Perception & Psychophysics
2003, 65 (7), 1029-1044

On the (non)categorical perception of lexical tones

ALEXANDER L. FRANCIS
Purdue University, West Lafayette, Indiana

and

VALTER CIOCCA and BRENDA KEI CHIT NG
University of Hong Kong, Hong Kong

Identification and discrimination of lexical tones in Cantonese were compared in the context of a traditional categorical perception paradigm. Three lexical tone continua were used: one ranging from low level to high level, one from high rising to high level, and one from low falling to high rising. Identification data showed steep slopes at category boundaries, suggesting that lexical tones are perceived categorically. In contrast, discrimination curves generally showed much weaker evidence for categorical perception. Subsequent investigation showed that the presence of a tonal context played a strong role in the identification of target tones and less of a role in discrimination. The results are consistent with the hypothesis that tonal category boundaries are determined by a combination of regions of natural auditory sensitivity and the influence of linguistic experience.

In tone languages, differences in fundamental frequency (f_0) patterns, perceived as pitch height and/or contour differences, can be sufficient for listeners to distinguish lexical items (see, e.g., Gandour, 1978).[1] For example, in Cantonese, a speaker can give two syllables with identical segmental qualities different lexical meanings simply by producing them with different f_0 patterns. The segmental string /ji/ means *doctor* when produced with a high-level tone (a roughly level f_0 contour approximately in the upper third of the talker's f_0 range); when produced with a low-level tone (a roughly level f_0 contour approximately in the bottom third of the talker's f_0 range), /ji/ means *two*.[2] Linguistically, tonal phonetic categories are generally considered to be psychologically equivalent to segmental phonetic categories (i.e., vowels and consonants; see, e.g., Ladd, 1996 pp. 1–2).

One well-known difference between vowels and consonants is that consonantal feature contrasts (e.g., voicing or place of articulation) are typically found to be perceived in a mostly categorical manner, especially for those features related to the perception of stop constonants (Liberman, Harris, Eimas, Lisker, & Bastian, 1961; Liberman, Harris, Hoffman, & Griffith, 1957; Liberman, Harris, Kinney, & Lane, 1961), whereas contrasts involving vowel features are not (Abramson, 1976; Fry, Abramson, Eimas, & Liberman, 1962; Stevens, Liberman, Studdert-Kennedy, & Öhman, 1969). The principle of categorical perception was developed, in part, to explain how listeners cope with the many-to-one mapping between acoustic patterns and phonological categories (the problem of *lack of invariance*). When producing multiple instances of a given phoneme, a talker is almost certain to produce a range of acoustic patterns. Despite this variability in the acoustic signal, listeners tend to hear similar acoustic patterns as belonging to the same category. When stimuli are perceived categorically, equivalent acoustic differences between two tokens are treated differently, depending on whether the two tokens are heard as members of the same category or as members of different categories. Two members of one category are less discriminable than are two tokens from two different categories with an equivalent acoustic difference between them. The idea is that, through experience with a given language, listeners learn the location of specific category boundaries along various acoustic continua. By increasing discrimination accuracy across these boundaries and/or reducing it within boundaries, listeners improve their ability to hear two acoustically similar but not identical members of one category as *the same* and, conversely, improve their ability to hear two acoustically similar members of distinct categories as *different.*

Studies of categorical perception often compare identification proportions with accuracy on a discrimination

Some of the data derive from part of a dissertation submitted by the third author in partial fulfillment of the requirements for the B.Sc. degree in speech and hearing sciences at the University of Hong Kong. Some of these results were first presented at the 139th meeting of the Acoustical Society of America, Atlanta, May 30th, 2000. Some of the experiments reported here were funded by Grants HKU 7193/00H to V.C. and HKU 10300/302 to V.C., A.L.F., and Lian Ma from the University of Hong Kong Research Grants Committee. This work was carried out while the first author was a postdoctoral fellow in the Department of Speech and Hearing Sciences at the University of Hong Kong. We thank José Benki, Bert Schouten, and one anonymous reviewer for their helpful comments on previous versions of this paper. Correspondence concerning this article should be addressed to A. L. Francis, Audiology and Speech Sciences, Purdue University, Heavilon Hall, West Lafayette, IN 47907 (e-mail: francisa@purdue.edu).

task, using stimuli that range in perceptually equal steps along an acoustic continuum. A stimulus continuum is typically considered to be perceived categorically if listeners' responses fit with two criteria. First, identification proportions should predict discrimination accuracy (Liberman et al., 1957). Second, peaks in discrimination sensitivity should correspond to the location of category boundaries as determined by identification, operationally defined as the point along a continuum at which the identification function for one category crosses the identification function for another (see, e.g., Repp, Healy, & Crowder, 1979).

With respect to the categorical perception of lexical tone contrasts, it has been reported that musically trained listeners exhibit categorical perception of nonspeech frequency continua (Howard, Rosen, & Broad, 1992; Zatorre & Halpern, 1979), suggesting that experience with the distinguishing of categories defined along an f_0 continuum (in this case, musical notes) may induce categorical perception. Furthermore, some researchers have argued that speakers of tone languages exhibit a decreased sensitivity to absolute frequency differences that correspond to within-category differences in tone (Stagray & Downs, 1993; Tanner & Rivette, 1964; although see Burns & Sampat, 1980). Thus, it is possible that experience in making category decisions on the basis of perceived pitch (whether phonological or musical) may lead to the categorical perception of f_0 continua.

Previous research has presented conflicting results with respect to the categorical nature of tone perception. Abramson (1979) investigated Thai listeners' perception of a continuum of three level tones. In the experiments reported by Abramson (1979), Thai listeners exhibited uniformly good discrimination across the continuum, suggesting that Thai level tone contrasts are not perceived categorically. On the other hand, Wang (1976) found evidence of a peak in discrimination accuracy corresponding to a tone category boundary in a stimulus continuum ranging from a Mandarin Chinese Tone 2 (high rising) to a Tone 1 (high level). These results were interpreted as suggesting that Mandarin Chinese listeners perceive tonal contrasts in a categorical manner. However, it is difficult to make clear comparisons between these studies, due to differences in the stimuli and languages under investigation. In order to resolve the discrepancy between the two sets of results, it would be useful to compare the categorical perception of level and contour tone contrasts within a single language. Cantonese, a Chinese language spoken in southern China, including Hong Kong and Macao, and in many overseas Chinese communities, presents an excellent case for such a study. Like Thai, Cantonese has three level tones: a high-level, a mid-level, and a low-level tone. Like Mandarin, Cantonese also contrasts level tones with contour tones.

Six experiments were carried out in the present study. The first three were designed to investigate the perception of level and contour tone continua within a single language. On the basis of the results of these experiments, three further experiments were conducted to investigate the effects of external context on the categorical perception of lexical tones.

EXPERIMENT 1

This experiment reprised Abramson's (1979) original experiment on Thai level tones, here using Cantonese level tones.

Method

Subjects. Thirty-one native Cantonese speakers, 26 women and 5 men, with no reported history of speaking or hearing disability, participated in this experiment. Nineteen were undergraduate speech pathology students in the Department of Speech and Hearing Sciences at the University of Hong Kong, and 12 were students and employees from other departments. Sixteen (14 women and 2 men) of these subjects were assigned to the identification task, and 15 (12 women and 3 men) were assigned to the discrimination task. Note that this ratio of gender is representative of the student body in the humanities, education, and social science faculties at the University of Hong Kong, from which the subjects for all of the following experiments were sought. There is no reason to expect a gender-based difference in performance on any of these basic listening tasks.

Stimuli. The stimuli for this and all the subsequent experiments were generated in the following manner. A male native speaker of Cantonese, 22 years old, with no speaking or hearing disability was recorded producing 10 samples of a complete sentence (adapted from a carrier sentence used by Zee, 1988), /ŋɔ25 wui25 tʰɔk22 ji pɛi25 lɛi24 tʰɛŋ55 / (*I will read* [ji] *for you to hear*), with each of the six contrastive Cantonese tones on the target syllable [ji] (60 utterances). Each such syllable is a distinct morpheme (and character) in Cantonese: /ji55/ (*doctor*), /ji25/ (*chair*), /ji33/ (*spaghetti*), /ji21/ (*child*), /ji23/ (*ear*), and /ji22/ (*two*). Note that we used a system based on Chao's tone numbers (Chao, 1947) to transcribe tones. This system reflects the relative pitch of the syllable at onset and offset within a 5-point scale from the *bottom* (1) to the *top* (5) of the talker's normal frequency range. However, we diverged from traditional (impressionistic) characterizations of some tones in order to reflect more accurately the actual fundamental frequency patterns of the talkers we have observed (and the data presented by Bauer and Benedict, 1997). Thus, 55 = high level, 25 = high rising, 33 = mid level, 21 = low falling, 23 = low rising, and 22 = low level.

Measurements were made of the maximum and minimum f_0 excursion in each syllable of each utterance, using GW Instruments' Sound Scope 16 running on an Apple Power Macintosh G3. Those six /ji/ syllables (one for each of the six tones of Cantonese) that represented the most extreme difference, in the appropriate direction, from the mean values for /ji/ were also identified. For example, the "best" /ji55/ (high level) syllable was the syllable that had the greatest positive difference in f_0 from the mean /ji/ values.

To generate the synthetic /ji/ syllables, a native speaker (the third author of this article) selected the /ji/ syllable that sounded segmentally the most natural from among the six /ji/ syllables identified as "best" according to their f_0 contours (the syllable with high-level tone). This most natural sounding of the "best" /ji/ syllables was subsequently used as a model for determining the formant frequencies, syllable duration, and amplitude envelope of the synthetic syllable. For synthesizing the six tones, a single /ji/ syllable was first synthesized with these formant frequencies, duration, and amplitude and with f_0 modeled on the best /ji33/ syllable. Synthesis parameters were then varied by trial and error, in consultation with native listeners, until optimum characteristics of a /ji/ syllable with a mid-level tone were achieved on the basis of perceptual judgment. The duration of this syllable had to be increased, even over the average duration of all measured /ji/ syllables, in order to make it acceptable to native speakers both in isolation and in sentence context. A duration of 300 msec was found to be acceptable in both contexts, although subsequently, some listeners reported that syllables presented in isolation had a "clipped" or "abrupt ending" quality. This syllable was then duplicated six times, and each resulting syllable

was given a different f_0 contour corresponding to a linearization of the measured f_0 values for the "best" natural syllables.

