns and Viemeister, 1976). When the temporal envelope of speech is derived from rectified speech with an envelope-smoothing filter which includes the voice f0 range, listeners can perceive the quasiperiodic amplitude modulations imposed on a noise carrier as pitch changes. The temporal regularities in the modulated noise can be referred to as periodicity information (Rosen, 1992). However, when the cutoff frequency of an envelope-smoothing filter falls below the fundamental frequency range, the periodicity information is not included in the envelope signal.

Whalen and Xu (1992) used signal-correlated-noise (SCN) to investigate the extent to which tone recognition can be aided by amplitude contour and duration. Their results showed high recognition scores for tones 2–4, averaging 87.6%, but only 38.5% correct for tone 1. To further examine the contribution of amplitude contour, they used stimuli with controlled duration, and found that recognition scores were still well above chance in the absence of the duration cue (45%, 55.3%, 69.5%, and 92.3% for tones 1, 2, 3, and 4, respectively). However, SCN is equivalent to wide-band noise modulated with an envelope extracted using a smoothing filter with a very high cut-off frequency on full-wave rectified speech. Therefore, SCN stimuli contain not only amplitude contour and duration cues, but also the periodicity of the natural tokens on which they were based, albeit in weakened form. Green *et al.* (2002) showed listeners to have some ability to label glides in fundamental frequency at low frequency ranges in stimuli that were similar to, although not identical to, SCN. Thus, periodicity information, directly related to the change in fundamental frequency, was likely to influence subject performance as well (Rosen, 1992; Van Tasell *et al.*, 1987).

Fu *et al.*, 1998 investigated the importance of periodicity and amplitude envelope to tone recognition by using amplitude-modulated noise. In this study, duration was not examined and retained its natural variations. When no spectral information was available (1-band condition), recognition scores of up to 80% could be achieved in a condition that preserved amplitude envelope, periodicity, and duration information, while accuracy was about 67% in a condition preserving only amplitude envelope and duration information. Tones 3 and 4 were identified best, nearly twice as well as tone 1 and tone 2. In another study, Fu and Zeng (2000) used SCN to further explore the contribution of periodicity, amplitude contour, and duration to tone recognition. These three cues were manipulated in isolation, in pairs, and all together. Results showed that the highest score, nearly 70%, was achieved in the condition with all three cues. Around 55% correct recognition was shown in conditions that preserved either the amplitude contour or periodicity cues. A condition preserving only the duration cue had the lowest recognition score of about 35% correct. The duration cue was found to contribute to recognition of tone 3, and the amplitude cue contributed to the recognition of both tones 3 and 4. The periodicity cue contributed to recognition of all four tones.

In a recent study by Xu *et al.* (2002), amplitude envelope and periodicity were manipulated by varying the low-pass cutoff frequency of an envelope-smoothing filter from 1 to 512 Hz in 1 octave steps. Tone recognition improved slightly but consistently as cutoff frequency increased in the 1-channel condition in which there was no spectral information available. Consistent with the results of Fu *et al.* (1998), around 50% correct recognition was achieved with only amplitude envelope and duration cues (when the envelope cutoff frequency was as low as 1 or 2 Hz), while accuracy approached 65% when periodicity cues were also available (with an envelope cutoff frequency at 512 Hz). When the duration cue was removed, scores dropped significantly, with tones 3 and 4 affected the most (the decreases were by 12.5, 8.5, 19.5, and 19.7 percentage points for tones 1, 2, 3, and 4, respectively). These decreases in performance indicated that duration did show a significant effect on tone recognition mainly on tones 3 and 4.

To sum up, results from the above-mentioned studies confirm that tone recognition can be assisted by temporal envelope cues. Although the pitch from modulated noise is considerably less salient than the pitch of harmonic signals for which there are spectral cues to pitch, listeners can perceive the periodicity information in the modulated noise and use it to recognize tone (Burns and Viemeister, 1976). The amplitude and duration cues also play significant roles for tone recognition, and contribute to tones 3 and 4 the most. The studies showed an inconsistency of effects of duration. Duration contributed to the identification of both tones 3 and 4 in Xu *et al.* (2002), but only affected tone 3 in Fu and Zeng (2000). It appears that duration did not consistently contribute to tone 4 recognition in the latter study because tone 4 was not significantly different from tone 1 in duration. In Xu *et al.*, tone 4 was the shortest tone. Therefore, although duration can sometimes be used to aid tone recognition, it varies in natural tokens so it may not always be a reliable cue.

The aim of the current study was to investigate the extent to which tonal contrasts can be recognized with different acoustic cues. Although these three main cues to tonal contrasts in Mandarin (f0, amplitude envelope, and duration) have been investigated in several studies, only the study by Fu and Zeng (2000) systematically combined different cues to examine their contributions. Modulated noises were used

to carry all possible combinations of the three cues. However, the pitch information was conveyed by the temporal fluctuations in the speech envelope which is a far less salient cue than that which arises from true periodic sounds. Here, we introduce a sawtooth wave form created period by period to match the f0 of original speech, and used this to carry pitch information. The f0-controlled sawtooth wave form contains both spectral and temporal cues to pitch, and was expected to give a clearer indication of voice pitch than would be possible from the temporal envelope of noise alone. The same three cues in all combinations were conveyed both by sawtooth and noise carriers [the latter as in Fu and Zeng (2000), but with some modifications in signal processing]. With the same testing materials, the results of the present study allow the comparisons of the level of correct tone recognition with different combinations of acoustic cues.

