Concurrent-vowel and tone recognitions in acoustic and simulated electric hearing

Xin Luo^{a)}

Communication and Auditory Neuroscience, House Ear Institute, 2100 West Third Street, Los Angeles, California 90057 and Department of Speech, Language, and Hearing Sciences, Purdue University, 500 Oval Drive, West Lafayette, Indiana 47907

Qian-Jie Fu

Communication and Auditory Neuroscience, House Ear Institute, 2100 West Third Street, Los Angeles, California 90057

(Received 16 July 2008; revised 27 February 2009; accepted 5 March 2009)

Because of the poor spectral resolution in cochlear implants (CIs), fundamental frequency (F0) cues are not well preserved. Chinese-speaking CI users may have great difficulty understanding speech produced by competing talkers, due to conflicting tones. In this study, normal-hearing listeners' concurrent Chinese syllable recognition was measured with unprocessed speech and CI simulations. Concurrent syllables were constructed by summing two vowels from a male talker (with identical mean F0's) or one vowel from each of a male and a female talker (with a relatively large F0 separation). CI signal processing was simulated using four- and eight-channel noise-band vocoders; the degraded spectral resolution may limit listeners' ability to utilize talker and/or tone differences. The results showed that concurrent speech recognition was significantly poorer with the CI simulations than with unprocessed speech. There were significant interactions between the talker and speech-processing conditions, e.g., better tone and syllable recognitions with the male-female condition for unprocessed speech, and with the male-male condition for eight-channel speech. With the CI simulations, competing tones interfered with concurrent-tone and syllable recognitions, but not vowel recognition. Given limited pitch cues, subjects were unable to use F0 differences between talkers or tones for concurrent Chinese syllable recognition.

© 2009 Acoustical Society of America. [DOI: 10.1121/1.3106534]

PACS number(s): 43.66.Hg, 43.66.Ts, 43.71.Es, 43.71.Ky [RYL]

Pages: 3223–3233

I. INTRODUCTION

Auditory sensory inputs are used to identify multiple sound sources within complex listening environments, a phenomenon described by Bregman (1990) as “auditory scene analysis.” One of the most challenging listening conditions is simultaneous presentation of competing speech. Segregation and streaming of sound sources allow a target talker to be understood in the presence of competing talkers. Many previous studies have measured identification of two simultaneously presented, synthesized vowel-like sounds, and have shown that the normal auditory system is able to stream and segregate concurrent vowels using acoustic cues such as fundamental frequency (F0) difference, harmonic misalignment, pitch period asynchrony, and formant transitions (e.g., Scheffers, 1983; Assmann and Summerfield, 1990; Summerfield and Assmann, 1991; Assmann, 1995).

For example, Scheffers (1983) and Assmann and Summerfield (1990) showed that normal-hearing (NH) listeners' identification of concurrent vowels was significantly better when the two vowels had different F0's (separated by one to four semitones) rather than the same F0. Based on such perceptual data, computational models have been proposed that

involve voice pitch estimation from the output of the auditory periphery, segregation of competing voices according to the estimated voice pitches, and vowel template matching for the segregated spectral patterns (Assmann and Summerfield, 1990; Meddis and Hewitt, 1992). It is possible that when two concurrent vowels have different F0's, listeners may better attend to the components of each vowel by making use of their respective harmonic and periodic structures. For example, Summerfield and Assmann (1991) found that concurrent-vowel recognition was improved by shifting the harmonics of a component vowel by half of its F0 relative to those of the other vowel, as long as the F0 was high enough (e.g., 200 Hz) and the harmonics were well separated in frequency. They also found that shifting the temporal waveforms of a component vowel by half of its period relative to those of the other vowel (i.e., introducing pitch period asynchrony) improved concurrent-vowel recognition, as long as the F0 was very low (e.g., 50 Hz) and the periods were well separated in time.

Frequency modulation (FM) with either linear gliding or sinusoidal functions may be another acoustic cue for auditory grouping and segregation. For example, Culling and Summerfield (1995) found that identification thresholds for a target vowel were lower when the masking vowel was unmodulated rather than modulated in the same way as the target vowel. Chalikia and Bregman (1989) measured recog-

^{a)}Author to whom correspondence should be addressed. Electronic mail: luo5@purdue.edu

dition of concurrent vowels, whose frequency components were either not modulated, modulated in parallel, linear glides (in log scale), or modulated in opposite, crossing glides. They found that the recognition with crossing glides was significantly better than that with un-modulated stimuli or with parallel linear glides. These results suggest that coherent gliding of frequency components within vowels, as well as incoherent gliding of components across vowels, might benefit segregation and identification of concurrent vowels.

In these above-cited studies, concurrent-vowel recognition has been measured using English vowels. While F0 differences generally benefit concurrent English vowel recognition, identifying F0 variation patterns is not an essential part of English syllable recognition. In contrast, F0 cues are lexically meaningful for tonal languages such as Mandarin Chinese. Mandarin Chinese has four lexical tones: tone 1 (high, flat pitch contour), tone 2 (rising contour), tone 3 (falling-rising contour), and tone 4 (falling contour). Concurrent Chinese syllable recognition provides a unique opportunity to investigate how competing tonal patterns (i.e., pitch contours) may influence concurrent-vowel recognition. Similarly, concurrent-tone recognition can be measured within the same mixture of Chinese syllables. Previous studies (e.g., [Chalikia and Bregman, 1989](#)) have largely focused on the contribution of FM cues to concurrent-vowel recognition, and have paid less attention to identification of modulation properties (e.g., frequency glide direction). Because F0 cues are lexically meaningful, it is also important to understand how Chinese listeners' concurrent-syllable recognition is affected by different vowel and tone pairs. Quite possibly, F0 differences and pitch contours may affect concurrent Chinese syllable recognition differently than concurrent English syllable recognition.