After the synthetic stimuli were played to a native speaker of Cantonese trained as a speech therapist, some adjustments were made to various parameters to make the synthesized speech sound more realistic. These parameters included those related to the perception of breathiness and nasality. With the fundamental frequency held completely level, the values for the level tones were the following: high level, 140 Hz; mid level, 115 Hz; and low level, 100 Hz. For constructing the level tone continuum, the extreme values (140 and 100 Hz) were taken as the endpoints. These frequency values were converted to mel units (Taylor, Caley, Black, & King, 1999), and the distance between them, in mels, was subdivided into nine equal steps. The resulting intermediate mel values were then reconverted to hertz to provide a 10-token continuum from a low-level tone to a high-level tone in perceptually equal steps (6.1 mel each, about 4.4– 4.5 Hz each). Exact frequency values are given in Table 1; each of these values was used as the fundamental frequency of a single stimulus syllable. These 10 stimuli were presented to listeners in isolation individually in the identification task.

The stimuli for the discrimination task consisted of all pairwise combinations of individual syllables separated by zero or one token along the continuum, with a 250-msec interstimulus interval (ISI). There was a total of 28 such pairs, including 10 identical pairs (1–1, 2–2, 3–3, etc.) and 18 adjacent pairs (1–2, 2–1, 2–3, 3–2, etc.).

Procedure. Two groups of listeners participated in this experiment, in two different tasks. In the identification task, the listeners heard a single syllable and identified it as /ji55/ (*doctor*), /ji33/ (*spaghetti*), or /ji22/ (*two*) by clicking on a button labeled with the appropriate Chinese character. The buttons were presented in the order above, from left to right, and the order did not vary. There were 11 blocks of 10 trials each (a single presentation of each of the 10 stimuli per block). The 1st block was treated as familiarization and was not scored, although the listeners were not aware of this at the time of testing. Order of stimulus presentation within blocks was random. Responses were collected automatically and were scored according to the lexical tone of the chosen character.

The discrimination task followed a similar format. There were 11 blocks, each with 28 trials, and the subjects were not aware that the 1st block was not scored. Each trial began with the presentation of a visual warning symbol on the computer screen. Subsequently, the listeners heard a warning tone (an amplitude-modulated complex tone with both fundamental frequency and harmonic structure significantly different from speech), followed after 500 msec by the presentation of a single pair of syllables separated by a 250-msec ISI. Following the presentation of a stimulus pair, the listeners were presented with two buttons arranged horizontally on the screen, labeled *same* and *different*. The response buttons were always presented in this order. The subjects were instructed to click on one of the buttons to indicate whether the syllables they heard were the same or different. Selection was followed by a 2-sec pause, and then the next trial began. The order of stimulus presentation within blocks was random. Responses were collected automatically and were scored according to whether they were correct or not. The subjects received no feedback on their responses in either the identification or the discrimination task.

Results

Figure 1 shows the identification group listeners' average response percentages on the identification task, superimposed on a plot of the discrimination group listeners' average percentage of correct discriminations for each adjacent pair. The proportion of correct discriminations for each pair was calculated as the average of the probability of a *different* response to different pairs, P(different | different), and the probability of a *same* response to same pairs, P(same | same). For example, the proportion correct for the 1–2 pair was the average of the proportion of *different* responses to the 1–2 and 2–1 pairs and the proportion of *same* responses to the 1–1 and 2–2 pairs. Note that the rate of false alarms, P(different | same), was low, averaging just 12% across all tokens and all listeners.

Figure 1 shows that the listeners did exhibit a crossover in their identification function, corresponding roughly to the expected location of boundaries between the three tone categories (at the 3–4 and 7–8 pairs). However, these pairs do not correspond to obvious peaks in sensitivity in the discrimination curve. A one-way repeated measures analysis of variance (ANOVA) of the arcsine-transformed proportions of correct discriminations showed no effect of pair, and trend analysis showed no significant linear trend for these means. Indeed, there was no significant difference between any two of the pairs across the continuum, according to post hoc Tukey HSD tests.

In order to compare identification with discrimination performance, we calculated the proportion of correct discriminations for each adjacent pair, as predicted by the results of the identification task, using Equation 1. This equation was taken from Liberman et al. (1957), who used it to predict the results of an ABX discrimination task. However, Pollock and Pisoni (1971) showed that the same equation can also be used to predict performance in a 2IAX (*same/different*) task:

$$p(disc_{ij}) = \frac{1}{2} + \frac{1}{4}\{[p_{LL}(i) - p_{LL}(j)]^2 + [p_{ML}(i) - p_{ML}(j)]^2 + [p_{HL}(i) - p_{HL}(j)]^2\} \quad (1)$$

where $p(disc_{ij})$ is the predicted probability of discriminating tokens i and j (e.g., 3 and 4, in pair 3–4), $p_{LL}(i)$ is the measured proportion of low-level responses to token i, and so forth (Liberman et al., 1957). Predicted and actual discrimination proportions were then transformed using the arcsine transformation and submitted to a two-way repeated measures ANOVA. The results showed that discrimination proportions predicted by identification scores ($\mu = .54$) were significantly lower than the actually measured proportions of correct discrimination [$\mu = .72$; $F(1,29) = 41.02$, $p < .001$]. There was also a significant effect of pair [$F(8,232) = 3.01$, $p = .003$] and a significant interaction between pair and group [$F(8,232) = 2.73$, $p = .007$]. Post hoc (Tukey HSD) analysis revealed that every point on the actual proportion correct discrimination curve was significantly greater than every point along the proportion correct discrimination curve predicted from the identification experiment ($p < .01$ in all cases).

Discussion

The results of the identification task showed sufficiently clear crossovers at the 3–4 and 7–8 pairs to indicate the presence of category boundaries. No peaks in discrimination accuracy were observed at these locations or any

other point along the discrimination continuum, and identification responses underpredicted discriminability across the continuum, suggesting that level tones in Cantonese, like those of Thai (Abramson, 1979), are not perceived categorically. In addition, although it cannot be seen from the figures presented here, it must be noted that Cantonese listeners were more sensitive to frequency differences between two tokens when the second syllable was *higher* in frequency than the first (low–high order) and were less sensitive when the second syllable was *lower* in frequency than the first (high–low order). In the low–high order, the listeners in the present experiment had a hit rate of 79.1%, whereas in the high–low presentation order, the listeners had a hit rate of only 35.5%. Note that these values are hit rate, not proportion correct; therefore, the average of these two numbers is somewhat lower than the average percentage correct given above, due to listeners' greater accuracy on the *same* pairs (correct rejections). This contrast effect, or asymmetry, which may be related to the observation that Cantonese speakers show considerable declination in frequency over the course of an utterance (the so-called *downstep*; cf. Vance, 1976) is explored in greater detail in a related article (Francis & Ciocca, 2003).

In summary, the results presented here for the perception of Cantonese level tones are qualitatively similar to those presented by Abramson (1979) for Thai level tones. The listeners showed evidence of the presence of category boundaries in an identification task but no corresponding peaks in discrimination accuracy. In addition, discrimination accuracy was better than was predicted by the identification task. These results do not match either of the two criteria commonly taken to indicate categorical perception. In the present experiment, identification performance significantly underpredicted discrimination performance (in particular, within-category discrimination was much higher than predicted), and there were no peaks in observed discrimination accuracy corresponding to the location of category boundaries in the identification function.

EXPERIMENT 2

Previous research (Wang, 1976) suggested that a continuum modeled on a Mandarin rising to high-level tone continuum was perceived categorically. Mandarin is usually described as having four tones (high level, rising, falling–rising, and falling; Chao, 1948), whereas Cantonese is typically considered to have three contour tones in addition to the three level tones already mentioned: low falling, low rising, and high rising (Matthews & Yip, 1994). Measurement of the natural utterances on which the present stimuli are based, as well as fundamental frequency measures by Bauer and Benedict (1997) and graphs displayed by Gandour (1983), suggested that the relationship between the Cantonese high-level and high-rising tones is qualitatively similar to that between the Mandarin high-level and rising tones, as used by Wang (1976). Thus, the present experiment was designed to investigate the categorical perception of contour tones, using a Cantonese high-rising to high-level tone continuum.

Method

Subjects. Twenty-seven native speakers of Cantonese (18 women and 9 men) participated in this experiment. All were undergraduate or graduate speech pathology students in the Department of Speech and Hearing Sciences at the University of Hong Kong, and none had participated in Experiment 1. The subjects were assigned to either the identification task (12 subjects, 6 men, and 6 women) or the discrimination task (15 subjects, 3 men and 12 women).

Stimuli. Stimulus generation was identical to that in the previous experiment, with the exception of the fundamental frequency values used to generate the test syllables. The tone categories used in this experiment were chosen to be similar to those used by Wang (1976). However, some differences between the two sets of stimuli must be noted. In particular, the stimuli used by Wang (1976) were 500 msec long and remained level in frequency over the first 100 msec before rising linearly to the end of the syllable. Recordings of Cantonese syllables with contour tones showed a similar period of level or, more often, slightly falling f_0, occurring over the first quarter to the first third of the syllable (cf. Bauer & Benedict, 1997; this observation was also made in the natural speech samples on which the present stimuli were based). In order to more closely emulate the natural f_0 contour, a third point was measured. The point at which the slope changed, which we termed the *inflection point* (cf. Moore & Jongman's, 1997, *turning point*), was estimated visually from an f_0 plot, and both the time (in terms of proportion of the syllable) and f_0 (in hertz) at that point were measured. These measured values for each syllable (onset, inflection, and offset) were then used as the starting points for synthesizing three new syllables (high rising, low rising, and low falling), using the base 300-msec syllable generated for the previous experiment (as compared with 500 msec for the stimuli used by Wang, 1976). These syllables were played, in the same synthesized sentential context, to three native speakers trained as speech therapists. On the basis of their judgments, the onset f_0, the offset f_0, and the time and frequency of the inflection point of the three syllables were adjusted (within the constraint that [1] frequency must change linearly between the onset and the inflection point and between the inflection and the offset and [2] all three syllables should be identical in the onset f_0 and in the time and f_0 of the inflection point) until all three judges determined each to be a good member of its intended category. Note that all three contour tones (high rising, low rising, and low falling) were synthesized together as a set, but only the high-rising tone stimulus was used in the present experiment.