Stimuli with explicit f0 information (f0 carried by sawtooth carriers) were expected to be recognized the best among all stimuli. For stimuli with the relatively weak temporal f0 information (f0 carried by temporal fluctuations of noise carriers), listeners were expected to be able to identify some tonal information, but their performance would not be as good as the performance when f0 was represented explicitly. It is well known that amplitude-modulated noise stimuli, with no regularity in fine structure, lead to weak pitch sensations. According to Lin's (1988) study, when explicit f0 is presented, amplitude envelope and duration are unlikely to contribute much to the recognition of tonal contrasts. However, when f0 information is absent or presented in a rather weak form, as in Whalen and Xu (1992) and Fu *et al.* (1998), amplitude envelope and/or duration could be of use. Even so, the effect of amplitude envelope and/or duration would be less significant.

II. METHODS

A. Speech stimuli

Four syllables (/i/, /ba/, /fu/ and /tɕʰi/) with each of four tones were used as stimuli. Four different syllable structures (vowel only, three point vowels with initial plosive, fricative, and affricative consonants) were included, so as to contain more variability and be more representative of the information in real life. All these syllables were Mandarin words with a high frequency of occurrence. Natural models for the stimuli were produced by two young adults, one male and one female, who were native speakers of Mandarin from Taiwan. All stimuli were randomized and produced in the carrier phrase "qǐ tiao chu-" ("Please choose—") to avoid inconsistency in pitch range. Six repetitions were produced for the four syllables with each of the four tones. The recording was conducted in an anechoic chamber with speech and laryngograph signals (*Lx*) recorded at the same time using a Sony DAT recorder at a 44.1 kHz sampling rate. The first two tokens sounded natural and contained no creaky voice, as judged by a native speaker of Mandarin (the first author), and so were selected as stimuli. There were a total of 64 speech stimuli (4 syllables X 4 tones X 2 tokens X 2 speakers).

TABLE I. Mean fundamental frequency (Hz), rms amplitude (dB re 1, where all amplitudes are within the range ±1), and duration for the four tones (standard deviations in parentheses).

		Fundamental frequency (Hz)				rms amplitude (dB re 1)	Duration (ms)
		Male		Female			
Tone 1	Average	158	(6)	257	(12)	−23.5 (3.5)	308.8 (38)
Tone 2	Initial	121	(16)	201	(11)	−27.7 (2.9)	347.7 (47)
	Final	143	(15)	222	(15)		
Tone 3	Initial	120	(13)	198	(15)	−30.7 (2.5)	323.7 (66)
	Turning point	91	(9)	113	(41)		
	Final	119	(9)	147	(37)		
Tone 4	Initial	171	(32)	264	(18)	−25.0 (3.0)	256.2 (72)
	Final	108	(14)	150	(42)		
Average						−26.7 (3.6)	309 (65)

B. Acoustic analysis of speech stimuli

Table I shows the f0 values of average pitch curves, mean rms amplitude, and duration for the four tones.

1. F0 contour

Figure 1 shows pitch contours for all speech tokens by the male and female speakers. Note that the pitch contours of tone 3 for the female appeared to be more variable. While the male speaker used the falling–rising pattern consistently, the female speaker used both falling–rising and low-falling patterns. Some of the pitch contours kept falling throughout while others rose slightly at the end.

2. Overall intensity

An analysis of variance (ANOVA) showed a significant difference of rms amplitude of the four tones [$F(3,60)=17.22$, $p<0.001$]. *Post-hoc* comparisons revealed that tone 3 was the least intense tone with significantly lower intensity than the other three tones. Tones 1 and 4 had no significant difference between their overall intensities, and both had significantly higher levels than tone 2.

3. Duration

Tone 2 had the longest duration, followed by tone 3, then tone 1, and finally tone 4. An ANOVA showed a significant effect of duration [$F(3,60)=7.3$, $p<0.001$]. *Post-hoc* comparisons revealed that tones 2 and 3 were significantly longer than tone 4.

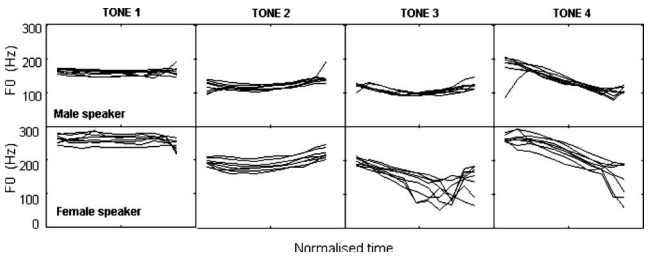

FIG. 1. F0 contours for speech stimuli from the two speakers (all the pitch contours were normalized using a linear time scaling to give the same duration).

TABLE II. Summary of stimuli used in the present study and the study of Fu and Zeng (2000). Cues *A*, *F*, and *D* represent amplitude contour, f0, and duration, respectively.

Cues		Carrier		
		Present study		Fu and Zeng 2000
		Sawtooth carrier	*Noise* carrier	*Noise* carrier
Natural duration	*AFD*	f0-controlled sawtooth carrier envelope extracted at 30 Hz	*AFDcis* Noise carrier × envelope extracted at 400 Hz *AFD* f0-controlled sinusoidal modulated noise × envelope extracted at 30 Hz	Noise carrier × envelope extracted at 500 Hz
	AD	Random-frequency sawtooth carrier envelop extracted at 30 Hz	Noise carrier × envelope extracted at 30 Hz	Noise carrier × envelope extracted at 50 Hz
	FD	f0-controlled sawtooth carrier	f0-controlled sinusoidal modulated noise	f0-controlled 100% amplitude modulated noise (no further detials provided)
Duration fixed at 309 ms	*AF*	Sawtooth carrier controlled by the 309 ms time-scaled f0 contour × time-scaled envelope extracted at 30 Hz	Time-scaled f0-controlled sinusoidal modulated noise × time-scaled envelope extracted at 30 Hz	Linear interpolation was applied to AFD-stimulito give a fixed duration at 400 ms (the range of f0 variation may change althoughthe same percentile of f0 change remains)
	A	309 ms random-frequency sawtooth carrier × time-scaled envelope extracted at 30 Hz	Noise carrier × time-scaled envelope extracted at 30 Hz	Linear interpolation was applied to *AD* stimuli togive a fixed duration at 400 ms (the range of f0 variation may change)
	F	Sawtooth carrier controlled by the 309 ms time-scaled f0 contour	Time-scaled f0-controlled sinusoidal modulatednoise	Linear interpolation was applied to *FD* stimuli(the range of f0 variation may change)
Duration cue only	*D*	Random-frequency sawtooth carrier with original duration	Noise carrier with original duration	Noise carrier with original duration

