Stimulus presentation order and the perception of lexical tones in Cantonese[a)]

Alexander L. Francis[b)]
Department of Audiology and Speech Sciences, Purdue University, West Lafayette, Indiana 47907

Valter Ciocca[c)]
Department of Speech and Hearing Sciences, University of Hong Kong

(Received 9 August 2002; revised 18 May 2003; accepted 30 June 2003)

Listeners' auditory discrimination of vowel sounds depends in part on the order in which stimuli are presented. Such presentation order effects have been argued to be language independent, and to result from psychophysical (not speech- or language-specific) factors such as the decay of memory traces over time or increased weighting of later-occurring stimuli. In the present study, native Cantonese speakers' discrimination of a linguistic tone continuum is shown to exhibit order of presentation effects similar to those shown for vowels in previous studies. When presented with two successive syllables differing in fundamental frequency by approximately 4 Hz, listeners were significantly more sensitive to this difference when the first syllable was higher in frequency than the second. However, American English-speaking listeners with no experience listening to Cantonese showed no such contrast effect when tested in the same manner using the same stimuli. Neither English nor Cantonese listeners showed any order of presentation effects in the discrimination of a nonspeech continuum in which tokens had the same fundamental frequencies as the Cantonese speech tokens but had a qualitatively non-speech-like timbre. These results suggest that tone presentation order effects, unlike vowel effects, may be language specific, possibly resulting from the need to compensate for utterance-related pitch declination when evaluating fundamental frequency for tone identification.
[DOI: 10.1121/1.1603231]

PACS numbers: 43.71.Hw [PFA]

I. INTRODUCTION

The discriminability of speech sounds has been shown to depend, in some cases, on the order in which stimuli are presented. For example, Repp, Healey, and Crowder (1979) described the results of experiments in which listeners were asked to judge the similarity of pairs of vowels selected from a continuum ranging from [i] to [I]. For a given pair of stimuli, when the initial vowel was more [i]-like, listeners tended to respond ''same'' significantly more often than when the order was reversed. Similar results were reported by Cowan and Morse (1986), and Macmillan, Braida, and Goldberg (1987). All of these authors suggested that such order effects might result from the decay of the memory trace of the initial token of the pair. According to this explanation, the memory trace of a vowel decays ''toward'' a vowel that is more centrally located in an abstract representation of the talker's vowel space (e.g., [ə]). This model could be termed a ''neutralization hypothesis'' because the memory trace becomes more like a neutral vowel (one that is not particularly high or low, front or back in the vowel space) in a manner similar to the reduction or neutralization of vowel quality in the production of unstressed English vowels (cf. Ladefoged, 2001, pp. 78–79). Thus, along an [i]–[ɪ] continuum, more [i]-like vowels, decaying toward [ə], become more [ɪ]-like, increasing the perceived similarity of the pair. When the initial token of the pair is more [ɪ]-like, its memory trace also decays toward [ə], but in this case that is ''away'' from [i] in the vowel space, leading to a decrease in the perceived similarity of the pair. There are two hypotheses regarding the cause of this observed tendency. One exploits the specific geometry of vowel spaces and one does not draw on linguistic factors at all.

Cowan and Morse (1986) argued that the direction and extent of decay of [i]-like stimuli is determined by the boundaries of the listeners' vowel space. However, their theory does not specify how these boundaries are defined. It is possible that they are determined entirely by the listener's experience, such that those previously encountered tokens with extreme first formant ($F1$) and second formant ($F2$) values determine the boundaries at any given time. Alternatively, these boundaries could be defined in an experience-independent (innate) manner based on the limits of the interaction of articulatory and auditory systems. As discussed by Lieberman and Blumstein (1988, pp. 171–183), the extreme articulatory configurations of [i], [ɑ], and [u] both impart significant acoustic stability to these vowels (cf. quantal theory, Stevens, 1972, 1989) and allow them to delineate the boundaries of the space of possible vowels. For example, the vowel [i] determines the lower bound for the first formant, and the upper bound for the second formant, because it is produced with the narrowest possible oral cavity constriction (for a vowel) and the widest possible pharyngeal cavity

[a)]Some of the material in this article was presented at the 141st meeting of the Acoustical Society of America, Chicago, IL, 7 June 2001.
[b)]Electronic mail: francisa@purdue.edu
[c)]Electronic mail: vciocca@hkusua.hku.hk

opening. No human articulatory configuration could produce a vowel with a greater distance between $F1$ and $F2$ than that of [i]. Thus, a listener's implicit knowledge of the role of articulation in determining acoustic properties of vowels plausibly entails an understanding of the limits of possible formant configurations—the boundaries of the space of possible vowels.

This second description seems more compatible with Cowan and Morse's (1986) model, in which the boundaries of vowel trace expansion appear to function as absolute limits on the extension of the vowel memory trace. According to their model, the memory trace of a vowel is best represented as a region of probability within the vowel space. This region expands over time, representing the gradual degradation of memory acuity. As the memory of a stimulus fades, the probability of accurately recalling its formant pattern or spectral shape also decreases, or, conversely, the probability of recalling incorrect features increases. However, in this model the boundary of a memory trace cannot expand beyond the boundary of the listener's vowel space. In other words, no matter how long a time there is between the presentation of a stimulus and its recall, listeners will not have any probability of recalling the formant frequency values or spectral shape of an unpronounceable vowel. Thus, memory for tokens near a ''point'' vowel (one located close to the intersection of two boundaries) such as [i] should expand disproportionately toward the center of the listener's vowel space. The probability that listeners exposed to a prototypical [i]-like stimulus (one with very low $F1$ and very high $F2$) will remember a more [ə]-like vowel (one with lower $F2$ and higher $F1$) is much greater than that they would remember an even less [ə]-like vowel (one with an even lower $F1$ and an even higher $F2$) because such a more extreme (less neutral) vowel could not have been produced by a human vocal tract. The memory trace for an [i] cannot expand very far in the direction opposite [ə] because in that direction it is already close to the outer bounds of the listener's vowel space. Thus, the first hypothesis can be characterized as proposing that the direction of memory trace decay is a function of the structure of the perceptual space under investigation.