The frequency contour of the final three contour syllables began at 102 Hz and dropped at a uniform rate to 98 Hz over the first 75 msec (first 25% of the syllable, as compared with 20% for the stimuli used by Wang, 1976). From this inflection point, the fundamental frequency of the syllable with a high-rising tone rose linearly to 148 Hz. Recall that, in Experiment 1, the best high-level stimulus had a level f_0 contour at 140 Hz. To construct a continuum from high rising to high level, it was necessary to determine a value for f_0 at the endpoint of the rising tone that could also function as the constant frequency of a natural-sounding high-level tone. Informal tests with native speakers suggested that 140 Hz was the highest acceptable level frequency contour in the existing sentence context. Thus, the f_0 values for the final high-rising tone were fixed at 102 Hz (onset), 98 Hz (inflection), and 140 Hz (offset), whereas those for the high-level tone were all fixed at 140 Hz. These values were converted to mel units (Taylor et al., 1999), and the distance between them, in mel, was subdivided equally (nine steps). The resulting intermediate mel values were then reconverted to hertz to provide a 10-token continuum from a high-rising tone to a high-level tone in perceptually equal steps (5.8 and 6.4 mel for the onset and the inflection continua, about 4.2 and 4.7 Hz, respectively). Exact frequency and mel values are given in Table 2.

The stimuli for the discrimination task consisted of all pairwise combinations of syllables separated by zero or two tokens along the continuum, with a 250-msec ISI. A two-step discrimination task was used instead of a one-step difference (as in the previous experiment),

Table 1
Fundamental Frequency Values for Stimuli in Experiment 1

Stimulus	f_0 Hertz	f_0 Mel	Tone Class
1	100.0	150.50	low level
2	104.4	156.61	
3	108.7	162.72	
4	113.1	168.83	mid level
5	117.5	174.94	mid level
6	122.0	181.05	
7	126.5	187.16	
8	130.9	193.27	
9	135.5	199.38	
10	140.0	205.49	high level

because the results from an initial one-step task using 6 listeners indicated that the listeners were almost incapable of hearing any difference between tokens separated by only one step along this continuum. Note that Wang (1976) also used two-step comparisons, resulting in a total difference in onset f_0 of 6 Hz between items in a pair in his experiments, as compared with between 8 and 9 Hz in the present experiment. With two-step pairs, there was a total of 26 pairs, including identical pairs (1–1, 2–2, 3–3, etc.), as well as two-step pairs (1–3, 3–1, 2–4, 4–2, etc.).

Procedure. As in the previous experiment, the identification task consisted of 11 blocks of 10 trials each. The 1st block was treated as familiarization and was not scored in the final analysis, although the subjects were not aware of this at the time of testing. Each identification trial began with the presentation of a visual warning symbol on the computer screen. Subsequently, the listeners heard a single presentation of a stimulus syllable. Following presentation of the stimulus, the listeners were presented with three buttons arranged horizontally on the screen. Each button was labeled with a Chinese character representing one of the three syllables /ji55/ (*doctor*), /ji25/ (*chair*), and /ji23/ (*ear*), from left to right. Although it was expected that the listeners would choose primarily between /ji55/ (*doctor*) and /ji25/ (*chair*), three choices were provided, for two reasons. First, it is possible that Cantonese listeners use the *slope* (rate of change), rather than the endpoint, of the fundamental frequency contour to differentiate the two rising tones. In that case, although the ending f_0 of the stimuli used here was most similar to that of a high-level or high-rising tone, it is possible that tokens toward the middle of the continuum, with a shallow but still rising f_0 contour, might be heard as low-rising tones. Second, providing three choices made this experiment equivalent to the previous one in terms of functional load. Although there was no a priori reason to expect that the cognitive load required by choosing among three responses would be greater than that required by choosing between only two, controlling for this variable was easily accomplished by including this third choice. Response buttons were always presented in the same order. The listeners were instructed to click on the button that corresponded to the word that they heard. After selection, followed by a 500-msec pause, the next trial began. Order of stimulus presentation within blocks was randomized. Responses were collected automatically and were scored according to the tone of the character that was clicked (high level, high rising, or low rising). The discrimination task followed a format similar to that in Experiment 1, with the exceptions that there was no auditory warning prior to stimulus presentation and, because of the two-step presentation, each of the 11 blocks had only 26 trials.

Results

Figure 2 shows the identification task listeners' average response proportions and the discrimination group's proportion correct on the discrimination task. From the identification curves, it can be seen that, at the left of the continuum, the listeners heard mostly low-rising (/ji23/, *ear*) and high-rising (/ji25/, *chair*) tones, whereas in the upper half of the continuum, the listeners heard mostly high-level (/ji55/, *doctor*) tones.

The discrimination curve in Figure 3 shows a gentle peak at the 5–7 pair, and a one-way ANOVA of the arcsine-transformed discrimination proportions correct confirmed this [$F(7,98) = 6.96$, $p < .001$]. There was a significant

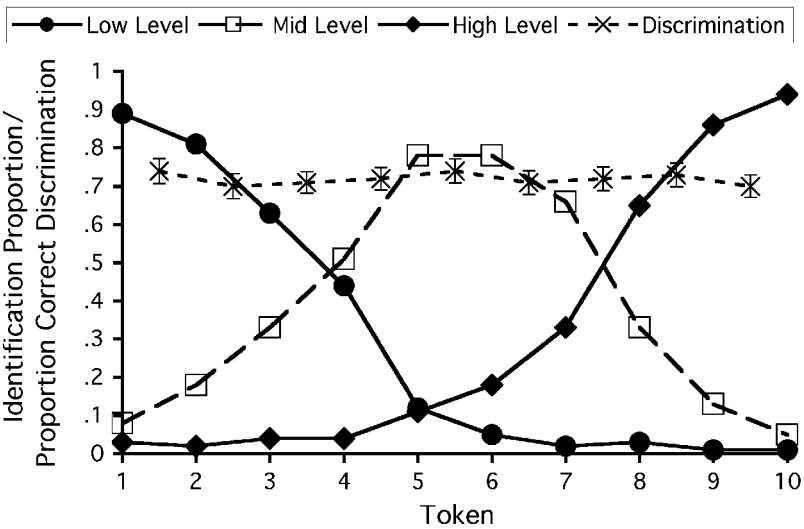

Figure 1. Identification and discrimination functions of a synthesized 10-token Cantonese level tone continuum ranging in identification from low-level (/ji22/) through mid-level (/ji33/) to high-level (/ji55/) tone categories. Error bars indicate ±1 standard error.

Table 2
Fundamental Frequency Values for Stimuli in Experiment 2

	Onset		Inflection		
Stimulus	Hertz	Mel	Hertz	Mel	Tone Class
1	102.0	153.31	98.0	147.67	high rising
2	106.1	159.10	102.6	154.09	
3	110.3	164.89	107.1	160.51	
4	114.5	170.68	111.8	166.93	
5	118.7	176.47	116.4	173.35	
6	122.9	182.26	121.1	179.77	
7	127.1	188.05	125.7	186.19	
8	131.4	193.84	130.5	192.61	
9	135.7	199.63	135.2	199.03	
10	140.0	205.42	140.0	205.45	high level

quadratic trend for these means [$F(1,14) = 12.60$, $p = .003$], supporting the conclusion that there was a central peak, albeit only a slight one. Post hoc (Tukey HSD) analysis showed that the listeners' proportion of correct responses for the 5–7 pair was significantly greater than that for all the pairs ($p < .05$) except the adjacent pairs 4–6, 6–8, and 8–9. Again, the rate of *different* response to same pairs was low, averaging just 8% across all tokens and listeners.

As in the previous experiment, predicted proportions of correct discriminations, based on the results for the identification group, were calculated using Equation 1 (above) and were compared with the actual proportions of correct discriminations for the discrimination group after both sets of proportions were transformed using the arcsine transformation. A two-way mixed factorial ANOVA was calculated with one repeated measure (pair) and one between-groups measure (identification vs. discrimination). The results showed a significant effect of pair [$F(7,175) = 15.35$, $p < .001$] but no significant differences between identification ($\mu = .58$) and discrimination ($\mu = .58$) and no interaction effect. Post hoc (Tukey HSD) analysis revealed that none of the measured discrimination proportions from the discrimination group were significantly different from those predicted on the basis of the identification responses for any of the pairs along the continuum.

Discussion

Overall, the mean proportion of correct discriminations in this experiment (.58) was lower than that in the previous experiment (.72), although the stimulus pairs in this experiment differed by nearly 9 Hz at their onset, as compared with less than 5 Hz in the previous experiment. This may be because the present stimuli all ended at the same frequency, meaning that the average frequency across the entire syllable was much closer than the differences in onset frequencies might suggest. Also, contour tones may be more difficult to discriminate than level tones, because listeners have been found to give more weight to the endpoint of a stimulus with changing frequency than they do to the starting frequency (Nábělek, Nábělek, & Hirsh, 1970). Despite this, discrimination sensitivity was found to peak across a pair of stimuli that span an apparent category boundary, as determined by identification, and actual proportion of correct discriminations was similar to that predicted by the results of the identification task. This pattern of results is similar to the findings of Wang (1976) and is consistent with the hypothesis that a rising-to-level f_0 continuum is perceived categorically.