While tone 3 was often reported as the longest tone in many early studies (e.g., Howie, 1976; Tseng, 1990; Whalen and Xu, 1992; Fu and Zeng, 2000), tone 3 had a shorter duration than tone 2 in this study (348 and 324 ms for tones 2 and 3, respectively). The short tone 3 might be due to the female speaker using both falling–rising and low-falling patterns. As mentioned in Sec. I, the falling–rising tone 3 had the longest duration and the low-falling tone 3 had the shortest duration. Both our speakers were native Mandarin speakers from Taiwan, while most speakers in earlier studies were native speakers from Beijing. The short tone 3 might be a dialect difference between Beijing and Taiwan.

Another possible reason for the short tone 3 was the speech stimuli used in the current study. The speech stimuli in our study included vowels in four different phonetic contexts (vowels only and with different initial consonants), while stimuli in most of early studies used vowels only. The potential effects of the four phonetic contexts might result in the differences in duration relative to other studies. A short tone 3 can also be found in previous studies. For instance, Howie (1976) reported some of the recorded syllables had longer tone 2 than tone 3 (e.g., syllable /hu/ had 260 and 235 ms vowel duration for tones 2 and 3, respectively; syllable /yan/ had 190 and 180 ms vowel duration for tones 2 and 3, respectively).

C. Signal processing

All stimuli in this study were simplified from the originally recorded speech stimuli. Table II summarizes the method used to generate these stimuli, compared to Fu and Zeng's (2000) study. Each speech stimulus was used to produce 15 simplified stimuli with all possible combinations of f0 (*F*), amplitude envelope (*A*), and duration (*D*) in isolation, pairs, and all together, and carried by either sawtooth or noise carriers. Figure 2 shows some examples of the simplified stimuli.

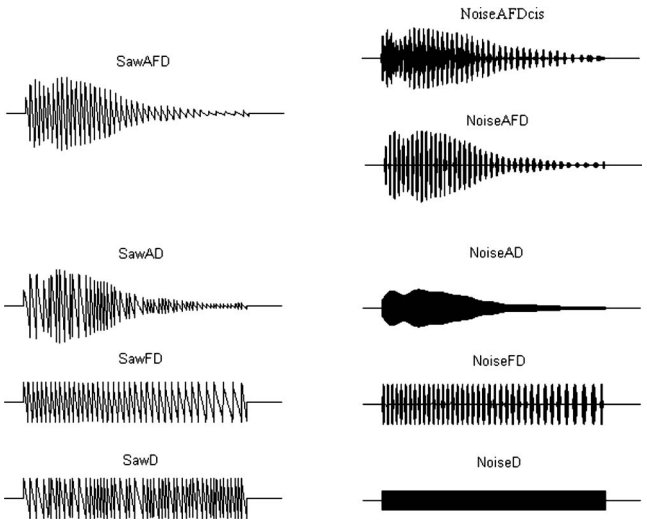

FIG. 2. Examples of some simplified stimuli (conditions *AFD*, *AD*, *FD*, and *D*, using both sawtooth and noise carriers and *NoiseAFDcis*) for the syllable /ba/ with tone 4 produced by the male speaker. Other simplified stimuli *AF*, *A*, and *F* are similar to *AFD*, *AD*, and *FD*, respectively, except with duration fixed at 309 m.

1. AFD (amplitude envelope, f0, and duration)

a. Sawtooth carrier The times between successive regular vocal fold closures (referred to as *Tx*) were measured from laryngograph signals, and used to generate a sawtooth wave form period by period. The f0-controlled sawtooth wave form preserved the f0 contour of the original speech with constant amplitude contour. The amplitude envelope was extracted by full-wave rectification, followed by forward and backward filtering with a 30 Hz cutoff frequency fourth-order elliptical low-pass filter. The f0-controlled sawtooth carrier was then multiplied by the amplitude envelope to generate the *SawAFD* stimulus.

b. Noise carrier In the present study, the signal processing used to generate the *AFD* stimulus by a noise carrier was different from that used in Fu and Zeng's study (2000). A sinusoidal wave form was generated from *Tx* values period by period, half-wave rectified, and then used to modulate a noise carrier. The *NoiseAFD* stimulus contained all three cues, but the temporal cue to pitch was expected to be rather weak. Another condition, which was almost the same as one used by Fu and Zeng (2000), except using a slightly different cutoff frequency (see Table II), was also generated. A noise carrier was multiplied by a speech envelope extracted by half-wave rectification and low-pass filtering with a 400 Hz cutoff frequency to generate the *NoiseAFDcis* stimulus ("cis" refers to "continuous interleaved sampling, CIS," one commonly used speech coding strategy in current cochlear implant systems). Since the range of voice pitch was included, the envelope with 400 Hz cutoff frequency would be expected to preserve some pitch information. The *NoiseAFD* stimulus, because of its clearer pattern of f0-related modulations, might contain slightly better pitch information than the *NoiseAFDcis* stimulus, although the difference should not be large (Green *et al.*, 2002). In any case, the use of an explicit modulator based on f0 allowed condition *NoiseAFD* to be comparable to condition *NoiseFD*, and also to make the f0 information contained in the noise conditions more comparable to that contained in the sawtooth carriers.