Unlike NH listeners, cochlear implant (CI) users have limited access to F0 cues. Contemporary implant systems typically consist of multiple electrodes (16–22) and use speech-processing strategies based on waveform representation (e.g., [Wilson et al., 1991](#)). Because of the limited number of electrodes and/or limited channel selectivity, the spectral resolution is too poor to resolve F0 and harmonic information. F0 is well represented in the temporal envelopes within individual frequency channels, but CI users are able to extract only some of the temporal pitch information, and only for relatively low F0's ([Green et al., 2004](#)). While other co-varying cues are available (e.g., overall duration and amplitude envelope), CI users' limited pitch perception capabilities result in only moderate levels of performance for Chinese tone recognition (e.g., [Fu et al., 2004](#); [Luo et al., 2008](#)). Given the limited Chinese tone recognition, it is unclear how competing tonal patterns may influence CI users' concurrent Chinese syllable recognition.

CI users' poor speech perception in the presence of competing talkers may also be due to limited F0 coding. [Qin and Oxenham \(2005\)](#) found that NH listeners' concurrent-vowel recognition performance with acoustic simulations of CI processing (even with 24 channels) was significantly poorer than with unprocessed speech. In these acoustic CI simulations, NH listeners were unable to use F0 differences be-

tween concurrent vowels to achieve better recognition performance. Similar results were reported by [Stickney et al. \(2007\)](#), who showed that, for unprocessed speech, increasing the F0 separation between the target and masker sentences gradually improved NH listeners' recognition of target sentences; for CI users or NH listeners listening to acoustic CI simulations, increasing the F0 separation provided no benefit.

In the present study, the effects of CI speech processing on concurrent-vowel and tone recognitions were acoustically simulated in NH listeners. While it might be of more practical interest to investigate target speech recognition in the presence of competing speech, concurrent-vowel and tone recognitions enable detailed analyses of performance and confusion patterns across different vowel and tone pairs, which might provide some insights into the strategies and cues used by NH listeners for sound source segregation in acoustic and simulated electric hearing. In the real CI case, patient-related factors (e.g., etiology of hearing loss, proximity of electrodes to healthy neural populations, and implant device differences) can result in significant inter-subject variability, making it sometimes difficult to measure the effect of a processing parameter change. Acoustic CI simulations allow better control of subject variables within the (presumably) more homogenous group of NH listeners, at least in terms of hearing health. The amount of spectral and temporal cues can be manipulated independently in NH listeners by varying the number of frequency channels and the temporal envelope cutoff frequency, respectively. These manipulations have great relevance for CI users, who must fuse the spectral and temporal cues delivered by electrical stimulation patterns. By using CI simulations in NH listeners, we can more cleanly measure the effect of a processing parameter change that might be important to the real CI case. CI processing was simulated here using a four- or eight-channel noise-band vocoder and a 500-Hz temporal envelope filter in each band. The number of frequency channels and temporal envelope cutoff frequency were chosen to produce overall performance similar to that of CI users with clinically assigned speech processors (e.g., [Friesen et al., 2001](#)), and in turn provide relevant implications for real CI speech perception. We hypothesized that increased spectral resolution would improve concurrent-vowel recognition, consistent with previous studies with single syllables (e.g., [Xu et al., 2002](#)). However, increased spectral resolution has been shown to have a much less effect on tone recognition than on vowel recognition with single syllables (e.g., [Xu et al., 2002](#)). Given the limited F0 cues in CI processing, it is unclear whether increased spectral resolution would improve concurrent-tone recognition. In this study, we explored the contribution of spectral cues (with unprocessed speech, eight- and four-channel CI simulations) to concurrent-vowel and tone recognitions. To investigate the effects of mean F0 difference on concurrent-vowel and tone recognitions, vowels were combined from a male and a female talker to produce a relatively large difference in mean F0, or within the same male talker to produce a nearly identical mean F0.

TABLE I. Ranges for F0, F1, and F2 values for the Chinese single-vowel stimuli, for the male and female talkers.

		Male talker			Female talker		
		F0 range (Hz)	F1 range (Hz)	F2 range (Hz)	F0 range (Hz)	F1 range (Hz)	F2 range (Hz)
/a/	Tone 1	147–155	930–960	1170–1290	277–290	1111–1261	1594–1775
	Tone 2	87–160	960–1050	1200–1322	178–280	1050–1231	1412–1654
	Tone 3	82–137	870–1080	1111–1231	130–202	1200–1260	1654–1805
	Tone 4	82–164	869–1111	1141–1292	129–293	1111–1382	1443–1684
/e/	Tone 1	157–170	356–597	1141–1352	277–296	446–870	1473–1594
	Tone 2	95–175	356–597	1231–1412	194–294	476–839	1352–1563
	Tone 3	82–121	446–658	1141–1292	148–202	507–809	1443–1533
	Tone 4	78–188	416–658	1141–1443	140–333	386–748	1503–1594
/u/	Tone 1	150–161	356–385	718–778	277–304	386–446	627–960
	Tone 2	92–183	356–476	567–809	186–284	386–537	627–960
	Tone 3	78–133	356–385	446–597	138–206	416–446	567–900
	Tone 4	83–206	356–446	446–688	110–320	386–537	658–1020
/i/	Tone 1	157–185	325–386	2348–2530	265–290	295–325	2681–3073
	Tone 2	96–157	235–295	2379–2439	186–266	265–326	3013–3254
	Tone 3	78–127	265–325	2228–2469	156–218	295–416	3194–3254
	Tone 4	78–195	265–356	2379–2439	150–332	356–356	3073–3284

II. METHODS

A. Subjects

Six adult native Chinese-speaking NH subjects (three males and three females) participated in the present study. All subjects had pure-tone thresholds better than 20 dB hearing level (HL) at octave frequencies from 125 to 8000 Hz in both ears. All subjects were very experienced with the acoustic CI simulations from previous experiments.