EXPERIMENT 3

The results of Experiment 2 provided tentative support for the hypothesis that contour tones are perceived categorically. However, although the apparent peak in discrimination accuracy observed in Experiment 2 could be the result of the presence of a linguistic category boundary, it could also be the result of a naturally high degree of

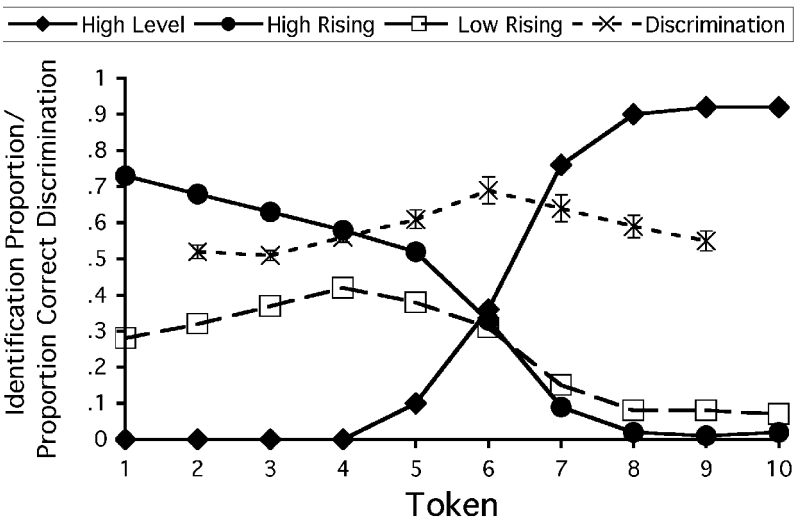

Figure 2. Identification and discrimination functions of a synthesized 10-token Cantonese rising tone continuum modeled on the Mandarin tone continuum used by Wang (1976). Error bars indicate ±1 standard error.

sensitivity to the difference between rising and level frequency contours (as was suggested by Gandour, 1978, with respect to Wang's, 1976, data). In other words, we cannot tell from this experiment whether the peak in identification (at the boundary between two linguistically relevant categories) was the cause of the peak in discrimination or whether the listeners exhibited category-like effects at this location along the continuum because it is a region of naturally higher sensitivity (i.e., to the difference between rising and level f_0 contours). In order to further investigate the interaction of linguistic category boundaries and regions of higher discrimination accuracy, it would be helpful to find a linguistic category boundary that does not lie in an obvious region of natural sensitivity. For example, in Cantonese, the distinction between the high-rising and the low-rising tones appears to be made primarily according to the f_0 at the end of the syllable. The fundamental frequency contours of both tones begin at approximately the same frequency, but high-rising tones typically rise much more (Bauer & Benedict, 1997). All of the fundamental frequency contours along a continuum from low-rising to high-rising are rising; all that differs between them is the extent (and thus the slope) of the rise. There is no suggestion that there should be a nonlinguistic perceptual discontinuity along such a range, and therefore, this distinction fits our requirements (although in the absence of other evidence, the possibility of a hitherto undiscovered nonlinearity in the perception of rising pitch glides cannot be ruled out). If listeners exhibit increased discrimination accuracy across the category boundary between low-rising and high-rising tones in Cantonese, this would suggest that these tones are perceived categorically, possibly as a consequence of the listeners' linguistic experience, rather than because of the presence of natural regions of heightened auditory sensitivity.

In addition to this boundary between categories with rising f_0 contours, the Cantonese system of contour tones also exhibits a category distinction made according to frequency contour differences to which we might expect listeners to be naturally sensitive, independently of their linguistic categories. The distinction between the low-falling and the low-rising tones is also apparently made primarily according to the frequency at the end of the syllable (Bauer & Benedict, 1997). However, this distinction also corresponds to the difference between falling (or level) and rising frequency contours—a highly salient auditory boundary at which we might expect to find a natural region of sensitivity (Schouten, 1985). Therefore, by testing listeners' sensitivity to tonal distinctions made along a Cantonese tone continuum ranging from low falling through low rising to high rising, we can investigate both listeners' natural sensitivity to tone contour differences and the possibility that contour tone contrasts are perceived categorically even in the absence of natural regions of sensitivity.

Method

Subjects. Twenty-eight native speakers of Cantonese participated in this experiment (22 women and 6 men). The subjects were assigned to either the discrimination group (16 subjects) or the identification group (12 subjects). None had participated in any of the previous experiments, and none reported any speaking or hearing disability.

Stimuli. Stimulus generation used the same natural speech models for the context sentence as those described in the previous experiments. Selection of the model natural contour syllables is described in the Method section of Experiment 2. All the stimuli started with an f_0 of 102 Hz, falling to the 98-Hz inflection point at 75 msec before rising or falling to the endpoint frequency. To construct the contour tone continuum, the extreme endpoint values for the low-falling and the high-rising tones (88 and 144 Hz, respectively) were taken as the ends of the continuum. These values were converted to mel (Taylor et al., 1999), and the distance between them, in mel, was subdivided equally (nine steps). The resulting intermediate mel values were then reconverted to hertz to provide a 10-token continuum from a low-falling tone to a high-rising tone in perceptually equal steps (9.2 mel each, about 6.5–6.7 Hz each). Exact endpoint frequency values are given in Table 3. Although this characterization of the continuum is not as obviously one-dimensional as that of the first experiment, like the stimuli in Experiment 2, these stimuli could be characterized in terms of the single acoustic dimension of the slope of the f_0 contour over the last three quarters of the syllable. According to this characterization, Stimulus 1 had a slope of −0.044 Hz/msec, Stimulus 2 had a slope of −0.015 Hz/msec, and so forth, with each step increasing in slope by 0.029–0.030 Hz/msec to Stimulus 10, with a slope of 0.22 Hz/msec.

The stimuli for the discrimination task consisted of all pairwise combinations of syllables separated by zero or one token along the continuum, with a 250-msec ISI. A one-step continuum was used in the present experiment (rather than a two-step continuum, as in the previous experiment) because the step size is somewhat larger (about 6.5 Hz, rather than 4.5 Hz) and because listeners are known to be more sensitive to frequency differences between pitch glides with different endpoints than they are to glides with differing onset frequencies (see Hombert, 1978, for a discussion). There was a total of 28 pairs, including identical pairs (1–1, 2–2, 3–3, etc.) as well as adjacent pairs (1–2, 2–1, 2–3, 3–2, etc.).

Procedure.The identification task procedure was identical to that in the previous experiment. Following presentation of the stimulus, the listeners were presented with three buttons arranged horizontally on the screen. Each button was labeled with a Chinese character representing one of the three syllables /ji25/ (*chair*, high-rising tone), /ji23/ (*ear*, low-rising tone), and /ji21/ (*child*, low-falling tone), from left to right. The response buttons were always presented in this order. Responses were collected automatically and were scored according to the tone of the character that was clicked (low falling, low rising, or high rising). The discrimination task followed exactly the format of Experiment 1, with the exception that, as in the second experiment, there were no audible warning tones preceding stimulus presentation.

Table 3
Fundamental Frequency Values for Stimuli in Experiment 3

Stimulus	Endpoint		Tone Class
	Hertz	Mel	
1	88.0	133.50	low falling
2	94.5	142.69	
3	101.0	151.88	
4	107.5	161.07	
5	114.2	170.26	
6	120.8	179.45	low rising
7	127.5	188.64	
8	134.3	197.83	
9	141.1	207.02	
10	148.0	216.21	high rising

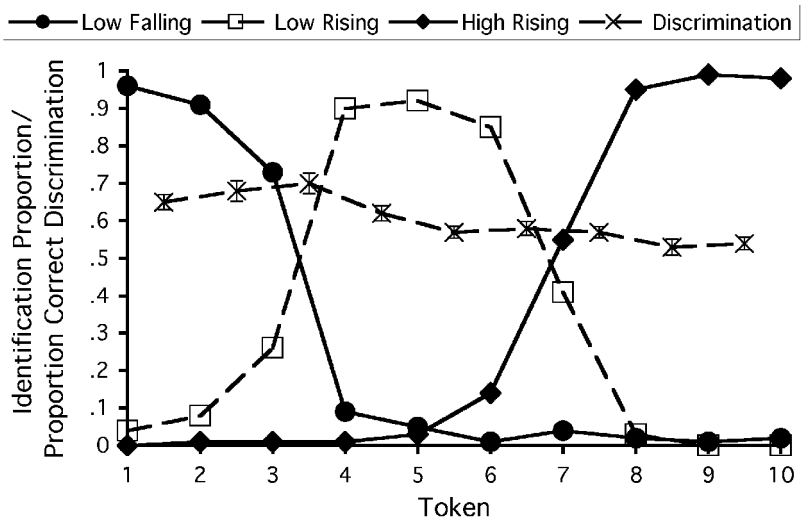

Figure 3. Identification and discrimination functions of a synthesized 10-token Cantonese contour tone continuum ranging from low falling (/ji21/) to high rising (/ji35/). Error bars indicate ±1 standard error.

Results

Figure 3 shows identification and discrimination curves for Experiment 3. This graph shows that listeners heard three distinct tone categories, with category boundaries between Tokens 3 and 4 (low falling to low rising) and between Tokens 6 and 7 (low rising to high rising).

The proportion of correct discriminations curve shows a peak between Tokens 3 and 4, but no evidence of a peak between Tokens 6 and 8. The rate of *different* responses to same pairs on the discrimination task was higher in this experiment than in the previous two, averaging 28% across the continuum. This may reflect the greater difficulty of discriminating between contour tones (as opposed to level tones in Experiment 1) and the greater difficulty of the one-step contrasts used in the present experiment (as opposed to the two-step contrasts in Experiment 2).

A one-way ANOVA of the arcsine-transformed percentages of correct discrimination responses was significant [$F(8,120) = 12.37, p < .001$]. Trend analysis showed that the expected quartic (double-peaked) trend among these points was not significant. However, a linear trend was found to be significant [$F(1,15) = 65.36, p < .001$]. Post hoc (Tukey HSD) analysis supported the observation of a generally linear trend falling from left to right (low to high frequency), so that the listeners' proportion of correct responses to the 3–4 pair was significantly greater ($p < .05$) than that for all the other pairs except 1–2 and 2–3. The results of this ANOVA did not support the appearance of a peak in accuracy at the 6–7 pair, for which the listeners' percentages of correct responses was not significantly greater than that for any other pair.

As in the previous experiments, the proportion of correct discriminations was compared with the proportion of correct discriminations predicted by applying Equation 1 (above) to the results of the identification task. A two-way ANOVA (one between-groups and one within-group measure) of the arcsine-transformed proportions showed a significant main effect of group, so that the discrimination predicted by identification ($\mu = .56$) was significantly smaller than that observed in the discrimination task [$\mu = .60$; $F(1,26) = 5.53, p = .03$], a main effect of pair [$F(8,208) = 16.87, p < .001$], and a significant interaction between the two [$F(2,208) = 6.35, p < .001$]. Post hoc (Tukey HSD) analysis showed a significant difference between identification and discrimination only at the 1–2 and 2–3 pairs, where the proportion of correct discriminations was significantly greater than was predicted by the identification task.