2. AD (amplitude envelope and duration)

a. Sawtooth carrier. A random-frequency sawtooth carrier was generated by taking *Tx* randomly from the *Tx* values of the 64 original speech signals across four tones and two speakers, with the same duration as the original speech. The random-frequency sawtooth carrier eliminated the f0 cue to tonal contrasts. A fixed frequency sawtooth carrier was not used because a constant f0 would sound like tone 1, the level tone. The random-frequency sawtooth carrier was then multiplied by the amplitude envelope to generate the *SawAD* stimulus.

b. Noise carrier. A noise carrier was multiplied by the amplitude envelope extracted at 30 Hz to generate the *NoiseAD* stimulus.

3. FD (f0 and duration)

a. Sawtooth carrier. The *SawFD* stimulus was the f0-controlled sawtooth carrier, which preserved the f0 contour and the duration of the original speech but no variation in amplitude.

b. Noise carrier. The *NoiseFD* stimulus was the f0-controlled sinusoidally modulated noise. This stimulus contained the temporal cue to pitch with constant amplitude envelope.

4. AF (amplitude envelope and f0)

a. Sawtooth carrier. A linear time scaling was applied to the f0 contour so as to give a constant duration of 309 ms (the average duration of the voiced parts of all original speech signals), thus eliminating the duration cue. To create a sawtooth wave form with the original f0 contour but a fixed duration at 309 ms, the following procedure was used. The fundamental periods measured from the larygograph signal were converted to a pitch contour first. Then, a linear time scaling was applied to the pitch contour to produce a new pitch contour with the same shape but duration fixed at 309 ms. The new pitch contour was converted back to a new set of *Tx* periods. The first *Tx* period was the same as that in the original speech. The second *Tx* period was determined by the f0 value in the time-scaled pitch contour at the end of this first period. Similarly, the third and subsequent *Tx* periods were determined from the f0 value of the time-scaled contour at the end of the immediately preceding period. The new sets of *Tx* values preserved the original pitch contours without changing the overall f0 range.

The amplitude envelopes extracted at 30 Hz were also scaled in time to give the 309 ms duration, and then multiplied against the sawtooth carriers controlled by time-scaled f0 contour. The *SawAF* stimulus preserved both the f0 and amplitude contours in the original speech but with duration fixed at 309 ms.

b. Noise carrier. A sinusoidal wave form was generated from the time-scaled *Tx* values and half-wave rectified, and then used to modulate a noise carrier. The time-scaled f0-controlled sinusoidal wave form was then multiplied by the time-scaled amplitude envelopes extracted at 30 Hz. The *NoiseAF* stimulus contained the same temporal pitch and amplitude information as in *NoiseAFD* stimulus but without the duration cue.

5. A (amplitude envelope only)

a. Sawtooth carrier. A random-frequency sawtooth carrier with duration fixed at 309 ms was multiplied by the time-scaled amplitude envelope extracted at 30 Hz, resulting

in the *SawA* stimulus.

b. Noise carrier. A noise carrier was multiplied by the time-scaled amplitude envelope extracted at 30 Hz, resulting in *NoiseA* stimulus.

6. F (f0 only)

a. Sawtooth carrier. The *SawF* stimulus was the constant amplitude sawtooth carrier generated from the time-scaled *Tx* values.

b. Noise carrier. The *NoiseF* stimulus was the time-scaled f0-controlled sinusoidal modulated noise. This stimulus contained the temporal cue to pitch with a constant amplitude envelope and a fixed duration.

7. D (duration only)

a. Sawtooth carrier. The *SawD* stimulus was a fixed-amplitude random-frequency sawtooth carrier with duration varied as in the original speech.

b. Noise carrier. The *NoiseD* stimulus was a noise carrier with the duration of the original speech.

D. Subjects

Eight normal hearing adults, three male and five female, participated in the experiment. All were native speakers of Mandarin from Taiwan, aged between 27 and 35.

E. Procedure

Stimuli were blocked by condition, so were presented in 15 blocks. A graphical user interface built in MATLAB was used for running the experiment. In each block, a learning session and a training session were given before a testing session started. Subjects were given sample stimuli for all 15 processing conditions in the learning session, and these stimuli were played randomly in the training session with feedback given. Subjects were allowed to spend as much time as desired in these two sessions. No stimuli used for familiarization were used in the testing session. In the testing session, the four corresponding Mandarin characters of the stimulus were shown first, and the stimulus was played. The Mandarin characters with the Chinese phonetic alphabet, the jùyīnfúhaù, were also shown on the screen. Four push buttons labeled with Tone 1 to Tone 4 were presented below the corresponding characters. Subjects made their identification response using a computer mouse to click one of the four buttons. Then the four corresponding characters of the next stimulus were shown, and the next stimulus was played. No feedback was given in the testing session.

Stimuli were presented through Sennheiser HD 414X headphones, with intensity varied randomly within a 3 dB range around the original level on each trial to reduce cues derived from the overall intensity of the stimuli. The 3 dB range was chosen as it approximated the standard deviation of the rms amplitude of the natural tokens. The order of the 15 processing conditions was randomized and counterbalanced across subjects. In each block, stimuli from male and female speakers were presented in a randomized order. The 15 blocks with 64 simplified stimuli in each resulted in a total of 960 stimuli.