B. Stimuli and speech processing

While synthesized vowels used in previous studies (e.g., [Scheffers, 1983](#)) allow precise control over vowel parameters (e.g., F0 and harmonics, formant frequencies and bandwidths, duration, amplitude, etc.), these parameters interact dynamically in natural speech. In addition, Chinese syllable synthesis faces a special challenge, namely, how to accurately synthesize tones. Most of the current tone models (data-driven or rule-based) work for continuous speech. To the best of our knowledge, there are no standard F0 contour models for generating isolated tones. In the present study, concurrent Chinese syllable recognition was measured using naturally produced vowels. Single-vowel stimuli were drawn from the Chinese Standard Database, recorded by [Wang \(1993\)](#). One male and one female talker each produced four Mandarin Chinese single-vowels (/a/, /e/, /u/, and /i/ in Pinyin) according to four lexical tones—tone 1 (high, flat), tone 2 (rising), tone 3 (falling-rising), and tone 4 (falling)—resulting in a total of 32 single-vowel syllables (4 vowels $\times$ 4 tones $\times$ 2 talkers). These single-vowel stimuli were digitized using a 16-bit analog/digital converter at a 16-kHz sampling rate, without high-frequency pre-emphasis. Table I

shows the ranges for F0, first formant (F1), and second formant (F2) frequencies for the Chinese single-vowel stimuli.

The concurrent Chinese syllables were constructed by summing either one single-vowel syllable from each of the male and female talkers (male-female condition) or two single-vowel syllables from the male talker (male-male condition). The male-female condition provided a relatively large difference in mean F0, while the male-male condition provided a relatively small difference in mean F0. To align the onsets and offsets of the two component vowels, all single-vowel syllables were normalized to have the same time duration (405 ms for the vowel segments). The duration-normalization was performed by time-stretching or -compressing the input vowel duration to 405 ms, without changing the input pitch and formant frequencies, using an algorithm in Adobe Audition. The time scaling factor for individual single-vowel syllables ranged from 0.6 to 1.3. The duration-normalization made duration cues unavailable for tone recognition, forcing subjects to attend to other cues such as pitch and amplitude envelope. After duration-normalization, the single-vowel syllables were normalized to have the same long-term root-mean-square (rms) amplitude (65 dB). Therefore, the single-vowel components in the concurrent syllables were equated in terms of overall duration and amplitude. After summation, the long-term rms amplitudes of the concurrent syllables were normalized to 65 dB. In both the male-female and male-male conditions, there were a total of 256 concurrent syllables (16 single-vowel syllables from the male talker $\times$ 16 single-vowel syllables from the competing talker). Note that in the male-male condition, each single-vowel syllable was paired with itself once, and each pair of different single-vowel syllables was presented twice.

FIG. 1. Mean Chinese syllable (left panel), vowel (middle panel), and tone recognition scores (right panel) for NH subjects listening to unprocessed speech, eight- or four-channel CI simulation, as a function of talker condition. The numbers on each vertical bar represent the corresponding mean score and standard deviation (shown in the bracket).

Noise-band vocoders with either eight or four frequency channels were used to simulate CI speech processing (Shannon *et al.*, 1995). After pre-emphasis (first-order Butterworth high-pass filter at 1200 Hz), the input speech signal was band-pass filtered into eight or four frequency channels (fourth-order Butterworth filters). The overall input acoustic frequency range was 100–6000 Hz. The analysis bands were evenly distributed in terms of cochlear location according to Greenwood's (1990) formula. The corner frequencies (–3 dB points) were 100, 222, 404, 676, 1083, 1692, 2602, 3964, and 6000 Hz for the eight-channel processor and 100, 404, 1083, 2602, and 6000 Hz for the four-channel processor. The temporal envelope from each band was extracted by half-wave rectification and low-pass filtering (fourth-order Butterworth filter at 500 Hz), and was used to modulate wide-band noise. The amplitude-modulated noise carriers were filtered using the same pass-bands used for the analysis filters. The band-limited, amplitude-modulated noise carriers from all frequency channels were summed to produce the CI simulation speech, which was then normalized to have the same long-term rms amplitude as the input speech signal.

C. Procedures

Both single-syllable and concurrent-syllable recognitions were measured using the original, unprocessed speech, as well as speech processed by acoustic CI simulations. Thus, there were a total of nine experimental conditions [3 talker conditions (single, male-male, and male-female talkers) $\times$ 3 signal-processing conditions (unprocessed speech, eight- and four-channel CI simulations)]. To minimize potential learning effects, the test order of the experimental conditions was randomized across subjects; no learning trends were observed in terms of test order.

Subjects were seated in a double-walled sound-treated booth and listened to stimuli presented in sound field over a single loudspeaker (Tannoy Reveal) at 65 dBA. A closed-set identification task (16 choices) was used to measure both single-syllable and concurrent-syllable recognitions. In each trial, a stimulus was randomly selected from the stimulus set (without replacement) and presented to the subject. In the single-syllable recognition tasks, subjects were instructed to identify the Chinese syllable by clicking on one of the re-

sponse choices shown on the screen: a1, a2, a3, a4, e1, e2, e3, e4, u1, u2, u3, u4, i1, i2, i3, and i4; note that the numbers in the response labels refer to the Chinese tones (1—high, flat, 2—rising, 3—falling-rising, and 4—falling). Responses were collected and scored as the percentage that the Chinese syllable, vowel, or tone was correctly identified. In the concurrent-syllable recognition tasks, subjects were instructed to identify the two Chinese syllables by making two consecutive choices; the order of choices was not important for scoring. Responses were collected and scored as the percentage that both syllables, both vowels, or both tones were correctly identified. No preview, feedback, or training was provided. Note that the subjects were very experienced with the CI simulations from their participation in previous studies, meaning that no training was required to familiarize subjects with the test procedure or the signal processing.