Discussion

On the present discrimination task, the listeners were more sensitive to differences at the low-frequency end of the continuum, as compared with differences at the high-frequency end. In this part of the continuum, the frequency contours change rapidly from clearly falling (Tokens 1 and 2) to nearly level (Token 3) to rising (Tokens 4 and up). One possible reason for increased sensitivity in this region is that listeners may simply be more sensitive to differences between frequency contours that change in different directions. In other words, the distinction between falling and rising contours may be psychoacoustically more salient than a similar magnitude of difference in endpoints between two rising contours. Alternatively, because the tonal inventory of Cantonese includes a low-level tone category in addition to the low-falling, low-rising, and high-rising categories, it is possible that this region of heightened sensitivity is due to the interaction (and perhaps overlap) of two category boundaries—the first between low-falling and low-level tones and the second between low-level and rising (both low-rising and high-rising) tones in

the present experiment. The observation that discrimination sensitivity in this low region is significantly greater than predicted by the identification task appears to lend provisional support to the hypothesis that auditory sensitivity, not linguistic experience, is the critical factor. However, because the identification task provided only two kinds of contours as possible responses, falling and rising, it is not possible to answer this question definitively on the basis of this evidence alone.

One clue to the resolution of this problem might be found in the lack of a significant peak in discrimination accuracy near the category boundary between the low-rising and the high-rising tones, despite evidence for a category boundary in this region of the identification curve. This pattern of results suggests that, like level tone contrasts in the first experiment, a contrast between tones that share the same general direction of change (in this case, rising) is not perceived categorically (shows no discrimination peak across category boundaries), whereas a contrast between tones that differ in terms of the direction of change in frequency may be perceived categorically. This is consistent with the results reported by Gandour (1981), who found that the degree of contour and the direction of change of f_0 accounted for more of the variance of a multidimensional scaling solution using Cantonese tonal stimuli than did the average f_0 of the contours. Gandour's (1981) results suggest that, for stimuli presented in isolation, Cantonese listeners give relatively more weight to the direction of change of f_0, a feature that distinguishes the low-falling tone from the low-rising tone, than they do to the average f_0 of the syllable, a feature that could be used to distinguish the low-rising tone from the high-rising tone (and between the low-level, mid-level, and high-level tones in Experiment 1 as well).

EXPERIMENT 4

Research by Wong and Diehl (in press), Moore and Jongman (1997), and Leather (1983, among others, suggests that listeners' ability to accurately locate boundaries between tone categories is facilitated by the presence of external contextual information about the talker's pitch range. Therefore, we might expect a tone continuum presented in sentential context to exhibit more evidence of categorical perception (sharper identification boundaries, stronger discrimination peaks across boundaries) than was observed in Experiment 1 for stimuli presented out of context. In this experiment, we repeated Experiment 1, using the same stimuli, but now presented in sentence context.

Method

Subjects. Twenty-eight native speakers of Cantonese (21 women and 7 men) participated in this experiment. The subjects were undergraduate speech pathology students in the Department of Speech and Hearing Sciences at the University of Hong Kong and friends or relatives of students. Sixteen listeners were assigned to the identification task (these were the same individuals who had participated in the out-of-context identification task in Experiment 1), and 12 were assigned to the discrimination task (these listeners had not previously participated in any tone perception experiments).

Stimuli. The stimuli for this experiment were identical to those in Experiment 1, with the exception of the inclusion of two different context sentences. For the identification task, the sentence /ŋɔ25 wui25 tʰɔk22 ji pɛi25 lɛi24 tʰɛŋ55/ (*I will read —— for you to hear*) was used. A sentence with the target syllable located in the middle was used to mitigate the influence of sentential intonation that might affect the fundamental frequency of the target syllable in final position (Vance, 1976; Zee, 1998). The identification sentence was recorded with a /ji/ syllable with a mid-level tone in the target position to ensure that appropriate coarticulatory cues were included in the pre- and posttarget syllables. The discrimination sentence was recorded in isolation. For the discrimination task, a different sentence was used: /kɔɪ23 tɛɪ22 hɑɪ22 mɑɪ22 jɑt55 j ŋ22 ___? *Are they the same?*). In this case, it was necessary to use a context sentence with the target located at the end, in order to accommodate the linguistically artificial task of discriminating between two acoustically similar syllables.

The sentences were recorded onto the hard disk of an Apple Power-Macintosh 7100/AV, using a Bruel & Kjaer Type 4003 microphone and a Type 2812 MK II microphone preamplifier, normalized in peak amplitude and used as a basis for synthesizing the synthetic context sentence. In the case of the identification sentence, the talker who produced the sentence was the same as the talker who produced all of the target syllables. In the case of the discrimination sentences, this talker was no longer available for recording, and so a substitute talker of the same age, gender, and sociolinguistic background produced the sentence. Following recording, the f_0 of the discrimination sentence was subsequently scaled linearly until the mean f_0 was identical to that of the identification sentence and then was resynthesized using the Praat 4.0 PSOLA algorithm (Boersma & Weenink, 2002).

Six measures were taken every millisecond (every 10 msec for the discrimination sentence) throughout the duration of each context sentence: f_0, the center frequencies of the first four formants ($F1$, $F2$, $F3$, and $F4$), and the amplitude. These raw measures were used as the initial settings for SenSyn (Sensimetrics Corp.), a Klatt-style cascade/parallel formant synthesizer (Klatt & Klatt, 1990) implemented on an Apple Macintosh G3 as for the target syllables of the first experiment. The identification stimuli and discrimination pairs of stimuli from Experiment 1 were then digitally spliced into the appropriate context sentences. The identification syllables were inserted in place of the /ji33/ syllable that was already there, whereas the discrimination pairs were inserted 150 msec after the end of the context sentence.

Procedure. The experimental procedures were identical to those in Experiment 1, with the exception that no auditory warning signal was given prior to stimulus presentation on each trial.

Results

Figure 4 shows the listeners' average response percentages on the identification task superimposed on a plot of the listeners' average proportion of correct discriminations for each adjacent pair. As in the results for Experiment 1, there is no evidence to suggest that the listeners were more accurate at discriminating stimulus differences across category boundaries: There were no peaks in discrimination between the members of the 3–4 or the 6–7 pair where the identification functions cross over. A one-way repeated measures ANOVA of the arcsine-transformed proportions of correct discriminations revealed a significant effect of pair [$F(8,88) = 4.35, p < .001$], and a trend analysis showed a significant linear trend between the points [$F(1,11) = 8.62, p = .01$]. Post hoc (Tukey HSD) analysis showed that the listeners' proportion correct for the 1–2 pair (.75) was significantly greater than that for the 7–8 (.65), 8–9 (.62), and 9–10 (.60) pairs, and the proportion correct for

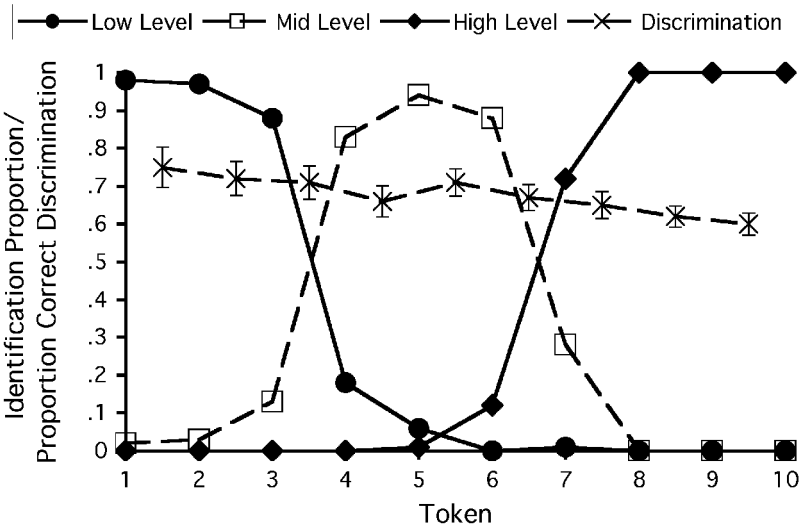

Figure 4. Identification and discrimination functions of a synthesized 10-token Cantonese level-tone continuum, as in Figure 1, except that these tokens were presented in meaningful sentence context (see the text for details). Error bars indicate ±1 standard error.

the 2–3 pair (.72) was also significantly greater than that for the 9–10 pair ($p < .05$ on all contrasts; no other contrasts reached significance). These results suggest that listeners are more sensitive to contrasts at the low-frequency end of the continuum than to those at the high end.

In order to compare identification with discrimination, we calculated the proportion of correct discriminations for each pair as predicted by the identification task, as in Experiment 1. The arcsine-transformed proportions correct for both the identification and the discrimination tasks were submitted to a two-way (mixed factorial) ANOVA with one factor between groups (identification vs. discrimination) and one factor within groups (pair). The results showed that discriminability as predicted by the identification task ($\mu = .58$) was significantly lower than the observed proportion of correct discriminations [$\mu = .68$; $F(1,26) = 9.02, p = .006$], that there was a significant difference between pairs across the continuum [$F(8,208) = 9.71, p < .001$], and that there was a significant interaction between pairs and group [$F(8,208) = 8.48, p < .001$]. Post hoc (Tukey HSD) analysis revealed significant differences between identification and discrimination at the 1–2, 2–3, and 5–6 pairs, for which identification significantly underpredicted discrimination by about 20 percentage points: For the 1–2 pair, the proportion of correct discriminations predicted by identification performance was .50, as compared with an actual discrimination proportion correct of .75; corresponding values for the 2–3 pair were .52 versus .72; and for the 5–6 pair, .52 versus .71.

Figure 5 shows a comparison of predicted discrimination based on identification within context (the present experiment) with that predicted by identification out of context (Experiment 1). A two-way repeated measures ANOVA (both factors, context and boundary pair, were within group) of these arcsine-transformed proportions showed significant effects of context [$F(1,15) = 36.83, p < .001$] and of pair [$F(8,120) = 12.80, p < .001$] and a significant interaction between the two [$F(8,120) = 12.55, p < .001$]. Post hoc (Tukey HSD) analysis showed that there was no significant difference between predicted discrimination accuracy in context versus out of context except at the 3–4 and 6–7 pairs, the expected locations of category boundaries. At these two locations, discrimination predicted by in-context identification (.78 at 3–4, .74 at 6–7) was significantly greater than that predicted by out-of-context identification (.54 at 3–4 and .52 at 6–7). The presence of context in the identification task clearly served to enhance cross-boundary sensitivity but did not appear to affect within-category discrimination.