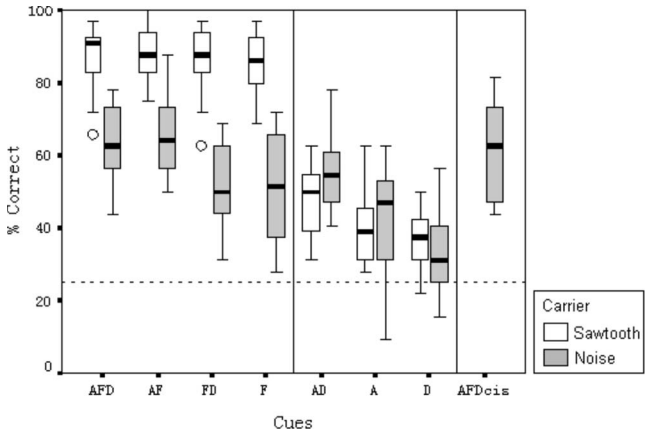

FIG. 3. Boxplots of percentage of correct tone recognition across conditions of different acoustic cues carried by sawtooth or noise carriers. The box represents the 25 to 75 percentile range of the data over subjects and speakers, and the bar within each box represents the median. The whiskers represent the range of data points, except for outliers which are shown as asterisks (more than 3 box lengths from the box edge) or open circles (more than 1.5 box lengths). The dashed line represents chance performance (25% correct).

III. RESULTS AND DISCUSSION

A. Contribution of f0, amplitude envelope, and duration

For stimuli in which f0 information is conveyed by sawtooth carriers (*SawAFD*, *SawAF*, *SawFD*, and *SawF*), a pitch percept can be evoked from both resolved lower harmonics and unresolved higher harmonics. For stimuli with f0 carried by noise carriers (*NoiseAFD*, *NoiseAF*, *NoiseFD*, *NoiseF*, and *NoiseAFDcis*), pitch can only be derived from regular fluctuations in the amplitude of the noise. Since pitch perception in the normal auditory system is mainly determined by resolved lower frequency harmonics (e.g., Moore and Peters, 1992; Ritsma, 1967), stimuli *SawAFD*, *SawAF*, *SawFD*, and *SawF* would be expected to give the most clear information about tonal contrasts. Figure 3 shows the percentage correct tone recognition for the various acoustic cues. Conditions with f0-controlled sawtooth carriers (*SawAFD*, *SawAF*, *SawFD*, and *SawF*) led to the highest levels of performance (around 90% correct), but conditions with other acoustic cues also allowed tone recognition to some extent (varying between 33% and 65%). Performance on all conditions was significantly above chance, except marginal effects for conditions with the duration cue only (36.3% and 33.6% for conditions *SawD* and *NoiseD*; a binomial distribution reveals that scores of 23 or more out of 64, or over 35.9%, are statistically different from chance performance, $p < 0.05$).

Conditions *NoiseAFDcis* and *NoiseAFD* would be expected to carry similar information for tonal contrasts, and an *a priori* comparison confirmed that there was no significant difference between performances on these two conditions. *A priori* comparisons were also carried out to examine the effect of periodicity, amplitude envelope, and duration, by using contrasts of conditions with and without one of these three cues (for instance, comparison of conditions *NoiseAFD/NoiseAF/NoiseFD* vs *NoiseAD/NoiseA/NoiseD* was used to examine the effect of periodicity; contrasts of

TABLE III. Results of Bonferroni post hoc comparisons for pairs of conditions with different acoustic cues. An asterisk indicates a significance level of 0.05.

	Sawtooth Carrier							Noise Carrier						
	AFD	AF	FD	F	AD	A	D	AFD	AF	FD	F	AD	A	D
AFD														
AF	-							-						
FD	-	-						-	*					
F	-	-	-					-	*	-				
AD	*	*	*	*				-	-	-	-			
A	*	*	*	*	-			-	-	-	-	-		
D	*	*	*	*	*	*		*	*	-	-	-	-	

NoiseAFD/AF/AD vs *NoiseFD/F/D*, and *NoiseAFD/FD/AD* vs *NoiseAF/F/A* were used for amplitude envelope and duration, respectively). These revealed that there were significant differences for the periodicity and amplitude envelope cues ($[F(1,7)=20.3,\ p<0.01]$ and $[F(1,7)=48.8,\ p<0.01]$, respectively), but not for the duration cue.

A repeated-measures ANOVA was performed for factors of cue, carrier, and speaker. This showed significant effects of carrier $[F(1,7)=131.3,\ p<0.001]$ and cue $[F(6,42)=51.3,\ p<0.001]$, and significant interactions of carrier by cue $[F(6,42)=27.3,\ p<0.001]$ and speaker by carrier $[F(1,7)=15.0,\ p<0.05]$. No other two-way or three-way interaction was significant.

The interaction of carrier by cue was almost certainly due to performance with the two carriers differing greatly only in those conditions in which f0 information was presented (*AFD*, *AF*, *FD*, and *F*). Bonferroni-corrected *post-hoc* comparisons, with an alpha of 0.05, were used to examine the effect of the two different carriers on each of seven conditions, revealing that this was indeed the case. All conditions involving f0 information (*AFD*, *AF*, *FD*, and *F*) were significantly different according to carrier, whereas conditions which had no f0 information (*AD*, *A*, and *D*) were statistically equal. This confirmed our expectation that a sawtooth carrier generated period by period from f0 provided better information for tone recognition than could possibly be conveyed by a modulated noise, with no significant difference for using a random-frequency sawtooth carrier and a noise carrier for conveying information about amplitude envelope and duration.