In contrast, Repp and Crowder (1990) argued that the effects described by Cowan and Morse (1986) are a psychophysical consequence of presentation order, and are not specific to memory for speech sounds, let alone language. In a series of experiments, Repp and Crowder (1990) found no consistent evidence that memory for vowels decays in a particular direction. For example, in their experiment 1, pairs consisting of a prototypical [ɛ] (called Pɛ) and a more [ə]-like [ɛ] (called N3ɛ) showed evidence of a decay toward [ə], in that listeners responded ''different'' less often to pairs ordered Pɛ–N3ɛ than to pairs ordered N3ɛ–Pɛ. In contrast, in pairs consisting of prototypical [i] (Pi) and a more [ə]-like [i] (N3i) showed very little evidence of a decay toward [ə], despite there being significant evidence of such effects in other experiments (Cowan and Morse, 1986; Repp *et al.*, 1979). Repp and Crowder's (1990) results suggest that the direction of vowel trace decay may depend on the particular set of stimuli used in a given experiment. According to this hypothesis, vowel contrast effects are a consequence of the gradual replacement of token-specific information with more generic information. That is, memory for exemplars is gradually replaced by information that is more representative of the category to which that exemplar most likely belongs, for example, the representation of the category prototype (cf. Hellström, 1985, and also Huttenlocher, Hedges, and Vevea, 2000). Repp and Crowder (1990) argued that Cowan and Morse's (1986) neutralization hypothesis merely represents a special case of stimuli for which the relevant generic information happens to be similar to the neutral vowel [ə]. However, Repp and Crowder (1990) conceded that their evidence was inconclusive. While some continua showed clear order of presentation effects, others did not, and the authors were unable to determine any systematic factor that might govern the observed pattern of effects. They concluded that the confusing nature of their results may be due in part to the greater acoustic complexity of speech stimuli as compared with the stimuli typically used in demonstrating effects related to the stimulus set (e.g., Braida *et al.*, 1984). Still, this second hypothesis remains plausible: The systematic distortion of memories for speech stimuli over time could result from specific experimental conditions, not as a function of listeners' knowledge of speech or language.

The goal of the present study is to more closely investigate the source of presentation order effects by examining sensitivity to small differences in fundamental frequency (f_0) across speakers of two languages. We examined the perception of stimuli differing only in f_0 by speakers of a tone language (Cantonese, where such differences can be lexically contrastive) and a nontone language (American English, where such f_0 differences are not lexically contrastive). Such stimuli have a number of advantages over vowel stimuli. First, it is possible to generate nonspeech stimuli that are acoustically quite simple (similar to those used in typical psychoacoustic studies of stimulus set-related effects), yet retain the crucial perceptual differences that cue tone category distinctions in speech stimuli. Thus, it is possible to more clearly assess the role of stimulus- or design-related factors. For example, in the examination of the present data the possibility arose that memory traces might, in some cases, decay in a unidirectional manner, regardless of category prototype location or the set of stimuli presented. According to this kind of formulation, the results presented by Cowan and Morse (1986) could be described as a decay of vowel memory traces toward the right side of the vowel space (toward a lower $F2$ value). A second advantage of using lexical tone-based stimuli is that monolingual speakers of American English do not possess a linguistically structured knowledge of pitch differences between syllables (and therefore do not have a linguistically structured ''tone space'' analogous to vowel space), while Cantonese speakers, and indeed, tone language speakers in general, do (cf. Gandour, 1981; Gandour and Harshman, 1978). By comparing the perception of small, barely suprathreshold f_0 differences by speakers of these two languages, it may be possible to determine whether contrast effects are purely a nonlinguistic consequence of the experimental stimulus set, as hypothesized by Repp and Crowder (1990), or rather a consequence of

TABLE I. Fundamental frequency values for stimuli for all experiments.

Stimulus	f_0 in Hz (mel)	Tone class
1	100.0 (150.5)	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

some aspect of linguistic knowledge, as implied by Cowan and Morse (1986).

II. EXPERIMENT 1

The first experiment examined native Cantonese speakers' sensitivity to small suprathreshold frequency differences in synthesized Cantonese syllables ranging in f_0 along a lexical tone continuum from the frequency of a low-level tone to that of a high-level tone (see Bauer and Benedict, 1997 for a thorough discussion of Cantonese tonal phonology). In particular, this experiment was designed to investigate whether Cantonese listeners showed a difference in sensitivity due to the order of presentation of the syllables in pairs that differed by about 4 Hz.

A. Methods

1. Subjects

Fifteen native speakers of Cantonese (12 women, three men) reporting no history of speaking or hearing disability participated in this experiment. Eight were undergraduate speech pathology students in the Department of Speech and Hearing Sciences at the University of Hong Kong, and seven were students and employees from other departments. All participated in the experiment on a voluntary basis.

2. Stimuli

Stimuli for this experiment consisted of a continuum of ten 300-ms syllables synthesized with the parallel branch of a Klatt-style formant synthesizer (Klatt and Klatt, 1990) called SenSyn (Sensimetrics Corp.), implemented on a PowerMac G3. All stimuli were modeled on real Cantonese words differing only in tone (low level, corresponding to token 1, mid level, corresponding to token 4 or 5, and high level, corresponding to token 10). These words were all segmentally [ji] according to standard IPA transcription (cf. IPA, 1999). All stimuli had level f_0 contours and differed in frequency in perceptually equal steps (6.1 mel, approximately 4.4–4.5 Hz). Exact frequency values are given in Table I; each of these values was used as the fundamental frequency for the entire duration of a single stimulus syllable. Selected synthesis parameters for token 1 are given in the Appendix. Subsequent tokens differed only in f_0, as shown in Table I.

On each trial a pair of stimuli was presented with a 250-ms interstimulus interval (ISI). All pairwise combinations of syllables separated by 0 or 1 token along the continuum were presented (total of 28 pairs), including ten identical pairs (1–1, 2–2, 3–3, etc.) and 18 adjacent pairs (1–2, 2–1, 2–3, 3–2, etc.). Stimuli 1, 4, 5, and 10 were identified as satisfactory exemplars of real Cantonese words comparable to those produced by the native speaker on whose productions these stimuli were modeled. Stimulus 1 was identified as the word /ji22/ ''two'' (low level tone), stimulus 4 and 5 were identified as the word /ji33/ ''spaghetti'' (mid level tone), and stimulus 10 was identified as the word /ji55/ ''clothing'' (high level tone). Note that tones are indicated numerically, according to a commonly accepted five-point scale where 1 indicates the lowest pitch of a talker's pitch range, and 5 the highest. Two numbers are used to indicate the starting and ending pitch of the syllable. To our knowledge there is currently no published information regarding the lexical frequency or familiarity of these words in spoken Cantonese. Thus, we cannot speculate as to whether our results might have been affected by these factors. However, it may be noted that these words were selected in part because they are easily recognized and understood by children (cf. Ciocca and Lui, 2003).