III. RESULTS

Figure 1 shows Chinese syllable, vowel, and tone recognition scores for the six NH subjects listening to single-talker, male-male, or male-female syllables, for processed and unprocessed speech. Note that the different recognition tasks had different chance performance levels. Chance level for syllable recognition with single syllables was 6.25% correct (1/16), while chance level for syllable recognition with concurrent syllables was 0.76% correct ($16/256 \times 1/256 + 240/256 \times 2/256$). Similarly, chance level for vowel or tone recognition with single syllables was 25% correct (1/4), while chance level for vowel or tone recognition with concurrent syllables was 10.94% correct ($4/16 \times 1/16 + 12/16 \times 2/16$). When listening to unprocessed speech, subjects achieved nearly perfect recognition performance with both single and concurrent syllables. When listening to the eight- or four-channel CI simulation, single-talker speech recognition worsened, but remained at relatively high levels (>70% correct). With the CI simulations, recognition performance with concurrent-talker speech ranged from ~15% (e.g., syllables) to ~65% correct (e.g., vowels).

TABLE II. Results from one-way RM ANOVAs performed on single-talker data, with speech processing as the factor. Significant effects are shown in bold. Significant differences ($p < 0.05$) from *post-hoc* Bonferroni t-tests are also shown in bold.

	Speech processing (unprocessed, 8ch, 4ch)			
	dF, res	F-ratio	p-value	Post-hoc $p < 0.05$
Vowels	2,10	29.7	<0.001	Unprocessed > 4ch; 8ch > 4ch
Tones	2,10	14.2	0.001	Unprocessed > 8ch, 4ch
Syllables	2,10	41.4	<0.001	Unprocessed > 8ch, 4ch; 8ch > 4ch

A. Effects of speech-processing and talker conditions

Vowel, tone, and syllable recognition performances were analyzed independently. Single-talker performance was analyzed using one-way repeated measures analyses of variance (RM ANOVAs) with speech processing as the factor; unprocessed and processed speech were treated as different levels within the speech-processing factor. Concurrent-talker performance was analyzed using two-way RM ANOVAs with speech-processing and talker conditions as factors.

Table II shows the results from one-way RM ANOVAs performed on single-talker speech performance. Vowel, tone, and syllable recognitions were all significantly affected by speech processing. For vowels, there was no significant difference between unprocessed speech and the eight-channel simulation, but performance with the eight-channel simulation was significantly better than that with the four-channel simulation. For tones, performance with unprocessed speech was significantly better than that with either of the CI simulations, but there was no significant difference between the eight- and four-channel simulations. For syllables, performance was significantly different between any two of the three processing conditions (unprocessed speech, eight- and four-channel simulations).

Table III shows the results from two-way RM ANOVAs performed on concurrent-talker speech performance. Vowel, tone, and syllable recognitions were all significantly affected by speech processing. For vowels, tones, and syllables, performance with unprocessed speech was significantly better than that with either of the CI simulations. For vowels and syllables, performance with the eight-channel simulation was

significantly better than that with the four-channel simulation. There was no main effect for talker conditions. There were significant interactions between speech-processing and talker conditions for tones and syllables. Tone and syllable recognitions were significantly better with the male-female condition for unprocessed speech, and with the male-male condition for eight-channel speech. However, the difference in mean performance between the talker conditions was quite small (<5%) within any of the speech-processing conditions.

B. Effects of vowel pairs

Figure 2 shows concurrent Chinese vowel and tone recognition scores with the CI simulations, as a function of vowel pairs in the concurrent syllables. Because there was no main effect for talker conditions and the detailed performance patterns were similar between the talker conditions, the male-male and male-female performance data were combined. Because syllable recognition performance was largely predictable from the vowel and tone scores, only vowel and tone recognition scores are shown. Also, because performance was nearly perfect with unprocessed speech, performance is shown only for the four- and eight-channel CI simulations.

Table IV shows the results from two-way RM ANOVAs performed on the data shown in Fig. 2. Concurrent-vowel recognition was significantly affected by the speech processing and vowel pairs, and there was a significant interaction between the speech processing and vowel pairs. Concurrent-tone recognition was also significantly affected by the vowel pairs; there was no significant interaction between the speech processing and vowel pairs.

Vowel response patterns were generated for the different vowel pairs used in the concurrent-syllable recognition tasks. Figure 3 shows the distribution of vowel responses for different vowel pairs. Although these vowel response patterns do not provide a totally unambiguous representation of confusions between vowel pairs, they may still provide useful information. The response patterns with unprocessed speech (data not shown) corresponded to nearly perfect recognition performance. When the concurrent syllables had the same vowel (i.e., /a/-/a/, /e/-/e/, /u/-/u/, and /i/-/i/), subjects chose

TABLE III. Results from two-way RM ANOVAs performed on concurrent-talker data, with speech-processing and talker conditions as factors. Significant effects are shown in bold. Significant differences ($p < 0.05$) from *post-hoc* Bonferroni t-tests are also shown in bold.