To examine the relative importance of the different acoustic cues to tonal contrasts, the effect of different combinations of cues on tone recognition was also compared, with sawtooth and noise carriers separately. Table III summarizes the results. For the sawtooth carrier, no significant difference was found between the four conditions with salient f0 (*SawAFD*, *SawAF*, *SawFD*, and *SawF*), and they were all significantly higher than the three conditions without f0 information. There was no significant difference between the two conditions with amplitude envelope cue presented (*SawAD* and *SawA*), and they were significantly higher than the condition with duration only (*SawD*). This indicated that information about f0 conveyed by a periodic sound contributed to tone recognition the most, and neither amplitude envelope nor duration affected tone recognition

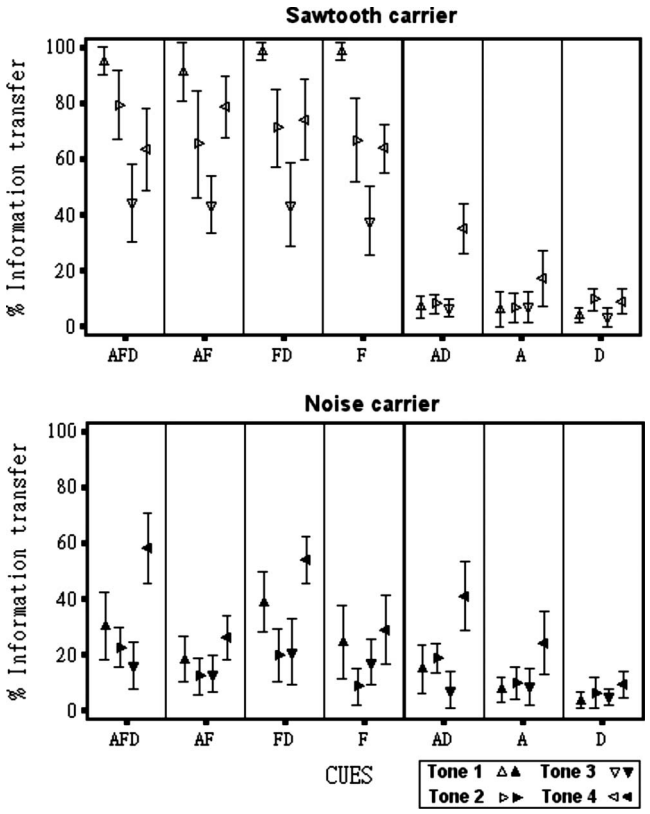

FIG. 4. Percentage of information transfer score of tone recognition for four tones across conditions with different combinations of three acoustic cues. Performance on conditions using the sawtooth carrier is displayed in the upper panel, and performance on conditions using the noise carrier is displayed in the lower panel. Error bars indicate the 95% confidence interval for the means averaging over 16 data points (8 subjects × 2 speakers).

performance while f0 was present. In the absence of salient f0, amplitude envelope appeared to contribute more to tone recognition than duration.

For noise carriers, only a few comparisons showed significant effects, even though several pairs of conditions had considerably different means (for instance, the mean score of *NoiseAFD* was more than 20% higher than that of *NoiseD*, yet, not significant). This was due to there being larger variations in some conditions than others, and the Bonferroni correction leads to relatively conservative tests.

B. Performance for the four tones

Previous studies have shown that the recognition for the four tones in Mandarin can vary with different acoustic cues available (e.g., Whalen and Xu, 1992; Fu *et al.*, 1998; Fu and Zeng, 2000). Here, performance for the four tones in conditions with different combinations of cues was further examined. To give an unbiased measure of identification performance for individual tones, information transfer scores were computed using the procedures of Miller and Nicely (1955) from tone confusion matrices (e.g., a 2 × 2 matrix classifying stimuli as tone 1 versus all other tones, and responses in the same way). The percentage of information transferred for the four tones across conditions is shown in Fig. 4.

A repeated-measures ANOVA was performed for factors of tone and condition, demonstrating a significant interac-

TABLE IV. (a) Response matrices (in percentage) for stimuli with the amplitude envelope cue only (results were summed over conditions *SawA* and *NoiseA*). (b) Cross correlations between amplitude contours and pitch contours for male and female speakers.

(a)		Response to male speaker				Response to female speaker			
	Stimulus	1	2	3	4	1	2	3	4
	Tone 1	31.3	43.0	16.5	9.4	35.9	28.9	16.4	18.8
	Tone 2	15.6	48.4	25.8	9.4	19.5	42.2	26.6	11.7
	Tone 3	7.8	44.5	38.3	9.4	17.2	19.5	25.8	37.5
	Tone 4	23.4	10.2	13.3	53.3	17.2	9.4	16.4	57.0
(b)	Amp	Male speaker F0				Female speaker F0			
		1	2	3	4	1	2	3	4
	Tone 1	0.38	−0.07	0.22	0.69	0.62	−0.01	0.69	0.72
	Tone 2	0.34	0.68	0.45	0.04	0.33	0.45	0.33	0.25
	Tone 3	0.43	0.40	0.78	0.51	0.43	−0.02	0.79	0.85
	Tone 4	0.41	−0.01	0.37	0.77	0.49	−0.19	0.88	0.89

tion $[F(39,273)=18.2,\ p<0.001]$, and significant main effects for condition $[F(13,91)=96.5,\ p<0.001]$ and tone $[F(3,21)=65.7,\ p<0.001]$. Bonferroni-corrected *post-hoc* comparisons were used to examine performance of the four tones on conditions with different combinations of cues. For conditions with explicit f0 included (*SawAFD*, *SawAF*, *SawFD*, and *SawF*), information transfer scores for the four tones showed similar patterns: Tone 1 was significantly higher than tones 2 and 4, and there was no significant difference between these two tones. All these three tones were significantly higher than tone 3. Tone 1 might be recognized best because it had a higher and more distinct frequency range compared to the other three tones. As for tone 3, the variation of its realization presumably was responsible for its relatively low scores. In the two conditions with the duration cue only (*SawD* and *NoiseD*), information transfer scores were all very low, with not much difference among the four tones. For the rest of conditions with little or no f0 presented, tone 4 was generally recognized well compared to the other three tones.