3. Procedure

Participants were seated in a single-wall IAC sound booth looking through a window at the monitor of an Apple Power Macintosh 7100/AV computer located outside the sound booth. Stimuli were presented via Sennheiser HD-545 headphones at a comfortable listening level (73-dBA peak level for target words). Stimulus presentation and response collection was controlled by a Hypercard (Apple Computers, Inc.) stack running on the computer. Sounds were output through an Audiomedia II sound card at a sampling rate of 44.1 kHz. Listeners participated in two tasks, an identification task and a discrimination task. The order of the two tasks was counterbalanced across subjects (eight participants completed the discrimination task first, followed by identification, while seven completed the experiment in the reverse order). For this article, only the results of the discrimination task will be considered (Identification results are discussed in detail by Francis, Ciocca, and Ng, to appear).

The discrimination task consisted of 11 blocks, each with 28 trials. The first block was treated as familiarization and not scored, though listeners were not aware of this at the time of testing. Each trial began with the presentation of a visual warning signal on the computer screen. Subsequently, listeners heard a warning tone (an amplitude-modulated complex tone with fundamental frequency, harmonic structure, and amplitude modulation significantly different from speech) followed after 500 ms by the presentation of a single pair of syllables separated by 250-ms ISI. Following the presentation of a stimulus pair, listeners were presented with two buttons arranged horizontally on the screen, labeled *same* and *different*. Response buttons were always presented in this order. Participants were instructed to click on one of the buttons to indicate whether the syllables they heard were the same or different. After selection, followed by a brief pause, the next trial began. Order of stimulus presentation within blocks was random. Responses were collected auto-

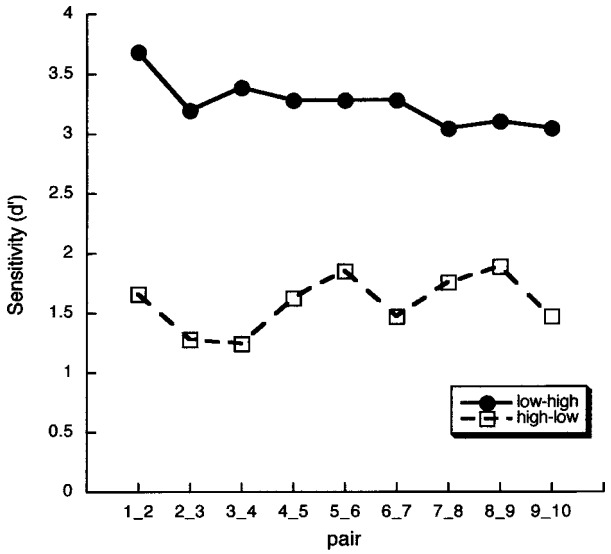

FIG. 1. Cantonese listeners' group sensitivity (d') calculated according to the roving methods (differencing model) described by Macmillan and Creelman (1991) for both orders of presentation of pairs of adjacent [ji] syllables along a continuum ranging in f_0 from 100 Hz (token 1) to 140 Hz (token 10) in perceptually equal steps (6.11 mel, approximately 4.5 Hz). Low–high order (solid line) indicates pairs in which the token with a lower f_0 is first. High–low (dotted line) indicates pairs in which the token with a higher f_0 is presented first.

matically, and scored according to whether responses were correct or not. Participants received no feedback on their responses.

B. Results

In this study we presented each stimulus pair ten times (not counting the first unscored block of trials). Group d' values based on the mean hit- and false-alarm rates of all subjects were calculated using the roving (differencing model) methods discussed by Macmillan and Creelman (1991, pp. 147–152, using Table A5.4, pp. 338–354, adapted from Kaplan, Macmillan, and Creelman, 1978). Macmillan and Creelman's Table A5.4 was generated by systematically varying the value of k (the response threshold) in the equations shown in Eq. (1) (where Φ is the normal distribution function), to arrive at a d' value for each possible H(it) and F(alse alarm) pair[1]

$$H = \Phi\left[\frac{(-k+d')}{\sqrt{2}}\right] + \left[\frac{(-k-d')}{\sqrt{2}}\right],$$
$$F = 2\Phi(-k\sqrt{2}). \tag{1}$$

Statistical analyses were carried out on the quantity (hit rate minus false alarm rate, or H-F) which served as an approximation of d' (Maddox and Estes, 1997).[2] Overall, listeners scored above chance (50%) across the continuum, ranging from 70% to 74% correct. This rate of accuracy may overestimate discrimination sensitivity because listeners appeared to be strongly biased toward responding "same" and exhibited a mean false-alarm rate of only 12% across all listeners and all stimuli.

Group d' for each stimulus pair in each order is shown in Fig. 1, where low–high indicates a pair in which the first token has a lower f_0 than the second (e.g., pair 1–2), while high–low indicates a pair in which the first token is higher in fundamental frequency than the second (e.g., pair 2–1). A two-way, repeated measures ANOVA on the difference between hit rate and false-alarm rate (H − F) showed a main effect of order of presentation (low–high mean = 0.65, high–low mean = 0.22), $F(1,14) = 19.88$, $p = 0.001$, but not of stimulus pair, $F(8,112) = 1.26$, $p = 0.27$. There was also a significant interaction between the two factors, $F(8,112) = 2.66$, $p = 0.01$. However, *post hoc* (Tukey HSD) analysis showed a significant difference ($p < 0.001$) between every one of the low–high points and every one of the high–low points, while none of the within-order pairwise comparisons reached significance ($p > 0.05$ for all). Similarly, a one-way ANOVA of the differences between the low–high and high–low scores at each pair showed a significant effect of pair, $F(8,112) = 2.66$, $p = 0.01$, but the only comparison to reach significance at the $\alpha = 0.05$ level (by Tukey HSD) involved pair 3–4, where the difference was significantly greater than that between pair 8–9.[3] Thus, the appearance of a greater overall difference at pairs 3–4 and 6–7 (near expected category boundaries) is not supported statistically.

C. Discussion

Cantonese listeners were, on the whole, more sensitive to small f_0 differences in speech stimuli when the first token in a pair had a lower f_0 (low–high order) than when the first token had a higher f_0 (high–low order). These results are consistent with the hypothesis that memory for pitch decays downward (in pitch) over time, such that pairs in the low–high order become increasingly distinct over time, while pairs in the high–low order become more similar (at least over the 250-ms ISI used in the present experiment). However, these results do not provide strong evidence for identifying the source of such a directional memory trace decay. In order to determine whether this asymmetry in sensitivity to pitch differences is related to properties of the stimuli or of the experimental procedure, as opposed to being due to the linguistic experience of the listeners, we examined the perception of listeners who had no experience making pitch-based phonological distinctions.