A comparison of the arcsine-transformed discrimination scores for the in-context discrimination group ($n = 12$, the present experiment) and the no-context discrimination group ($n = 15$, Experiment 1), using a mixed factorial ANOVA (where context was between groups and stimulus pair was within subjects), showed a significant effect of pair [$F(8,200) = 4.68, p < .001$] and a significant interaction of context and pair [$F(8,200) = 3.22, p = .002$] but no effect of context. Post hoc (Tukey HSD) comparisons showed no significant difference between discrimination of any pair presented in context versus out of context, except for the 8–9 pair, for which discrimination in context (.62) was significantly lower than that out of context (.73). In this experiment, the presence of a preceding sentence context did not provide an overall increment to discrimination sensitivity. Indeed, the significant interaction between context and pair and examination of

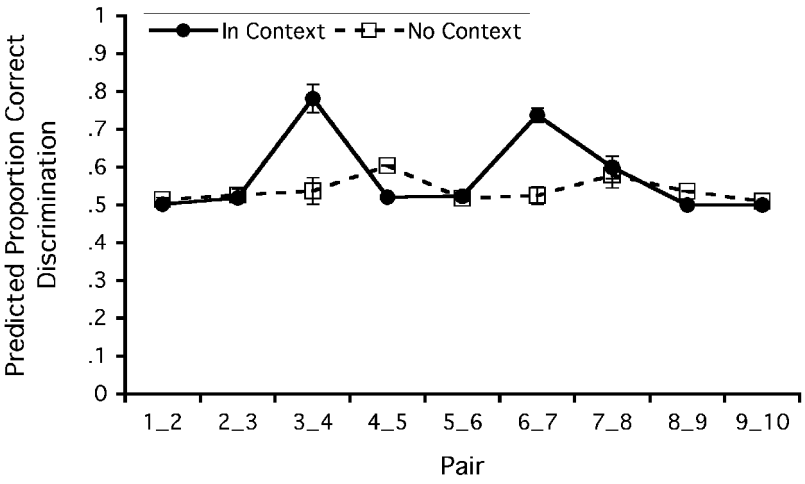

Figure 5. Predicted proportions of correct discriminations calculated from response proportions obtained with presentation of level tone stimuli in sentence context (in context) and in isolation (no context). The predicted proportion of correct discriminations was calculated according to Equation 1 (see the text), based on Liberman, Harris, Hoffman, and Griffith (1957). Error bars indicate ±1 standard error.

the post hoc analyses suggest that, in this case, presenting stimuli in context may have caused a slight *decrease* in sensitivity at the high-frequency end of the continuum.

Discussion

Research by Moore and Jongman (1997) and Wong and Diehl (in press) showed that identification of level tone categories is facilitated by knowledge of the talker's pitch range. The results of the present identification experiment are consistent with the hypothesis that listeners use extrinsic information about the talker's pitch range to estimate the location of tone category boundaries. Predicted discrimination accuracy across the category boundaries was significantly higher when the stimuli were presented in context in the present experiment, as compared with presentation in isolation in Experiment 1. However, it should be noted that the present discrimination experiment provided no evidence to support the hypothesis that level tones are perceived categorically, even though these stimuli were presented in the context of a meaningful carrier sentence. In the discrimination task, the listeners showed no peak in sensitivity across category boundaries, as would be predicted were level tones perceived categorically. The listeners do appear to have been somewhat affected by the presence of a sentence context, because there was a slight decrease in sensitivity to pitch differences at the higher frequency end of the continuum in context. However, the lack of any peak in sensitivity across category boundaries even in the presence of extrinsic talker information suggests that the listeners did not make use of category-relevant aspects of contextual information in a *same/different* decision task. This pattern of behavior seems to be consistent with the results of a wide variety of experiments that have suggested that listeners tend to have more access to contextual information when performing identification-type tasks than when performing discrimination tasks (cf. Macmillan, 1987).

One possible cause of the observed decrease in sensitivity at higher frequencies following the sentence context is that Cantonese listeners typically expect pitch to drop over the course of an utterance, a phenomenon called *tone declination* or *downdrift* (Vance, 1976; Wong, 1999). Because the frequency specifications of the stimuli used in this experiment were based on recordings of syllables produced in the middle of a sentence, not at the end, tokens toward the lower frequency end of the continuum probably lie closer to where the center of the talker's pitch range would be at the end of a sentence. However, tokens at the higher frequency end of the continuum are probably higher in frequency than listeners would expect to hear at the end of this sentence spoken by this talker. It is possible that the listeners' expectations of downdrift shifted the focus of their attention away from higher frequency regions of the continuum, thereby slightly reducing sensitivity to pitch differences in those regions. In principle, the talker-related information from the context sentence that induced the expectation of downdrift could have allowed the listeners to estimate the location of category boundaries in the discrimination task, although the boundary locations might have been shifted downward under the influence of the expectation of downdrift. However, this does not appear to have happened, since there was no context-related improvement in discrimination sensitivity across any category boundary.

It is also possible that differences in voice quality between the context sentence and the target stimuli used in the in-context discrimination task encouraged the listeners to ignore the context and make their similarity judg-

ments as if the target stimuli had been presented in isolation. Although no subject mentioned such a disjunction of percept and the observed decrease in sensitivity at the higher end of the continuum suggests that information from the context did play some role in the listeners' judgments, we cannot rule out this possibility. From this experiment alone, we cannot be certain whether the listeners' apparent failure to use available extrinsic information about level tone category boundary locations in a *same/different* discrimination task is due to stimulus-related factors or whether it is a general characteristic of discrimination tasks.

EXPERIMENT 5

The results of comparing in-context and out-of-context identification can also be used to further explore the perception of contour tone stimuli. Since the f_0 pattern of a contour tone extends through a linguistically determined subset of the talker's pitch range, the f_0 contour of a syllable with a contour tone could provide intrinsic information about the talker's pitch range. This intrinsic information about pitch range might, in turn, be sufficient to allow listeners to normalize tones across talkers even in the absence of extrinsic pitch information. If listeners show little or no difference between in-context and out-of-context cross-boundary sensitivity to contour tones, this would support the hypothesis that listeners do not require the presence of extrinsic context to correctly identify contour tones. In this experiment, we examined the effect of the presence of context on cross-boundary sensitivity, using the contour tone stimuli from Experiments 2 and 3.

Method

Subjects. Thirty-two native speakers of Cantonese participated in this experiment (28 women and 4 men), all students at the University of Hong Kong. Sixteen listeners were randomly assigned to each of two groups: the onset group (using the stimuli from Experiment 2) and the offset group (using the stimuli from Experiment 3). None of the subjects in this experiment had participated in any of the previous experiments. The experiment lasted approximately 20 min, and the subjects were compensated for their time.

Stimuli. The stimuli were identical to those in Experiments 2 and 3. All the stimuli were presented in the same sentence context as that in Experiment 4.

Procedure. Both identification tasks were identical to those in the previous experiment.

Results

As in the previous experiments, identification performance was used to calculate predicted discrimination of each pair. Figure 6 shows the listeners' predicted discrimination in context (the present experiment) and out of context (Experiment 2) across the rising-to-level tone continuum, and Figure 7 shows the same values for the low-falling to high-rising continuum (in-context data from the present experiment, no-context data from Experiment 3). Note that the rising-to-level group in the present experiment also made two-step comparisons, as in Experiment 2.

Because we had no a priori reason to assume that the listeners' sensitivity to the three types of stimuli should be comparable, responses to the two sets of stimuli were analyzed separately. For the rising-to-level continuum (from Experiment 2), a two-way mixed factorial ANOVA of the arcsine-transformed proportions showed a significant effect of context, so that the predicted proportion of correct discriminations for the no-context group ($\mu = .58$)

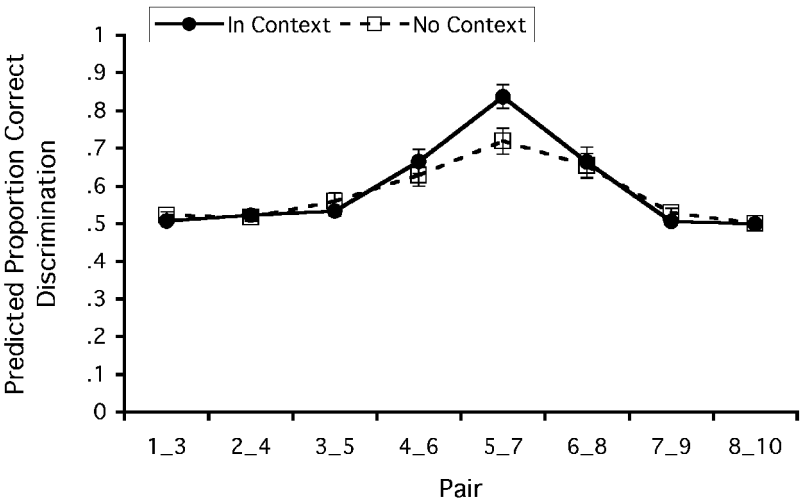

Figure 6. Predicted proportions of correct discriminations calculated from response proportions obtained with presentation of stimuli from the rising-to-level tone continuum in sentence context (in context) and in isolation (no context). The predicted proportion of correct discriminations was calculated according to Equation 1 (see the text), based on Liberman, Harris, Hoffman, and Griffith (1957). Error bars indicate ±1 standard error.