	Speech processing (Unprocessed, 8ch, 4ch)			Talker condition (M-M, M-F)			Speech processing × talker condition		
	dF, res	F-ratio	p-value	dF, res	F-ratio	p-value	dF, res	F-ratio	p-value
Vowels	2,10	158.7	<0.001	1,10	1.2	0.33	2,10	2.7	0.11
<i>Post-hoc</i> $p < 0.05$	Unprocessed > 8ch > 4ch								
Tones	2,10	602.4	<0.001	1,10	0.2	0.65	2,10	11.5	0.003
<i>Post-hoc</i> $p < 0.05$	Unprocessed > 8ch, 4ch						Unprocessed: M-F > M-M; 8ch: M-M > M-F		
Syllables	2,10	812.6	<0.001	1,10	0.9	0.40	2,10	19.6	<0.001
<i>Post-hoc</i> $p < 0.05$	Unprocessed > 8ch > 4ch						Unprocessed: M-F > M-M; 8ch: M-M > M-F		

FIG. 2. Mean concurrent Chinese vowel (left panel) and tone recognition scores (right panel) for NH subjects listening to the eight- or four-channel CI simulation, as a function of vowel pairs in the concurrent syllables. The error bars represent one standard deviation of the mean.

the target vowel nearly 100%. When the concurrent syllables had two different vowels (e.g., /a/-/e/, /a/-/u/, /a/-/i/, etc.), subjects' responses were evenly split between the two target vowels (~50%). There was a broad distribution of vowel responses with the CI simulations, with subjects choosing vowels that were not present in the concurrent syllables and/or favoring one of the component vowels. For example, for vowel pair /a/-/u/ with the eight-channel CI simulation, subjects most often heard /a/ (55%), seldom heard /u/ (17%), and sometimes heard /e/ (19%), which was not present. Similarly, with four channels, subjects most often heard /a/ (53%), seldom heard /u/ (15%), and sometimes heard /e/ (27%).

To analyze whether subjects were biased toward responding with the same vowel for individual syllable pairs, the percentage of responses with the same vowel within a pair was calculated. Averaged across both talker and CI simulation conditions, subjects made such responses only 30% of the time, i.e., close to the percentage of concurrent syllables actually consisting of the same vowel (25%). Therefore, there was no strong bias toward responding with the same vowel within a pair.

C. Effects of tone pairs

Figure 4 shows concurrent Chinese vowel and tone recognition scores with the CI simulations, as a function of tone pairs in the concurrent syllables. Similar to Fig. 2, the male-male and male-female performance data were combined. Compared with the vowel pairs (left panel of Fig. 2), the

different tone pairs had a smaller effect on vowel recognition (left panel of Fig. 4). Conversely, the different tone pairs had a stronger effect on tone recognition (right panel of Fig. 4) than did the different vowel pairs (right panel of Fig. 2).

Table V shows the results from two-way RM ANOVAs performed on the data shown in Fig. 4. Concurrent-vowel recognition was significantly affected by the speech processing, but not by the tone pairs; there was a significant interaction between the speech processing and tone pairs. Except for tone pair 1-1, concurrent-tone recognition was significantly better when tone pairs consisted of the same tone rather than different tones. For tone pairs consisting of the same tone, performance was significantly better for tone pairs 3-3 and 4-4 than for tone pair 1-1.

Tone response patterns were generated for the different tone pairs used in the concurrent-syllable recognition tasks. Figure 5 shows the distribution of tone responses for different tone pairs. Similar to the vowel response patterns, the tone response patterns do not provide a totally unambiguous representation of tone pair confusions, but they may still provide useful information. Again, the response patterns with unprocessed speech are not shown as subjects achieved nearly perfect recognition performance. There was a broad and/or uneven distribution of tone responses with the CI simulations. For example, for tone pair 1-4, subjects more often responded with tone 4 (56%) than with tone 1 (32%). For tone pair 2-4, subjects responded with tone 2 (33%), tone 4 (37%), and tone 1 (23%).

Different from concurrent-vowel recognition, subjects

TABLE IV. Results from two-way RM ANOVAs performed on the concurrent-talker data shown in Fig. 2, with speech processing and vowel pairs as factors. Significant effects are shown in bold. Significant differences ($p < 0.05$) from *post-hoc* Bonferroni t-tests are also shown in bold.

	Speech processing (8ch, 4ch)			Vowel pairs (a-a, a-e, a-u, a-i, e-e, e-u, e-i, u-u, u-i, i-i)			Speech processing × vowel pairs		
	dF, res	F-ratio	p-value	dF, res	F-ratio	p-value	dF, res	F-ratio	p-value
Vowels	1,45	28.2	0.003	9,45	6.7	<0.001	9,45	6.1	<0.001
<i>Post-hoc</i> $p < 0.05$		8ch > 4ch		e-u, e-i, u-u, u-i, i-i > a-u; u-u > a-a, a-i, e-e			8ch: a-a, a-e, a-i, e-i, u-u, u-i > a-u 4ch: e-u, u-u, u-i > a-u, a-i, e-e; i-i > a-u; u-u > a-a, a-e		
Tones	1,45	5.9	0.06	9,45	3.0	0.008	9,45	1.5	0.19
<i>Post-hoc</i> $p < 0.05$				i-i > e-u					

FIG. 3. Mean distribution of vowel responses (in percentage of responses) for the different vowel pairs in the concurrent syllables, with the eight-channel (left panel) or four-channel CI simulation (right panel). The error bars represent one standard deviation of the mean.

FIG. 4. Mean concurrent Chinese vowel (left panel) and tone recognition scores (right panel) for NH subjects listening to the eight- or four-channel CI simulation, as a function of tone pairs in the concurrent syllables. The error bars represent one standard deviation of the mean.

TABLE V. Results from two-way RM ANOVAs performed on the concurrent-talker data shown in Fig. 4, with speech processing and tone pairs as factors. Significant effects are shown in bold. Significant differences ($p < 0.05$) from *post-hoc* Bonferroni *t*-tests are also shown in bold.