III. EXPERIMENT 2

Existing research on the consequences of memory trace decay suggests that biases in stimulus recall may arise from factors specific to either particular category structures (e.g., prototypes, Huttenlocher, Hedges, and Duncan, 1991), particular perceptual spaces (e.g., the geometry of vowel spaces, Cowan and Morse, 1986), or the content of particular experimental stimulus sets (Repp and Crowder, 1990). While evidence presented by Polka and Bohn (1996) suggests that language-specific vowel category prototypes are not likely to play a detectable role in determining the biased recall of vowels (at least by infants), the other two possibilities are still equally plausible. Indeed, it is even possible that tones, unlike vowels, may be influenced by language-specific category prototypes. However, the most obvious theoretical distinction is between auditory (or nonlinguistic) and linguistic

sources of contrast effects. Listeners with no experience hearing a tone language do not possess a linguistically structured ''pitch space,'' nor do they have any exclusively pitch-based phonetic categories. If these listeners show the same asymmetric pattern of discrimination as Cantonese speaking listeners, then we may conclude that this asymmetry results from the interaction of stimulus properties and human auditory capabilities. If, on the other hand, there are noticeable differences between the response patterns of the English- and Cantonese-speaking listeners, then we may conclude that these order of presentation effects are related to listeners' linguistic knowledge or experience (whether in the form of a linguistically structured perceptual space, or language-specific category prototypes).

A. Methods

1. Subjects

Nine people (five men and four women) participated in this experiment. All were native speakers of North American dialects of English. Five participants were undergraduate and graduate students and alumni from the University of Chicago, while four were newly arrived faculty members and visitors to the University of Hong Kong who had been in Hong Kong for less than a month. All of the participants reported having no knowledge of Cantonese or other tone language.

2. Stimuli

All stimuli were identical to those used in experiment 1.

3. Procedures

All procedures were identical to those described in experiment 1 including the counterbalanced participation in an additional identification experiment not reported here. However, the present experiment was run on a Macintosh iBook and stimuli were presented via Sennheiser HD-570 (three participants) or HD-580 (six participants) headphones in a quiet room. No warning tone was provided prior to stimulus presentation.[4] Stimuli were played at a comfortable listening level (approximately 75-dBA peak level for target words).

B. Results

The mean percent-correct score for American English-speaking participants ranged from 66% to 76% across the continuum, similar to Cantonese listeners. Again, American English listeners were strongly biased toward ''same'' responses, with a false-alarm rate of just 6%. As in experiment 1, group d' was calculated for each stimulus pair in each order (Macmillan and Creelman, 1991, pp. 147–151), shown in Fig. 2. For statistical comparison of the two speaker groups, both Cantonese and American English listeners' sensitivity (d') to each pair regardless of order of presentation was calculated (based on hit rate and false-alarm rate for each pair ignoring differences in order of presentation). American English listeners' mean d' across the continuum was 2.60, compared with 2.53 for Cantonese speakers, and this difference was not significant, $t(22)=0.11$, $p=0.92$. For the American listeners, a two-way repeated measures

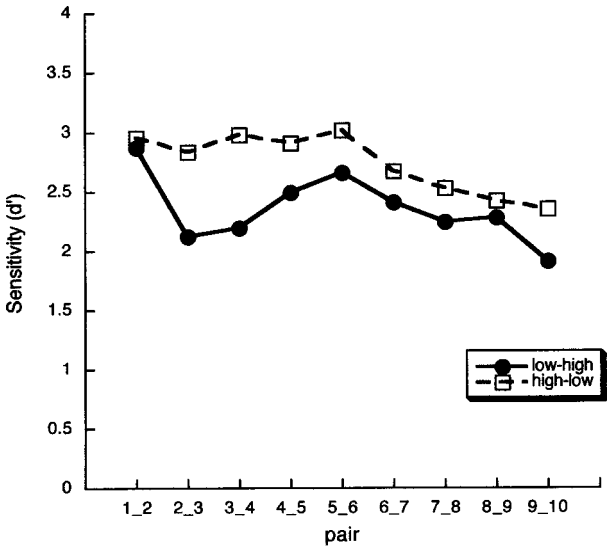

FIG. 2. American English listeners' group sensitivity (d') (calculated as for Fig. 1) for both orders of presentation of pairs of adjacent [ji] syllables along a continuum ranging in f_0 from 100 Hz (token 1) to 140 Hz (token 10) in perceptually equal steps (6.11 mel, approximately 4.5 Hz). Low–high order (solid line) indicates pairs in which the token with a lower f_0 is first. High–low (dotted line) indicates pairs in which the token with a higher f_0 is presented first.

ANOVA [calculated using (H-F) values for each order separately], showed no effect of order of presentation, $F(1,8)=3.13$, $p=0.11$, or stimulus pair, $F(8,64)=1.72$, $p=0.11$, and no interaction between the two factors, $F(8,64)=1.17$, $p=0.33$.

C. Discussion

The results of experiment 2 suggest that monolingual American English-speaking listeners are as sensitive as native speakers of Cantonese to small (subcategorical) differences in fundamental frequency. However, unlike Cantonese listeners, American listeners showed no evidence of a contrast effect in pitch discrimination. That is, there is no evidence that American English speakers are differentially sensitive to pitch differences depending on the order of presentation (low–high versus high–low). This pattern of results, when contrasted with the results of experiment 1, suggests that the source of contrast effects in Cantonese speakers' perception of tone is language specific. In this case, order-of-presentation effects are not a purely psychophysical consequence of a particular set of stimuli or experimental procedure. Experience with perceiving and speaking Cantonese apparently leads to differences in the way listeners store and/or retrieve memory traces of the pitch of auditory stimuli as compared with listeners without such experience.

The observation that speaking a tone language affects pitch perception is not new. Stagray and Downs (1993) demonstrated that speakers of a tone language (Mandarin Chinese) were less sensitive to differences between pure tones around 1000 Hz than were speakers of a nontone language (English). Stagray and Downs (1993) argued that their results can be best accounted for in terms of the categorical perception of tones. Because tone language speakers are used to making categorical decisions based on pitch, Mandarin

speakers exhibit the decreased within-category sensitivity characteristic of categorical perception of segments (cf. Macmillan, 1987 for discussion of categorical perception in signal detection-theoretic terms). However, Stagray and Down's (1993) results go beyond the usual claims of studies of categorical perception, in that they suggest that experience categorizing speech can affect sensitivity to differences between nonspeech sounds. One implication of this claim is that some kinds of speech experience may affect the function of basic auditory processes. Similar claims that linguistic experience can affect ''preattentive'' aspects of auditory processing have recently been advanced (Allen, Kraus, and Bradlow, 2000; Sharma and Dorman, 2000; Trembley *et al.*, 1997).