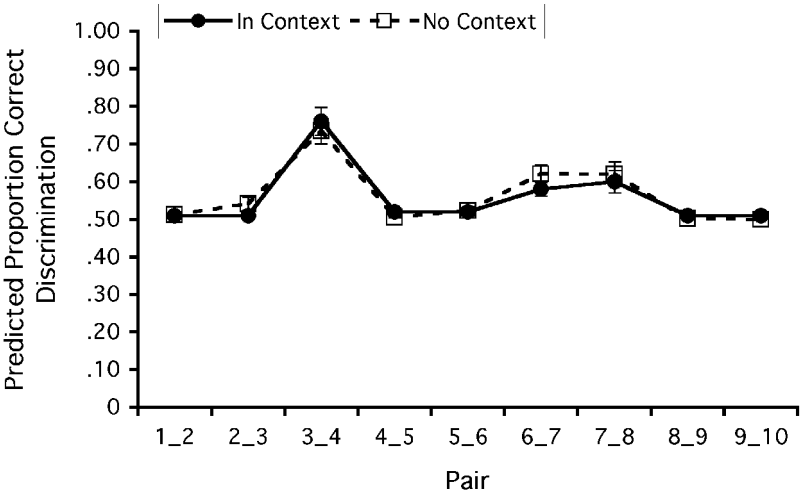

Figure 7. Predicted proportions of correct discriminations calculated from response proportions obtained with presentation of stimuli from the low-falling to high-rising tone continuum in sentence context (in context) and in isolation (no context). The predicted proportion of correct discriminations was calculated according to Equation 1 (see the text), based on Liberman, Harris, Hoffman, and Griffith (1957). Error bars indicate ±1 standard error.

was significantly smaller than that for the in-context group [$\mu = .59$; $F(1,26) = 4.66$, $p = .04$], a main effect of pair [$F(7,182) = 24.56$, $p < .001$], and a significant interaction between the two [$F(7,182) = 2.20$, $p = .04$]. Planned comparison of means at the category boundary (the 5–7 pair) showed that discrimination predicted from in-context identification was significantly greater than that predicted by identification in isolation [$F(1,26) = 5.64$, $p = .03$].

For the low-falling to high-rising continuum (from Experiment 3), a two-way mixed factorial ANOVA of the arcsine-transformed proportions showed a significant effect of pair [$F(8,208) = 26.74$, $p < .001$] but no significant difference between the in-context ($\mu = .56$) and the out-of-context ($\mu = .56$) group performances and no interaction between group and pair. A planned comparison of the means showed no significant difference between the groups at either of the two boundary locations.

One further observation is worthy of note, relating to the broadness of the peak in identification sensitivity observed at the boundary between the low-rising and the high-rising categories (Figure 7), encompassing at least two pairs of stimuli (6–7 and 7–8). An analysis of the individual listeners' response patterns suggested that, in the in-context condition, some of the listeners showed a strong peak at 6–7, whereas others showed a strong peak at 7–8, but nearly all the listeners showed a clear peak between only two tokens (either 6–7 or 7–8). Only 1 listener showed a broad peak encompassing both the 6–7 and the 7–8 pairs in the in-context condition, similar to the group results. In contrast, in the no-context condition, most of the listeners tended to exhibit a shallow elevation in proportion of correct discriminations somewhere in the upper half of the continuum, but the exact location ranged from 5–6 to 8–9 and usually encompassed one or two neighboring pairs (e.g., 6–7 and 7–8, or 7–8 and 8–9). Thus, although group performance looked quite similar across the two conditions, an analysis of individual response patterns provided some evidence that presenting these stimuli in context can improve the determination of category boundaries.

Discussion

The presence of extrinsic context sharpened the category boundary along a rising-to-level continuum but had no effect on a contour tone continuum varying in endpoint f_0. These results support the hypothesis that contour tones may be self-normalizing, although only in the case of the stimuli that begin at a similar pitch. The manner in which this hypothesis fails in the case of the rising-to-level continuum may shed light on aspects of the mechanism of intrinsic tone normalization (the way in which f_0 contours contribute to normalization).

In studies of Cantonese tone production (e.g., Bauer & Benedict, 1997), the low-rising and high-rising tones are primarily distinguishable according to their ending f_0. All three contour tones start at about the same level in the talker's pitch range (at Level 2). Therefore, the presence of a rising (or a falling) contour may provide sufficient information to allow listeners to estimate the dimensions of the talker's tone space on the basis of the initial pitch of the syllable. Once the syllable is identified as having a changing contour (of any sort), listeners may assume that the onset frequency of the syllable indicates that the talker's Pitch Level 2—and in principle, the other pitch levels—and the boundaries between tone categories based on pitch level could be predicted on that basis. In the case of the

rising-to-level continuum used here, the result of such a computation would be misleading. On the basis of intrinsic information alone, each stimulus would predict a somewhat different location for the category boundary. Providing an extrinsic context, as in the present experiment, may have allowed the listeners to partially override this misleading intrinsic information, resulting in a sharper category boundary. In the case of the low-falling to high-rising continuum, all the stimuli started at the same (talker-appropriate) frequency, thereby allowing an intrinsic normalization process based on syllable-initial pitch to function correctly. It should be noted that the rising-to-level continuum still showed evidence of the presence of category boundaries in the identification task, even when the stimuli were presented in isolation. The effect of extrinsic context may occur merely as a further increase in sensitivity across a boundary that is already identifiable on the basis of intrinsic information and/or as a consequence of the presence along the continuum of a region of naturally heightened auditory sensitivity.

GENERAL DISCUSSION

The results presented in Experiments 1–3 reflect the apparently conflicting findings of Abramson (1979) and Wang (1976). Abramson (1979) found that a Thai level tone continuum was not perceived categorically, whereas Wang (1976) found that a Mandarin continuum from rising to level tone was. In corroborating the findings of these previous researchers, the results of the experiments presented here suggest that, with respect to categorical perception, contrasts between level tones behave more like vowel-related contrasts than like consonantal contrasts. On the other hand, contrasts between contour tones seem to be more similar to consonantal feature contrasts in this respect. One exception to this rule may be the contrast between the low-rising and the high-rising tone categories. This contrast does not appear to be perceived categorically under any circumstances, possibly accounting for the relative difficulty many native speakers exhibit in distinguishing between these two tones (e.g., Ciocca & Lui, 2003; Fok Chan, 1974, p. 110). Furthermore, the present results suggest that extrinsic contrast influences the perception of level and contour tone contrasts to differing degrees. Here, we argue that the overall pattern of results supports a model of tone perception in which at least three factors interact to facilitate the determination of category boundaries within a talker-specific pitch space: (1) the presence of regions of natural auditory sensitivity, (2) learned associations between particular acoustic (f_0) patterns and linguistic categories, and (3) the estimation of a talker's tone space on the basis of intrinsic and extrinsic acoustic information.

First, the boundaries between some contour tones in Cantonese appear to be located in regions of perceptual space where it seems highly likely that listeners will exhibit naturally heightened sensitivity (changes in the sign of the slope of a pitch glide). The transitions from falling to rising and rising to level pitch contours are auditorily salient boundaries that have been demonstrated in listeners unfamiliar with tone languages listening to speech and nonspeech stimuli (Klatt, 1973; Schouten, 1985).The argument that the location of linguistic category boundaries may be influenced by the location of nonlinearities in auditory perception is not unique to tone perception. For example, listeners show a category boundary effect around 20 msec on a voice onset time continuum, corresponding to the voiced/voiceless category boundary in English and many other languages, even when those listeners speak a language that does not have a category boundary there (Williams, 1977) and when those listeners are not human (Kuhl & Miller, 1975, 1978; see Kuhl, 1987, and Rosen & Howell, 1987, for discussions). Following Rosen and Howell, it seems plausible that regions of natural auditory sensitivity may function as *attractors* for category boundaries in the socio-historical development of linguistics systems, so that category boundaries located in psychophysically sensitive regions of perceptual space may be cross-linguisticially more stable and/or common. This cannot be the only determinant of tone language inventories, of course, because the vast majority of tone languages exhibit a contrast between only two (phonologically)level tones and many languages with complex tonal inventories exhibit more level tones than contour tones. However, in cases in which perceptually stable contrasts arise, it seems plausible that they should also prove linguistically stable as well (see Maddieson, 1978, and Wang, 1967, for further discussions of this topic).

Although an auditory sensitivity account may be sufficient for the results observed in Experiment 3 (at the boundary between the low-falling and the low-rising tones), the results presented by Wang (1976) suggest that a purely psychophysical explanation may not be sufficient to account for the results observed in Experiment 2 at the boundary between a rising and a level tone. In Wang's (1976) study, two naive Chinese listeners showed a strong identification boundary between rising and level tones at about the midpoint of the continuum, with a corresponding peak in discrimination accuracy. A second pair of Chinese listeners who had experience doing psychophysical experiments showed a comparably heightened discrimination accuracy at the center of the continuum, but their discrimination accuracy remained relatively high toward the level end of the continuum. Finally, two highly trained listeners (the experimenters) showed no peak in discrimination across the category boundary but still showed a high degree of discrimination accuracy toward the level end of the continuum in a manner similar to that exhibited by three native speakers of English unaccustomed to hearing Chinese. Wang's (1976) discussion of these results did acknowledge that there was evidence for a language-independent region of heightened sensitivity along the perceptual continuum from rising to level pitch glides but also invoked the presence of learned associations between f_0 patterns and linguistic categories. In effect, Wang (1976) argued that Chinese listeners have learned to suppress a

natural auditory sensitivity to the difference between rising and truly level pitch in order to facilitate the acquisition of a similarly heightened sensitivity to differences between strongly rising and *almost* level pitch glides, because the inherent variability of pitch production made it unlikely that talkers could produce a truly level f_0 contour in a consistent manner. Thus, listeners were best served by shifting the location of an intrinsic sensitivity to the distinction between rising and level pitch glides. English speakers and Chinese listeners trained to ignore the influence of learned patterns of sensitivity showed the influence only of natural auditory sensitivity. Although care must be taken in interpreting results from only a small number of subjects, Wang's (1976) results are consistent with a model of tone perception in which a learned association interacts with natural regions of heightened auditory sensitivity.

In any case, in addition to the role of regions of natural sensitivity, listeners must also make use of learned associations between f_0 patterns and linguistic tone categories. Although listeners may be naturally predisposed to accurately discriminate one set of sounds from another, they must still learn that the two sets of sounds represent different linguistic categories. In cases in which listeners do not appear to exhibit naturally heightened auditory sensitivity across category boundaries, such as among level tones or between rising tones, listeners must use some other principle to define the location of boundaries, and this is the role of the third factor involved in lexical tone perception. Those category boundaries that are not defined in terms of the location of regions of naturally heightened auditory sensitivity appear to be defined in a talker-dependent manner. The results presented by Wong and Diehl (2003) showed that Cantonese listeners are strongly influenced by contextual information when making level tone category decisions, providing a basis for the observation that listeners have difficulty identifying level tone categories presented in isolation. The differences observed here between the discrimination performance predicted by the level tone identification task in and out of context (see Figure 5) strongly support this hypothesis. Listeners showed significantly sharper category boundaries when level tone stimuli were presented in a context that provided information about the pitch range of the talker.