	Speech processing (8ch, 4ch)			Tone pairs (1-1, 1-2, 1-3, 1-4, 2-2, 2-3, 2-4, 3-3, 3-4, 4-4)			Speech processing $\times$ tone pairs		
	dF, res	F-ratio	<i>p</i> -value	dF, res	F-ratio	<i>p</i> -value	dF, res	F-ratio	<i>p</i> -value
Vowels	1,45	22.6	0.005	9,45	1.2	0.31	9,45	2.4	0.02
<i>Post-hoc</i> $p < 0.05$		8ch > 4ch						8ch: 2-4, 3-4 > 4-4	
Tones	1,45	2.8	0.15	9,45	18.2	<0.001	9,45	1.5	0.16
<i>Post-hoc</i> $p < 0.05$				2-2, 3-3, 4-4 > 1-3, 1-4, 2-3, 2-4, 3-4; 3-3, 4-4 > 1-1, 1-2					

FIG. 5. Mean distribution of tone responses (in percentage of responses) for the different tone pairs in the concurrent syllables, with the eight-channel (left panel) or four-channel CI simulation (right panel). The error bars represent one standard deviation of the mean.

had a strong bias toward responding with the same tone for individual syllable pairs. The percentage of such responses averaged across both talker and CI simulation conditions was 50%, which was two times the percentage of concurrent syllables actually consisting of the same tone (25%).

IV. DISCUSSION

Concurrent Chinese syllable, vowel, and tone recognitions were nearly perfect with unprocessed speech. With unprocessed speech, concurrent-syllable and tone recognition scores were slightly ($\sim 5\%$) but significantly better for the male-female condition, showing some benefit of the larger F0 separation. For the Chinese vowel stimuli used in the present study, different vowels produced by the same male talker had different instantaneous F0's because of the different tonal patterns and variations in production. With unprocessed speech, these F0 differences (along with pitch period asynchrony and formant transitions) may have been sufficient to produce nearly perfect vowel recognition performance in the male-male condition, similar to that in the male-female condition. Also, with unprocessed speech, there was little variability in concurrent recognition performance across the different vowel and tone pairs, due to ceiling effects.

The reduced spectral resolution in the acoustic CI simulations had a more detrimental effect on concurrent-syllable recognition than on single-syllable recognition (Fig. 1). This confirms that while gross spectral and temporal representations may support good speech understanding in quiet, they do not provide sufficient acoustic cues for sound source segregation. The single-talker vowel and tone recognition scores with the CI simulations in the present study were slightly higher than previously reported with real CI listeners (Luo *et al.*, 2008), possibly because the present stimulus set was a subset of the stimuli used in Luo *et al.*, 2008 and because of inherent differences between simulated and real CI listening. The present concurrent-vowel recognition results with the 8-channel CI simulation were comparable to those reported by Qin and Oxenham (2005) with 24-channel processing, possibly due to the smaller number of vowel choices used in the present study [four, as compared to five in the Qin and Oxenham (2005) study].

In the single- and concurrent-talker conditions, when the number of frequency channels was increased from four to eight, vowel and syllable recognitions significantly improved, while tone recognition was unchanged. These findings extend those for single-syllable recognition with 1–4 frequency channels (Fu *et al.*, 1998) or up to 12 frequency channels (Xu *et al.*, 2002). In the study by Fu *et al.* (1998), the effects of spectral resolution may have been obscured by the unequal distribution of each tone type in different conditions. In the present study, doubling the number of frequency channels from four to eight may have improved spectral envelope representations, but not to the point of resolving F0 and harmonics. Thus, single- or concurrent-vowel recognition improved with the number of frequency channels while tone recognition remained unchanged. As suggested by Fu *et al.* (1998), when spectral resolution is severely limited,

either single- or concurrent-tone recognition in electric hearing strongly relies on temporal envelope cues.

When listening to the CI simulations, NH listeners' vowel recognition was much better than tone recognition for concurrent syllables. In contrast, vowel and tone recognition performances were similar for single syllables, possibly due to ceiling effects. When two Chinese syllables are presented simultaneously, interference is created in both the spectral and temporal domains. In the CI simulations, the degraded spectral envelopes (i.e., formant structures) of the two vowels were mixed together and further smeared, resulting in poor segregation and recognition of individual vowels. In each frequency channel, the temporal waveforms of the two vowels were also mixed, creating some degree of modulation detection interference (e.g., Richardson *et al.*, 1998). It is plausible that such modulation detection interference may have adversely affected tracking of amplitude envelopes and periodicity fluctuations, resulting in poor segregation and recognition of individual tones. Because concurrent-vowel recognition was so much better than concurrent-tone recognition with the CI simulations, listeners may have been more susceptible to interference in the temporal domain than in the spectral domain. These results indicate that compared with English syllable recognition, which does not require tone recognition, concurrent Chinese syllable recognition with CI may be more challenging, due to the strong interference between concurrent tones.

In general, there was little difference in performance between the male-male and male-female conditions with the CI simulations. Note that with the eight-channel simulation, concurrent-syllable and tone recognition scores were slightly but significantly better for the male-male condition. In the male-male condition, each single-vowel was paired with itself once. Such pairs of two exactly same vowels may not be informative for concurrent-syllable recognition, but they introduced minimum temporal waveform interference and may have produced better tone recognition. The small number of frequency channels in the CI simulations may have limited listeners' ability to take advantage of the larger F0 separation in the male-female condition. Previous studies have also shown that increasing the F0 separation between concurrent talkers provides no benefit for CI users' masking release from competing talkers (Stickney *et al.*, 2007), or for recognition of concurrent, synthesized vowels in CI simulations (Qin and Oxenham, 2005). Recently, Carlyon *et al.* (2007) showed that CI users could not exploit pulse asynchrony or rate differences between concurrent channels to segregate sounds, suggesting that F0 differences may help segregation only when harmonics are resolved by the peripheral auditory system. However, in a study using sequential (rather than concurrent) presentation of vowels, Gaudrain *et al.* (2008) observed F0-based auditory segregation at a much faster-than-normal presentation rate (7.5 vowels/s) in NH subjects listening to a 12-channel (rather than 8-channel) CI simulation. Such F0-based auditory segregation has been attributed to spectral envelope cues instead of temporal periodicity cues.