IV. EXPERIMENT 3

One way to demonstrate that experience with a tone language can affect basic (pitch-processing) properties of the auditory system would be to show a discrepancy in contrast effects between Cantonese and English speakers using stimuli that are not speech-like. Stagray and Downs (1993) argued that categorical perception of lexical tones was reflected in their listeners' performance on a pure-tone pitch discrimination task. If we were to observe pitch contrast effects in the processing of *nonspeech* sounds by Cantonese speakers, and if we fail to observe these contrast effects in an English-speaking population, then we may conclude that these contrast effects must result from differences in basic auditory processes related to differences in linguistic experience. This experiment was designed to compare the performance of Cantonese and English listeners on a nonspeech task equivalent to the first and second experiments.

A. Methods

1. Subjects

Two groups of listeners participated in this experiment, one Cantonese speaking, the other American English speaking. The first group consisted of nine female native speakers of Cantonese from the University of Hong Kong community, none of whom had participated in experiment 1. The second group consisted of seven native speakers of American English (four men, three women), all undergraduate or graduate students at the University of Chicago, of whom three had participated in experiment 2 two days prior to the present experiment. All participants reported normal hearing, and one Cantonese participant reported having perfect pitch.

2. Stimuli

Stimuli consisted of ten complex tones modeled on the synthetic speech stimuli used in experiments 1 and 2 and synthesized using the PowerSynthesiser application (Russell and Darwin, 1991). All stimuli were 300 ms long, consisting of eight equal-amplitude harmonics (harmonics 1, 3, 5, 6, 7, 8, 9, and 11). Harmonics 2, 4, and 10 were omitted to make the sound clearly less speech-like. Each stimulus had amplitude rise and decay times of 5 ms. The only difference between the ten stimuli was the fundamental frequency, which varied along the identical continuum as the stimuli in the first two experiments (see Table I). Complex tones were synthesized at a sampling rate of 44.1 kHz on a Power Macintosh 7100/AV equipped with an Audiomedia II sound card.

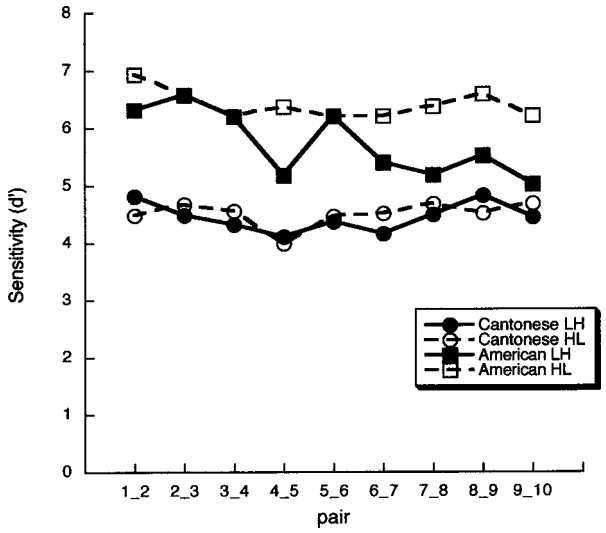

FIG. 3. Group sensitivity (d') for Cantonese listeners (circles) and American English listeners (squares) (calculated as for Figs. 1 and 2) for both orders of presentation of pairs of adjacent nonspeech tone complexes along a continuum ranging in f_0 from 100 Hz (token 1) to 140 Hz (token 10) in perceptually equal steps (6.11 mel, approximately 4.5 Hz). Low–high (solid lines) indicates pairs in which the token with a lower f_0 is first. High–low (dotted lines) indicates pairs in which the token with a higher f_0 is presented first.

3. Procedure

Procedures were identical to experiment 2 for all listeners, except that all American English listeners used Sennheiser HD-570 headphones, and all Cantonese listeners used Sennheiser HD-545 headphones and the sound presentation hardware from experiment 1.

B. Results

Group d' values were calculated for each language group based on each order of presentation, as shown in Fig. 3. In order to derive a statistical test of differences between the two groups, each listener's sensitivity (d') regardless of order of presentation was calculated for each pair (based on hit rate and false-alarm rate for each pair, ignoring differences in order of presentation) using the roving calculations described by Macmillan and Creelman (1991). Regardless of order of presentation, American English listeners' mean d' across the continuum was 5.59, compared with 4.56 for Cantonese speakers. Note that these measures of mean sensitivity were noticeably higher than those measured for speech stimuli in experiments 1 and 2 (2.60 for American listeners in experiment 2, and 2.53 for Cantonese listeners in experiment 1). A three-way mixed factorial ANOVA (between group factor of language, and repeated measures factors of pair and order) was calculated using H–F, the hit rate minus false-alarm rate statistic used in the previous two experiments. There was a significant main effect of language, $F(1,14)=6.75$, $p=0.02$, showing that American English lis-

teners were more sensitive to f_0 differences than Cantonese listeners. No other main effects or interactions were significant.

C. Discussion

Both English- and Cantonese-speaking listeners in experiment 3 were considerably more sensitive to frequency differences between these nonspeech complex tones as compared with the sensitivity of comparable groups of listeners to equivalent frequency differences in synthesized speech stimuli. This may reflect differences in the complexity of the stimuli. However, Flanagan and Saslow (1958) found difference limens (DLs) for f_0 differences between synthetic vowels (about 0.63 Hz) to be slightly *smaller* than those identified by Harris (1952) for pure tones (about 0.75 Hz) (see Klatt, 1973, for discussion).[5] In other words, increasing the acoustic complexity of stimuli does not necessarily increase the difficulty of perceiving differences in the pitch of those stimuli. Note also that DLs reported for both speech and pure tones are considerably smaller than the differences between stimuli used in the present experiments. Aside from this overall greater sensitivity to frequency differences between nonspeech stimuli, neither American English-speaking nor Cantonese-speaking listeners showed any discernible effect of order of presentation in their discrimination of nonspeech complex tones. This suggests that, whatever the source of the contrast effects observed in experiment 1, experience with Cantonese does not affect the perception of the pitch of complex tones in the same way that it affects pitch-based discrimination of speech sounds.