There are, however, at least two sources of information about pitch range available to listeners. In some cases, the f_0 pattern of a single syllable may provide sufficient intrinsic information about the talker's pitch range to allow accurate estimation of category boundary locations (e.g., in Cantonese, syllables with changing pitch contours all start at about the same pitch level, so that the identification of a change in pitch can indicate significant aspects of the structure of a given talker's pitch space), whereas in other cases, listeners must seek out supplementary information about the talker identity from the pitch patterns of surrounding syllables to determine talker-specific boundary locations (e.g., in Cantonese, syllables with level tone contours provide no intrinsic information relevant to determining the location of category boundaries). Although natural auditory sensitivity may have played a role in determining the location of the boundary between rising and level tones (in Experiment 2), the presence of extrinsic context appears to have provided an additional benefit by further improving the sharpness of the identification boundary between them (in Experiment 5) and to have enabled the determination of boundary locations between level tone categories in Experiment 4. In contrast, the presence of an extrinsic context did not seem to improve identification performance on the low-falling to low-rising to high-rising continuum (in Experiment 5). This could mean that these category boundaries are sufficiently well determined by natural auditory sensitivities and/or intrinsic information about talker pitch range. With respect to the low-rising versus high-rising contrast, it seems likely that neither an account based solely on the role of intrinsic information nor one based soley on the presence of a region of natural sensitivity would be entirely sufficient to completely determine the category boundary, since the presence of extrinsic context did improve sharpness of the boundary for each individual and because Wang (1976) showed that the true region of natural sensitivity is more likely to be closer to the truly level contour. Thus, the degree of categoriality with which a particular tonal contrast is perceived depends on a complex interaction among regions of naturally heightened auditory sensitivity, learned associations between particular acoustic (f_0) patterns and linguistic categories, and talker-specific speech processing based on intrinsic and extrinsic acoustic information about the talker's pitch range.

REFERENCES

Abramson, A. S. (1976). Thai tones as a reference system. In W. Gething, G. Harris, & P. Kullavanijaya (Eds.), *Tai linguistics in honor of Fang-Kuei Li* (pp. 1-12). Bangkok: Chulalongkorn University Press.

Abramson, A. S. (1979). The noncategorical perception of tone categories in Thai. In B. Lindblom & S. Öhman (Eds.), *Frontiers of speech communication research* (pp. 127-134). London: Academic Press.

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese phonology*. Berlin: Mouton de Gruyter.

Boersma, P., & Weenink, D. (2002). *Praat program*. Retrieved May 9, 2002 from University of Amsterdam, Institute of Phonetics Sciences Web site at http://www.fon.hum.uva.nl/praat/manual/Praat_program.html.

Burns, E. M., & Sampat, K. S. (1980). A note on possible culture-bound effects in frequency discrimination. *Journal of the Acoustical Society of America*, **68**, 1886-1888.

Chao, Y. R. (1947). *Cantonese primer*. Cambridge, MA: Harvard University Press.

Chao, Y. R. (1948). *Mandarin primer*. Cambridge, MA: Harvard University Press.

Ciocca V., & Lui, J. Y.-K. (2003). The development of the perception of Cantonese lexical tones. *Journal of Multilingual Communication Disorders*, **1**, 141-147.

Fok Chan, Y. Y. (1974). *A perceptual study of tones in Cantonese*. Hong Kong: University of Hong Kong, Centre of Asian Studies.

Francis, A. L., & Ciocca, V. (2003). *Stimulus presentation order and the perception of lexical tones in Cantonese*. Manuscript submitted for publication.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). Identification and discrimination of synthetic vowels. *Language & Speech*, **5**, 171-189.

GANDOUR, J. T. (1978). The perception of tone. In V. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 41-76). New York: Academic Press.
GANDOUR, J. T. (1981). Perceptual dimensions of tone: Evidence from Cantonese. *Journal of Chinese Linguistics*, **9**, 21-36.
GANDOUR, J. T. (1983). Tone perception in far eastern languages. *Journal of Phonetics*, **11**, 149-175.
HOMBERT, J.-M. (1978). Consonant types, vowel quality, and tone. In V. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 77-111). New York: Academic Press.
HOWARD, D., ROSEN, S., & BROAD, V. (1992). Major/minor triad identification and discrimination by musically trained and untrained listeners. *Music Perception*, **10**, 205-220.
INTERNATIONAL PHONETIC ASSOCIATION (1999). *Handbook of the International Phonetic Association.* Cambridge: Cambridge University Press.
KLATT, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. *Journal of the Acoustical Society of America*, **53**, 8-16.
KLATT, D. H., & KLATT, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, **87**, 820-857.
KUHL, P. K. (1987). The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad (Ed.), *Categorical perception* (pp. 355-386). Cambridge: Cambridge University Press.
KUHL, P. K., & MILLER, J. D. (1975). Speech perception by the chinchilla: Voiced–voiceless distinction in alveolar plosive consonants. *Science*, **190**, 69-72.
KUHL, P. K., & MILLER, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America*, **63**, 905-917.
LADD, D. R. (1996). *Intonational phonology.* Cambridge: Cambridge University Press.
LEATHER, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics*, **11**, 373-382.
LIBERMAN, A. M., HARRIS, K. S., EIMAS, P. D., LISKER, L., & BASTIAN, J. (1961). An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language & Speech*, **4**, 175-195.
LIBERMAN, A. M., HARRIS, K. S., HOFFMAN, H. S., & GRIFFITH, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358-368.
LIBERMAN, A. M., HARRIS, K. S., KINNEY, J. A., & LANE, H. (1961). The discrimination of relative onset time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, **61**, 379-388.
MACMILLAN, N. A. (1987). Beyond the categorical/continuous distinction: A psychophysical approach to processing modes. In S. Harnad (Ed.), *Categorical perception* (pp. 53-88). Cambridge: Cambridge University Press.
MADDIESON, I. (1978). Universals of tone. In J. H. Greenberg (Ed.), *Universals of human language: Vol. 2. Phonology* (pp. 335-365). Stanford, CA: Stanford University Press.
MATTHEWS S., & YIP, V. (1994). *Cantonese: A comprehensive grammar*. London: Routledge.
MOORE, C. B., & JONGMAN, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America*, **102**, 1864-1877.
NÁBĚLEK, I. V., NÁBĚLEK, A. K., & HIRSH, I. J. (1970). Pitch of tone bursts of changing frequency. *Journal of the Acoustical Society of America*, **48**, 536-553.
POLLOCK, I., & PISONI, D. (1971). On the comparison between identification and discrimination tests in speech perception. *Psychonomic Science*, **24**, 299-300.
REPP, B. H., HEALY, A. F., & CROWDER, R. G. (1979). Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 129-145.
ROSEN, S., & HOWELL, P. (1987). Auditory, articulatory, and learning explanations of categorical perception in speech. In S. Harnad (Ed.), *Categorical perception* (pp. 113-160). Cambridge: Cambridge University Press.
SCHOUTEN, M. E. H. (1985). Identification and discrimination of sweep tones. *Perception & Psychophysics*, **37**, 369-376.
STAGRAY, J. R., & DOWNS, D. (1993). Differential sensitivity for frequency among speakers of a tone and a non-tone language. *Journal of Chinese Linguistics*, **21**, 144-163.
STEVENS, K. N., LIBERMAN, A. M., STUDDERT-KENNEDY, M., & ÖHMAN, S. E. G. (1969). Crosslanguage study of vowel perception. *Language & Speech*, **12**, 1-23.
TANNER, W. P., & RIVETTE, G. L. (1964). Experimental study of "tone deafness." *Journal of the Acoustical Society of America*, **36**, 1465-1467.
TAYLOR, P., CALEY, R., BLACK, A. W., & KING, S. (1999). *Edinburgh Speech Tools Library, System Documentation Edition 1.2 for 1.2.0.* Retrieved November 22, 2002 from http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/x13778.htm.
VANCE, T. J. (1976). An experimental investigation of tone and intonation in Cantonese. *Phonetica*, **33**, 368-392.
WANG, W. S.-Y. (1967). The phonological features of tone. *International Journal of American Linguistics*, **33**, 93-105.
WANG, W. S.-Y. (1976). Language change. In S. R. Harnad, H. D. Steklis, & J. Lancaster (Eds.), *Origins and evolution of language and speech* (Annals of the New York Academy of Sciences, Vol. 280, pp. 61-72). New York: New York Academy of Sciences.
WILLIAMS, L. (1977). The perception of stop consonant voicing by Spanish–English bilinguals. *Perception & Psychophysics*, **21**, 289-297.
WONG, P. C. M. (1999). The effect of downdrift in the production and perception of Cantonese level tone. *Proceedings of the XIVth International Congress of Phonetic Sciences*, **3**, 2395-2398.
WONG, P. C. M., & DIEHL, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, & Hearing Research*, **46**, 413-421.
ZATORRE, R. J., & HALPERN, A. R. (1979). Identification, discrimination, and selective adaptation of simultaneous musical intervals. *Perception & Psychophysics*, **26**, 384-395.
ZEE, E. (1998). Resonance frequency and vowel transcription in Cantonese. In C. F. Sun (Ed.), *Proceedings of the 10th North American Conference of Chinese Linguistics and the 7th Annual Meeting of the International Association of Chinese Linguistics* (pp. 90-97). Los Angeles: Graduate Students in Linguistics (GSIL) at USC.

NOTES

1. In this paper, we distinguish between *fundamental frequency*, *pitch*, and *tone*. Fundamental frequency (f_0) is an acoustic measure that refers to the periodicity of a sound. Pitch is the perceptual correlate of this frequency, and tone refers to the linguistic (phonological, categorical) use of pitch to distinguish between lexical items. It should be noted that, as with other linguistic categories, there may be multiple acoustic parameters sufficient to cue tonal distinctions, including, but not limited to, f_0 differences.

2. Throughout the paper, standard IPA transcription (International Phonetic Association, 1999) is used for transcribing segments. Tones are transcribed in a manner similar to the Chao tone number system used by Bauer and Benedict (1997).

(Manuscript received February 13, 2001;
revision accepted for publication March 19, 2003.)