C. The use of amplitude envelope for tone recognition

Table IV(a) shows subject response for stimuli with the amplitude envelope cue only. The confusion matrices were summed over conditions *SawA* and *NoiseA* for male and female speakers separately. Percent correct for each of the four tones is shown along the diagonal of each table. Results from previous studies have shown that listeners were able to use the amplitude cue to label tones to some extent, though performance based on the amplitude envelope cue was often highly variable across tones and speakers. Whalen and Xu (1992) suggested that listeners either recognized the consistent correlation between pitch and amplitude contours or interpreted amplitude change as f0 change. The variation in performance might arise from the fact that a significant correlation between pitch and amplitude contours was only found for certain tones and speakers (Whalen and Xu, 1992; Fu and Zeng, 2000). To examine if subject responses could be explained by the similarity of pitch and amplitude contours, the correlations between pitch and amplitude contours were calculated and compared with the frequency of subject response for stimuli with the amplitude envelope cue only.

The similarity of the amplitude contour of one tone to the pitch contour of each of the four tones was examined by calculating a cross correlation. The average pitch and amplitude contours, averaged over eight normalized contours with the duration of 309 ms of the four tones for the male and female speaker, were used (see Fig. 5). To obtain approximately independent samples, points along a contour were sampled separated by an interval determined by calculating autocorrelation coefficients for pitch and amplitude contours of several sentences. Since a sample point and its neighbor become less related to each other when they are further apart,

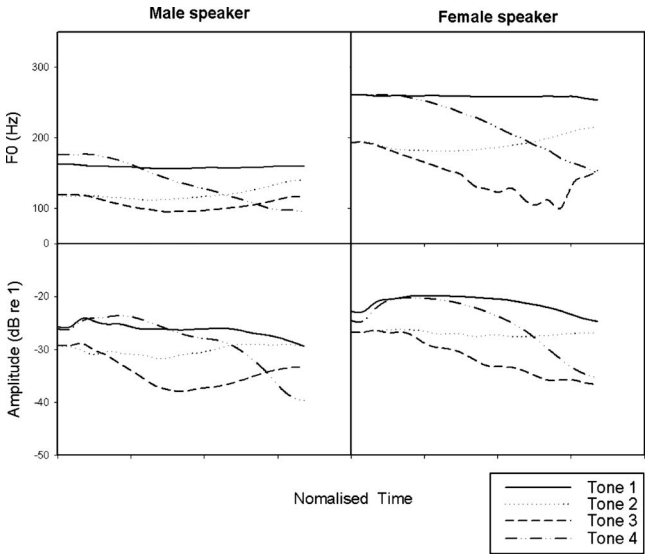

FIG. 5. Average pitch and amplitude contours for the four tones for the two speakers. Each line was averaged over eight speech tokens after a linear time scaling.

the autocorrelation coefficients were often not significant when the time lag was over 55 ms. Therefore, six samples for each of normalized amplitude and pitch contours were calculated by dividing one contour into six sections and taking the average value in each section. Correlations for the lags of −1, 0, and +1 were calculated, and the maximum value was selected. Table IV(b) shows the average correlation between the amplitude contour of one tone and the pitch contours of the four tones. The amplitude contour of a tone did not always correlate most highly with the pitch contour of that same tone. For instance, the amplitude contour of tone 1 for the male speaker was highly correlated to the pitch contour of tone 4 (r=0.69) rather than that of tone 1 (r =0.38). Also, for the female speaker, the amplitude contour of tone 3 was highly correlated with the pitch contours of both tone 3 and tone 4 (r=0.79 and 0.85, respectively), so as the amplitude contour of tone 4 to the pitch contours of tone 4 and tone 3 (r=0.89 and 0.88, respectively). This might be due to, as shown in Fig. 1, the female tone 3 sometimes appeared as low-falling pattern which could correlated well with tone 4, the falling tone.

To further investigate whether response frequencies to stimuli with only the amplitude envelope cue [Table IV(a)] could be explained by the similarity between pitch and amplitude contours [Table IV(b)], the relationship between subject responses and the correlation of pitch and amplitude was examined. Generally speaking, high response frequencies were associated with high correlation coefficients. Although a high correlation coefficient did not always lead to a high proportion of subject responses, a low correlation coefficient seldom did. This indicated that a certain relationship between identification responses and the correlation between amplitude and pitch contour exists.

IV. GENERAL DISCUSSION

While lexical tones are mainly distinguished by their f0 patterns, early studies often suggested that other acoustic cues such as amplitude and duration were of little importance for tone recognition. Recent studies have reported that these temporal cues have a more substantial contribution when f0 information is weak or completely absent. However, results from the present study clearly demonstrate that performance with any combination of these temporal cues was still much lower than with explicit f0 information. The results here might be used to indicate the extent to which information about tonal contrasts could possibly be obtained by current cochlear implant users with CIS-like speech processors, and what may possibly be achieved in future devices. While voice f0 is indubitably the most essential cue for recognizing tonal contrasts, it is not transmitted sufficiently well through current implants. The f0 information available in implant systems is very similar to that conveyed by temporal fluctuations of a noise carrier in the present study. Without the presence of explicit f0 information, this temporal pitch information could help tone recognition, as could amplitude envelope and duration. However, even with all available temporal cues, tone recognition performance was hardly above 70% correct. This is consistent with clinical results, which have reported that users of tonal languages often encounter great difficulty in tone recognition (e.g., Barry *et al.*, 2002; Peng *et al.*, 2004, Wei *et al.*, 2004).