One final observation in experiment 3 remains puzzling. In experiment 3, Cantonese listeners exhibited an overall lower sensitivity to pitch differences (in either order) when compared with that of American English listeners. These results show some support for observations made by Stagray and Downs (1993), who found that speakers of Mandarin Chinese (also a lexical tone language, but with a different tonal inventory from Cantonese) showed significantly larger frequency difference limens than did native speakers of English on a pure-tone discrimination task. Stagray and Downs (1983) attributed this difference to Mandarin speakers' categorical perception of tone. They argued that the frequency differences between the stimulus pairs used in their experiment were always well within a single category, within which acuity should be lower than across category boundaries according to standard theories of categorical perception [e.g., Liberman *et al.* (1957)]. However, there are two problems with using categorical perception to account for the results of experiment 3. First, although there was a significant difference between the sensitivities of the two language groups on the nonspeech complex tone discrimination task (experiment 3), the difference between their discrimination sensitivity on the speech task (experiment 1 versus experiment 2) was not significant. If Cantonese speakers' poorer discrimination of nonspeech tones was due to their greater experience with making pitch-based category judgments of speech stimuli, then there should be a similar, if not greater, difference in sensitivity between the two language groups when judging speech-like stimuli. However, the two groups were not significantly different in their overall sensitivity to f_0 differences between speech stimuli. They only showed a difference in sensitivity to nonspeech sounds. Second, the frequency range employed in these experiments was selected to encompass the three level tone categories of Cantonese. In particular, token pairs 3–4 and 6–7 explicitly span the category boundaries. Although the results of experiment 3 contribute to the mounting evidence that tone language speakers tend to be less sensitive than English speakers to fundamental frequency differences between nonspeech sounds (Stagray and Downs, 1993; Tanner and Rivette, 1964, though see Burns and Sampat, 1980 for a counterexample to this trend), there appears thus far to be little evidence that this tendency is related to the categorical perception of tone as typically specified.

V. GENERAL DISCUSSION

The results of experiment 1 suggest that Cantonese listeners' ability to discriminate between level fundamental frequency contours is strongly influenced by the order in which pairs of stimuli are presented. When the first token in a pair is lower in f_0 than the second (low–high order), listeners are considerably more sensitive to the difference than when the order of pairs is reversed (high–low order). One way to account for this difference is in terms of a gradual decay of the memory trace of the initial token, such that it is recalled as having a lower pitch for the purposes of comparison with the later-occurring token. Although the results of experiment 1 suggest that such a memory trace decay would be directional (toward lower values), we found no strong evidence to suggest that stimulus set related properties could have contributed to the appearance of these effects. If the direction of memory trace decay were the result of properties of the stimulus set, we would expect the directionality to be symmetrical, either with relation to the edges of the continuum or with respect to some more centrally located region along the continuum. With respect to the role of edges in the directionality of memory trace decay, Macmillan, Braida, and Goldberg (1987) characterize Berliner *et al.*'s (1977) *bias edge effect* in terms of the boundaries (edges) of the stimulus continuum. Berliner *et al.* (1977) found that when listeners were presented with two tokens of overall low intensity (close to the low-intensity end of the continuum) they tended to hear the first token as louder than the second. But, when the same intensity difference was presented using high-intensity stimuli (at the high-intensity end of the continuum), listeners tended to hear the first token as quieter than the second. Thus, the location of the stimulus pair along the continuum affected the direction of the bias—the first token of a quiet pair was heard as louder while the first token of a loud pair was heard as quieter, suggesting that in both cases the memory trace of the first token decayed toward a more intermediate (centrally located) value along the stimulus continuum.

In the present case, we do observe something like a bias edge effect at the right (higher frequency) end of the continuum, in that the first token in a higher-frequency pair (e.g., 8–9 or 9–8) is generally heard as having a higher pitch (such that 8–9 is treated more like 9–9, a ''same'' pair by being

only poorly discriminable, while 9–8 is still quite easily discriminated). This pattern can clearly be characterized as a decay of the memory for pitch of the first token toward a lower (more central) value along the continuum. However, there is no corresponding upward decay of the memory of the pitch of the first token in a lower frequency pair (e.g., 2–3). In this case we would expect to observe a tendency for the first token to be remembered as *higher* in pitch (e.g., for pair 2–3 to be poorly discriminated), but this is not what was observed. Instead, we see a tendency for the first token to be treated as *always* having a lower pitch than the second, regardless of where along the continuum the two tokens are located. Similarly, there does not seem to be any location along the continuum toward which, or away from which, memory traces seem to decay. The trend is always in a downward direction across the entire continuum. Thus, although there is clear evidence that Cantonese listeners' contrast effects can be described in terms of a general decay of memory for pitch toward lower values, there is no explicit evidence that this effect is due to properties of the stimulus set itself.

The results of experiment 2 further support the hypothesis that asymmetric discrimination of f_0 differences in synthetic speech stimuli by Cantonese listeners is not due to some property of the stimulus set, but rather is related to listeners' knowledge of Cantonese, a lexical tone language. In experiment 2, American English-speaking listeners with no experience with any lexical tone language showed no evidence of any order of presentation effects when tested with the same stimuli and procedures as in experiment 1. It is important to note that the American listeners of experiment 2 were (in both orders of presentation) about as sensitive to the f_0 differences between these stimuli as were the Cantonese listeners of experiment 1. The results of experiment 1 and 2, taken together, suggest that the contrast effects shown by the Cantonese listeners are a consequence of their knowledge of Cantonese.

The results of experiment 3 suggest that, whatever the specific mechanism that causes contrast effects in Cantonese listeners' perception of spoken pitch, it does not appear to have affected their ability to discriminate the pitch of nonspeech sounds. Cantonese listeners, like American English listeners, showed no asymmetry in sensitivity to f_0 differences between nonspeech stimuli that were identical in f_0 to the speech stimuli used in experiments 1 and 2. These results provide further support that the order of presentation effects demonstrated by Cantonese listeners are a consequence of their linguistic processing of speech stimuli.