With the CI simulations, the variability in performance across different vowel pairs was much larger for vowel rec-

ognition than for tone recognition (Fig. 2). The performance patterns (across vowel pairs) were quite different between concurrent-vowel and tone recognitions, suggesting that the percepts may have not been strongly related. In other words, better concurrent-vowel recognition with some vowel pairs was not necessarily associated with better concurrent-tone recognition, or vice-versa. Conversely, with the CI simulations, concurrent-tone (rather than vowel) recognition greatly varied across the tone pairs (Fig. 4). Specifically, concurrent-tone recognition was much better for tone pairs consisting of the same tone (except for tone pair 1-1), while concurrent-vowel recognition was quite similar across the tone pairs. These results are different from those of [Chalikia and Bregman \(1989\)](#), who found that NH listeners' concurrent-vowel recognition was better with crossing pitch contours than with parallel pitch contours. The CI simulations did not preserve low-order, resolved harmonics, which may have made listeners unable to benefit from the different tonal patterns within concurrent syllables ([Qin and Oxenham, 2005](#); [Carlyon et al., 2007](#)). Although there are perceptual trade-offs between spectral and temporal cues (e.g., [Xu and Pfingst, 2008](#)), vowel recognition in electric hearing strongly relies on spectral envelope cues, while tone recognition depends more strongly on temporal envelope cues. The limited spectro-temporal fine structure cues available in CI speech processing may not have provided sufficiently salient temporal envelope cues to segregate spectral envelope cues (and vice-versa). Therefore, concurrent-vowel and tone recognitions with the present CI simulations were independent of each other.

The distribution of vowel responses (Fig. 3) indicates that performance was poorer for some vowel pairs than for others. With the CI simulations, subjects' perception was typically dominated by one of the component vowels, especially when the other vowel was /u/. For example, subjects more often heard /a/ and /i/ for vowel pairs /a/-/u/ and /i/-/u/, respectively; /u/ was seldom heard. The first two formant frequencies of /u/ are relatively low and closely spaced (see Table I), and may have been masked by the competing vowel /a/ or /i/. [McKeown \(1992\)](#) found that for NH subjects listening to concurrent vowels, /u/ was also perceptually dominated by the other component vowel (e.g., /a/ or /i/). [McKeown \(1992\)](#) suggested that such vowel masking may have occurred at the auditory periphery in the spectral domain, and at a more central level during cognition and attention. Another common error was subjects' perception of a vowel that was not present in the component vowel pair. For example, with the four-channel CI simulation, subjects often responded with /e/ when presented with vowel pair /a/-/i/. The perception of the vowel /e/ may have been due to the combination of the first formant of /i/ with the second formant of /a/ (see Table I). Interestingly, in both the male-male and male-female conditions, vowel pairs consisting of the same vowel (except for vowel pair /u/-/u/) were not always better recognized than those consisting of different vowels. For vowel pairs /e/-/e/ and /i/-/i/, subjects sometimes responded with /u/ in addition to the target vowel (/e/ or /i/). The noise-band carriers used in the present CI simulations may have produced the perceptual illusion of the vowel /u/.

FIG. 6. Example of temporal envelope interference between concurrent syllables. A Chinese vowel /a/ in tone 2 (left panel) is combined with another Chinese vowel /a/ in tone 4 (middle panel). The combined temporal waveforms show a flat amplitude envelope (right panel).

Because of the closely spaced F1 and F2 values, CI simulations of /u/ have most energy in only one or two adjacent frequency channels (e.g., the second lowest channel of the four-channel CI simulation), and thus may sound similar to narrow-band noise.

Different from vowel recognition, concurrent-tone recognition with the CI simulations was significantly better when the two syllables had the same tone (except for tone pair 1-1), due to subjects' tendency to respond with the same tone for concurrent syllables within a pair. Different pitch contours between concurrent syllables did not aid in sound source segregation, but rather produced poorer tone recognition. The distribution of tone responses (Fig. 5) revealed that, similar to vowel recognition, concurrent-tone recognition with the CI simulations also exhibited two types of errors, both of which may be explained by temporal envelope interference. Note that for Chinese single-vowel syllables, the amplitude envelope and F0 contour have similar shapes, which is an important temporal cue for Chinese tone recognition with limited spectral resolution ([Luo and Fu, 2004](#)). When tone 1 (flat) was presented simultaneously with another tone (e.g., tone 4, falling), the combined temporal envelope largely followed tone 4 and may have resulted in the perceptual dominance of tone 4. When tone 2 (rising) was presented simultaneously with tone 4 (falling), the combined temporal waveforms may have resulted in a flat amplitude envelope, indicating a flat tone (i.e., tone 1). Not surprisingly, listeners often responded with tone 1 for tone pair 2-4. An example of temporal envelope interference is shown in Fig. 6.

The present study used four and eight frequency channels to simulate the typical spectral resolution available to CI users. The simulation results with NH listeners suggest that CI users might not utilize F0 differences across talkers and/or tones to recognize concurrent syllables. However, the strength of temporal envelope pitch may have been reduced in the simulations as compared to the real implant case, due to the noise-band carrier (which might reduce temporal envelope saliency) and the employment of synthesis band-pass filters ([Laneau et al., 2006](#)). Also, different from the NH subjects, CI performance can be greatly affected by patient-related factors. While simulation studies with NH listeners

presumably reduce these patient-related factors, they are an important consideration for future studies with real CI users.