While these temporal cues (temporal coding of f0, amplitude envelope, and duration) could aid in tone recognition to a certain extent and could be used by implant users, none of them is always reliable. Listeners often vary greatly in their ability to make use of these temporal cues, as shown in this study (see Fig. 3) and in previous research. For instance, in Green *et al.* (2002), glide labeling performance based on temporal envelope cues was limited and varied greatly across subjects. Also, the amplitude envelope and duration cues in Mandarin have been reported to be highly variable across speakers and different tones (Fu and Zeng, 2000). The performance in those conditions with noise carriers may represent what can be achieved when listeners are able to perceive one, two, or all of these three cues. The average performance across conditions varies greatly between 35% and 65%, and this low level is about what has been observed in many implant users. Furthermore, these temporal cues are likely to be even less salient in a real life situation of running speech in less-than-ideal acoustic environments.

V. SUMMARY

As would be expected, the explicit f0 cue contributed to tone recognition the most, irrespective of the presence of amplitude envelope and/or duration cues. In the absence of explicit f0, the temporal f0, amplitude envelope, and duration cues also contributed to tone recognition to different extents: the temporal coding of f0 and amplitude envelope both demonstrated substantial contributions to tone recognition, while duration had only a marginal effect. The differences in the utility of duration relative to other studies might be due to: (a) a dialect difference between Beijing and Taiwan and/or (2) the speech syllables used here including different phonetic contexts (vowels only and with different initial consonants). Performance based on these secondary cues varied greatly across subjects, indicating listener variability in perceiving these temporal cues. The contribution of the amplitude envelope cue to the identification of certain tones, especially for tone 4, might be explained by the relatively high correlation between their amplitude and pitch contours.

ACKNOWLEDGMENTS

Thanks go to the Chiang Ching-kuo Foundation for International Scholarly Exchange (Taiwan) for the Ph.D. dissertation fellowship to the first author (DF 020-U-02). We also thank Andrew Oxenham and two anonymous reviewers of the current manuscript and the anonymous reviewers of the early version of this paper, for their helpful comments.

Barry, J. G., Blamey, P. J., Martin, L. F. A., Lee, K. Y. S., Tang, T., Ming, Y. Y., and van Hasselt, C. A. (**2002**). "Tone discrimination in Cantonese-speaking children using a cochlear implant," Clin. Linguist. Phonetics **16**, 79–99.

Blicher, D. L., Diehl, R. L., and Cohen, L. B. (**1990**). "Effects of syllable duration on the perception of the Mandarin Tone2/Tone 3 distinction: Evidence of auditory enhancement," J. Phonetics **18**, 37–49.

Burns, E. M., and Viemeister, N. F. (**1976**). "Nonspectral pitch," J. Acoust. Soc. Am. **60**, 863–869.

Chao, Y. R. (**1948**). *Mandarin Primer: An Intensive Course in Spoken Chinese* (Harvard University Press, Cambridge, MA).
Chao, Y. R. (**1968**). *A Grammar of Spoken Chinese* (University of California Press, Berkeley, CA).
Chen, M. Y. (**2000**). *Tone Sandhi: Patterns Across Chinese Dialects* (Cambridge University Press, Cambridge).
Fon, J., and Chiang, W. Y. (**2000**). "What does Chao have to say about tones?—A case study of Taiwan Mandarin," Can. J. Phys. **27**, 13–37.
Fu, Q. J., and Zeng, F. G. (**2000**). "Identification of temporal envelope cues in Chinese tone recognition," J. Speech Lang. Hear. Res. **5**, 45–57.
Fu, Q. J., Zeng, F. G., Shannon, R. V., and Soli, S. D. (**1998**). "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am. **104**, 505–510.
Gårding, E., Kratochvil, P., Svantesson, J. O., and Zhang, J. (**1986**). "Tone 4 and Tone 3 discrimination in modern Standard Chinese," Lang. Speech**29**, 281–293.
Green, T., Faulkner, A., and Rosen, S. (**2002**). "Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants," J. Acoust. Soc. Am. **112**, 2155–2164.
Howie, J. M. (**1976**). *Acoustic Studies of Mandarin Vowels and Tones* (Cambridge University Press, Cambridge).
Lin, H. B., and Repp, B. (**1989**). "Cues to the perception of Taiwanese tones," Lang Speech **32**, 25–44.
Lin, M.-C. (**1988**). "The acoustic characteristics and perceptual cues of tones in Standard Chinese," Chin. Yuwen **204**, 182–193.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Moore, B. C. J., and Peters, R. W. (**1992**). "Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity," J. Acoust. Soc. Am. **91**, 2881–2893.
Peng, S. C., Tomblin, J. B., Cheung, H., Lin, Y. Y., and Wang, L. S. (**2004**). "Perception and production of Mandarin tones in prelingually deaf children with cochlear implants," Ear Hear. **25**, 251–264.
Ritsma, R. J. (**1967**). "Frequencies 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 linguistics aspects," Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.
Shen, X. S., and Lin, M. (**1991**). "A perceptual study of Mandarin Tones 2 and 3," Lang Speech **34**, 145–156.
Shih, C. L. (**1988**). "Tone and intonation in Mandarin," Working Pagers Cornell Phonetics Lab. **3**, 83–97.
Tseng, C. Y. (**1990**). *An Acoustic Phonetic Study on Tones in Mandarin Chinese* (The Institute of History and Philology Academia Sinica, Taipei).
Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). "Speech waveform envelope cues for consonant recognition," J. Acoust. Soc. Am. **82**, 1152–1161.
Wei, C. G., Cao, K., and Zeng, F. G. (**2004**). "Mandarin tone recognition in cochlear-implant subjects," 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**). "Spectral and temporal features of stimulation affecting tonal-speech perception: Implication for cochlear prostheses," J. Acoust. Soc. Am. **112**, 247–258.
Yip, M. (**1980**). "The tonal phonology of Chinese," Ph.D. dissertation, MIT; published (Garland, New York, **1990**).
Zee, E. (**1978**). "Duration and intensity as correlates of F0," J. Phonetics **6**, 213–220.