A. The role of language experience in memory trace decay

One way to account for the difference in the effects of order of presentation for speakers of tone vs nontone languages might be to conclude that American listeners, lacking experience with pitch-based lexical distinctions, do not have a mental representation of a ''tone space.'' As a result, American English listeners' memories for words do not decay in a manner that is affected by the boundaries of such a space. Cantonese listeners, on the other hand, can be described as basing their perceptual judgments on relative relations between the mental representations of tokens in a tone space (cf. Gandour, 1981). Following the model proposed by Cowan and Morse (1986), we might expect the boundaries of Cantonese listeners' tone space to cause an asymmetric expansion of the memory trace for the pitch of the earlier-presented syllable away from the nearer boundary of the tone space (in the same way that memory for [i]-like stimuli is proposed to expand disproportionately away from the high-front corner of a listener's vowel space). There is, however, one significant problem with such an account. The stimuli used in experiments 1 and 2 range in frequency from a prototypical low-level tone to a prototypical high-level tone, encompassing the vast majority of the normal spoken frequency range of the speaker on whose productions they are modeled. Any explanation of order of presentation effects based on the influence of the boundaries of tone space on memory traces would predict opposing effects at either end of a frequency continuum spanning that space. That is, tokens at the high end of the continuum should decay toward a lower pitch as their memory trace, located near the top end of the space, cannot expand very far in a higher direction. Conversely, tokens at the low end of the continuum should decay toward a higher pitch. Contrary to this hypothesis, however, all tokens along the continuum, from lowest to highest, show an asymmetry in discriminability between the low–high and high–low orders, such that a decay-based account would have to conclude that the memory trace of every token appears to decay toward a lower pitch value.

Another possibility is that listeners' memory for pitch might decay, or be biased, toward tone-category prototypes (or away from category boundaries) as suggested by Huttenlocher and her colleagues (Huttenlocher *et al.*, 1991; Huttenlocher *et al.*, 2000). However, on a tone identification task using the same stimuli (Francis *et al.*, to appear) Cantonese listeners showed category boundaries between tokens 3 and 4 (between the categories low level and midlevel) and 7 and 8 (between the categories midlevel and high level).[6] Thus, if memory for pitch decayed toward category prototypes or away from category boundaries, we would expect listeners' sensitivity to pairs spanning these boundaries to be relatively high regardless of order of presentation. For example, if the 3–4 pair were presented in the low–high order (first token 3, then token 4), then the memory for 3 should decay toward the prototype for the low-level category (somewhere around token 1). Discrimination should be accurate because the perceived difference between the two syllables should increase due to category-related bias. If this pair were presented in high–low order (first token 4, then token 3), memory for token 4 should decay toward the prototype for the midlevel category (somewhere around token 5), similarly improving discrimination. However, discrimination was considerably better for the low–high order for this, and most other, pairs of tokens (including pair 7–8, the other cross-boundary pair). Thus, the results of experiment 1 suggest that neither the geometry of tonal space in general, nor listeners' language-specific phonological inventory, play a primary role in determining the directionality of the observed contrast effects. None of the results of experiment 1 provides explicit support for a memory trace decay model. There is no evi-

dence that the order of presentation effects shown in experiment 1 result from basic psychoacoustic properties of the stimulus or task. In contrast, the results of experiments 2 and 3 demonstrate that order of presentation effects in level tone perception occur only when native speakers of Cantonese are listening to speech.

B. Discrimination asymmetries and f_0 declination

The mechanism of memory trace decay could account for the present results if we arbitrarily assume that memories for pitch decay toward lower pitches. However, since there is, at present, no corroborating evidence to support this assumption, it may prove useful to investigate other mechanisms that could account for these data. One aspect of Cantonese speech production that may prove useful in this regard is the phenomenon of f_0 declination or downdrift, the gradual declination of fundamental frequency over the course of an utterance (Ohala, 1978; Pierrehumbert, 1979; Umeda, 1982; Vaissière, 1995; Vance, 1976; Wong, 1999). Downdrift has been argued to be a universal, language-independent (and even cross-species) characteristic of speech production (Hauser and Fowler, 1992; Ohala, 1978). In English, short declarative utterances without prominent focal stress, such as those typically elicited in laboratory speech experiments, tend to exhibit clear declination of f_0 over time (Umeda, 1982). Furthermore, American English listeners show evidence of compensating for downdrift in the perception of the relative pitch of early- and late-occurring syllables in sentences. For example, Pierrehumbert (1979) showed that listeners perceived a syllable occurring early in an utterance as having equal pitch with a later-occurring syllable even though the later syllable had a lower f_0. This was interpreted as evidence that listeners were able to compensate perceptually for an expected declination in pitch over the course of the utterance.

There is evidence that Cantonese speakers also exhibit f_0 declination in speech production (Vance, 1976), and some suggestion that Cantonese listeners perceptually compensate for this expected declination. For example, Wong (1999) presented listeners with a target word preceded by a context sentence in which the f_0 had been manipulated in one of two ways. The sentence was divided in half and the f_0 of either the earlier-occurring portion or the later-occurring portion of the context sentence was shifted. Results suggested that Cantonese listeners based their judgments of the tone of the target word on the pitch of more recent (later-occurring) pitch information in the sentence. For example, if the f_0 of the second half of the sentence was shifted down, listeners responded as if the target word had a high level tone. When the f_0 of the second half of the sentence was shifted upward, listeners tended to respond as if the word had a low-level tone, although in both cases the f_0 of the target world remained the same. However, when the f_0 of the second half of the sentence was shifted upward, the expectation of downdrift was violated. In this case listeners did not respond as strongly according to the more recent (second half) pitch information. That is, they made fewer than expected identifications of the target syllable as having a low-level tone. Wong (1999) argued that listeners responded less strongly in this condition because they were confused by the violation of their expectations for downdrift.

It is possible to account for the results found in experiment 1 in terms of a compensation for a learned expectation that utterances will tend to exhibit a slight declination in f_0 from beginning to end. In order to correctly identify the tone of a syllable, Cantonese listeners must be able to take into account the position of that syllable within the utterance. Syllables at the end of the sentence must be perceptually raised in pitch, otherwise they risk being identified as having a lower tone than the speaker intended. In the case of the high–low order of presentation in experiment 1, this raising of the perceived pitch results in the two tokens sounding similar or identical. In contrast, in the low–high presentation order this raising results in a heightened perception of the difference. This hypothesis is supported by the observation that, on average, Cantonese listeners were slightly more sensitive than American listeners to pairs presented in the low–high order, but less sensitive than American listeners to pairs presented in the high–low order.

One problem remains. American English listeners in the present experiment showed no evidence of an expectation that pitch should fall over the course of an utterance, although such expectations have previously been demonstrated in English listeners' judgments of the pitch of syllables in sentential context (Pierrehumbert, 1979). It is possible that American English listeners did not treat the syllable pairs used in this experiment as a single utterance. One possible reason for this is that the syllable [ji] is not an English word. However, Pierrehumbert (1979) achieved her results using nonsense sentences made up of repetitions of the syllable [ma]. A more likely explanation is based on the observation that English is a stress-timed language (Laver, 1994, p. 157), while Cantonese is typically considered to be syllable-timed (Bauer and Benedict, 1997, p. 316). Thus, American English listeners presumably expect utterances to exhibit a pattern of more or less alternating stressed (louder, higher pitched, longer) and unstressed (quieter, lower pitched, shorter) syllables.