Complex perceptual tasks such as speech recognition in the presence of competing speech require high degrees of spectral resolution. Even with 24 channels, concurrent English vowel recognition performance is much poorer than that with unprocessed speech (Qin and Oxenham, 2005). Although implanted with 16–22 electrodes, CI users can access only approximately eight channels, due to electrode interactions. Advanced speech-processing strategies that restore spectro-temporal fine structure cues to CI users may enhance their sound source segregation. Binaural cues, via bilateral CIs or a hearing aid in the non-implanted ear, may also improve CI users' sound source segregation. Previous studies with NH listeners have shown improved concurrent-vowel recognition when the two vowels were presented to different ears (i.e., dichotic hearing instead of diotic hearing; Zwicker, 1984) or when adding interaural time differences (Shackleton and Meddis, 1992). Recently, Long *et al.* (2006) found that bilateral CI users' signal detection in noise was significantly better when the 500-Hz temporal envelopes delivered to a single electrode in each ear were out of phase rather than in phase. Thus, it is of interest to investigate concurrent-vowel and tone recognitions in bilateral CI users and listeners with bilaterally combined electric and acoustic stimulation.

V. CONCLUSIONS

The present study measured NH listeners' recognition of concurrent Chinese syllables, vowels, and tones produced by one male and one female talker or by the same male talker. Performance was measured with original, unprocessed speech, and with speech processed by four- or eight-channel acoustic CI simulation. Concurrent-syllable, vowel, and tone recognitions were significantly poorer with the CI simulations than with original, unprocessed speech. Syllable and tone recognitions were significantly better with the male-female condition for unprocessed speech, and with the male-male condition for eight-channel speech. With the CI simulations, concurrent-vowel and syllable recognitions were significantly different across different vowel pairs, while tone recognition remained largely unchanged. In contrast, concurrent-vowel recognition was not significantly affected by the different tone pairs. Tone and syllable recognitions were significantly better when the two syllables had the same tone. With the CI simulations, concurrent-vowel and tone recognitions were independent of each other. The weak pitch coding in the CI simulations may preclude enhanced concurrent recognition performance derived from large F0 separations between the talkers and/or different pitch contours between the syllables. Furthermore, with the CI simulations, concurrent Chinese syllable recognition may be more challenging than English syllable recognition, due to the strong interference between tonal envelope cues.

ACKNOWLEDGMENTS

We are grateful to all subjects for their participation in these experiments. We thank John J. Galvin III for editorial

assistance. We would also like to thank three anonymous reviewers for their constructive comments on an earlier version of this paper. Research was supported in part by NIH (Grant Nos. R03-DC-008192 and R01-DC-004993).

- Assmann, P. F. (1995). "The role of formant transitions in the perception of concurrent vowels," *J. Acoust. Soc. Am.* **97**, 575–584.
- Assmann, P. F., and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *J. Acoust. Soc. Am.* **88**, 680–697.
- Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge, MA).
- Carlyon, R. P., Long, C. J., Deeks, J. M., and McKay, C. M. (2007). "Concurrent sound segregation in electric and acoustic hearing," *J. Assoc. Res. Otolaryngol.* **8**, 119–133.
- Chalikia, M. H., and Bregman, A. S. (1989). "The perceptual segregation of simultaneous auditory signals: Pulse train segregation and vowel segregation," *Percept. Psychophys.* **46**, 487–496.
- Culling, J. F., and Summerfield, Q. (1995). "The role of frequency modulation in the perceptual segregation of concurrent vowels," *J. Acoust. Soc. Am.* **98**, 837–846.
- Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X.-S. (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., Hsu, C.-J., and Horng, M.-J. (2004). "Effects of speech processing strategy on Chinese tone recognition by nucleus-24 cochlear implant users," *Ear Hear.* **25**, 501–508.
- 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.
- Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (2008). "Steaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation," *J. Acoust. Soc. Am.* **124**, 3076–3087.
- Green, T., Faulkner, A., and Rosen, S. (2004). "Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants," *J. Acoust. Soc. Am.* **116**, 2298–2310.
- Greenwood, D. D. (1990). "A cochlear frequency-position function for several species-29 years later," *J. Acoust. Soc. Am.* **87**, 2592–2605.
- Laneau, J., Moonen, M., and 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.
- Long, C. J., Carlyon, R. P., Litovsky, R. Y., and Downs, D. H. (2006). "Binaural unmasking with bilateral cochlear implants," *J. Assoc. Res. Otolaryngol.* **7**, 352–360.
- Luo, X., and Fu, Q.-J. (2004). "Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants," *J. Acoust. Soc. Am.* **116**, 3659–3667.
- Luo, X., Fu, Q.-J., Wei, C.-G., and Cao, K.-L. (2008). "Speech recognition and temporal amplitude modulation processing by Mandarin-speaking cochlear implant users," *Ear Hear.* **29**, 957–970.
- McKeown, J. D. (1992). "Perception of concurrent vowels: The effect of varying their relative level," *Speech Commun.* **11**, 1–13.
- Meddis, R., and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.* **91**, 233–245.
- 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.
- Richardson, L. M., Busby, P. A., and Clark, G. M. (1998). "Modulation detection interference in cochlear implant subjects," *J. Acoust. Soc. Am.* **104**, 442–452.
- Scheffers, M. T. M. (1983). "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. thesis, Groningen University, The Netherlands.
- Shackleton, T. M., and Meddis, R. (1992). "The role of interaural time difference and fundamental frequency difference in the identification of concurrent vowel pairs," *J. Acoust. Soc. Am.* **91**, 3579–3581.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Stickney, G. S., Assmann, P. F., Chang, J., and Zeng, F.-G. (2007). "Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences," *J. Acoust. Soc. Am.* **122**, 1069–1078.

- Summerfield, Q., and Assmann, P. F. (1991). "Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony," *J. Acoust. Soc. Am.* **89**, 1364–1377.
- Wang, R.-H. (1993). "The standard Chinese database," University of Science and Technology of China, internal materials.
- Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (1991). "Better speech recognition with cochlear implants," *Nature (London)* **352**, 236–238.
- Xu, L., and Pfingst, B. E. (2008). "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," *Hear. Res.* **242**, 132–140.
- 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.
- Zwicker, U. T. (1984). "Auditory recognition of diotic and dichotic vowel pairs," *Speech Commun.* **3**, 265–277.