According to this explanation, there are two reasons that speakers of a syllable-timed language would not show perceptual compensation for downdrift in the stimuli used here. First, American listeners might not have treated the two syllables in experiment 2 as a single utterance, perhaps because both syllables were equally loud. Alternatively, they may have treated it as a single utterance consisting of a single word, perhaps because the alternation in pitch suggested the presence of one stressed and one unstressed syllable. In the first case listeners might have expected a reset of the pattern of pitch declination with the start of the second utterance. The resetting of pitch declination at utterance breaks is a common phenomenon according to Pierrehumbert (1979), and might plausibly enable listeners to accurately distinguish small f_0 differences between the two syllables because both syllables are located at the start of the expected declination curve. In the second case, it is possible that American English speakers only compensate for pitch declination between *stressed* syllables. Since stressed syllables are typically sepa-

rated by at least one unstressed syllable (as in Pierrehumbert's experiments), American English listeners might retain the ability to distinguish small f_0 differences between adjacent syllables. Indeed, such an ability might contribute to the ability to distinguish between stressed and unstressed syllables. Further research on the perception of pitch, and stress, in American English utterances would be necessary to distinguish between these two possibilities.

ACKNOWLEDGMENTS

Some of the results presented in experiment 1 were first noted in a dissertation submitted by Brenda Ng in partial fulfillment of the requirements for the B.Sc. degree in Speech and Hearing Sciences at the University of Hong Kong. We are grateful to Howard Nusbaum and Kimberly Fenn at the University of Chicago for their assistance with running experiments 2 and 3, and to Neil Macmillan for insightful discussions concerning signal detection theory. Some of the results discussed here were first presented at the 141st meeting of the Acoustical Society of America, Chicago, IL, 7 June 2001. This research was conducted while the first author was a postdoctoral fellow at the University of Hong Kong, in the Department of Speech and Hearing Sciences, and was supported in part by the Hong Kong Research Grants Council, HKU 7193/00H, to Valter Ciocca.

APPENDIX: SYNTHESIS PARAMETERS

Stimulus synthesis used parameters measured at 1-ms intervals from a natural speech sample. After initial synthesis, parameter values were subsequently smoothed by hand, resulting in roughly linear contours for major frequency and amplitude parameter values. Stimuli were synthesized using an update interval of 5 ms. Parameter AV (amplitude of voicing) rose from a value of 50 to 60 dB over the first 210 ms of the utterance, and then declined to a value of 42 dB at the end of the utterance. $F1$ rose from 360 to 368 Hz over the first 100 ms and remained level for the remainder of the utterance. $F2$ rose from 2308 to 2392 Hz over the first 100 ms, and then fell to 2193 Hz by the end of the utterance. $F3$ began at 3712 Hz and fell to 3574 Hz in the first 100 ms, then fell more gradually to 2929 Hz at the end of the utterance. $F4$ began at 4126 Hz, and fell to 3826 Hz by 195 ms, then to 3620 Hz at the end of the utterance. $F5$ began at 4586 Hz, and fell to 4279 Hz by the end of the utterance. A1V (amplitude of $F1$) began at 55 dB and rose to 59 dB by 170 ms, remained level until 220 ms, then fell to 56 dB at 260 ms, and then fell more sharply to 42 dB by the end of the utterance. AV2, AV3, and AV4 all began at 46 dB. AV2 rose to 57 dB at 140 ms, then fell to 56 dB at 160 ms, remained level to 205 ms, fell to 55 dB by 210 ms, remained level until 275 ms, then fell to 42 dB by the end of the utterance. AV3 rose to 57 dB at 145 ms, then starting at 260 ms fell to 42 dB by the end of the utterance. AV4 rose to 55 dB at 145 ms, then remained level until 260 ms, at which point it fell to 42 dB by the end of the utterance.

[1]In this experiment we are interested in separately analyzing pairs of stimuli with different orders of presentation. Therefore, there are half as many trials for each "different" pair *in each order* (10) as there are "same" trials (20). Because there is a maximum of ten possible hits for a given order of presentation, but there are up to 20 possible false alarms, correcting for perfect scores according to the methods proposed by Macmillan and Creelman (1991, p. 10) would lead to different z scores for a hit rate of 1.0 and a false-alarm rate of 1.0 (and similarly for scores of 0.0). This is not a problem for analyzing group data, since there were no perfect scores on any pair. However, some individual subjects did score either perfect hit rates (1.0) or perfect false-alarm rates (0.0) on one or another stimulus pair. Therefore, individual d' scores were not calculated.

[2]Because this is not a commonly used statistic, all analysesusing the H−F statistic were also repeated using arcsine-transformed percent correct [P(C)] values. Unless otherwise noted, tests reported as significant based on the H−F score were also found to be significant using the P(C) score, while results reported as not significant were also found not to be significant using the P(C) score, although exact F- and p-values did differ between tests using the two scores.

[3]Note that analysis of arcsine-transformed P(C) data showed a significant difference between pairs 1–2 and 8–9 and 3–4 and 8–9.

[4]Although this difference in procedure could conceivably have contributed to the pattern of results observed here, we have also subsequently conducted additional tests with Cantonese listeners and without a warning tone and have observed results qualitatively similar to those reported in experiment 1.

[5]Flanagan and Saslow (1958) used a set of synthetic vowels (/ɑ/) with a standard f_0 of 120 Hz and 70 dB SPL and found a median of 0.63 Hz, while Harris (1952) used pure tones with a standard reference tone of 125 Hz at 30 dB SL and found a median of 0.74 Hz.

[6]Although the experiments conducted by Francis *et al.* (to appear) provided clear evidence for the presence of category boundaries using an identification paradigm, the discrimination testing results they report showed no evidence for the categorical perception of level tones in Cantonese. The question of whether lexical tones are perceived categorically is discussed in detail by Francis *et al.* (to appear).

Allen, J., Kraus, N., and Bradlow, A. (**2000**). "Neural representation of consciously imperceptible speech sound differences," Percept. Psychophys. **62**(7), 1383–1393.

Bauer, R. S., and Benedict, P. K. (**1997**). *Modern Cantonese Phonology* (Mouton de Gruyter, Berlin).

Berliner, J. E., Durlach, N. I., and Braida, L. D. (**1977**). "Intensity perception. VII. Further data on roving level discrimination and the resolution of bias edge effects," J. Acoust. Soc. Am. **61**, 1577–1585.

Braida, L. D., Lim, J. S., Berliner, J. E., Durlach, N. I., Rabinowitz, W. M., and Purks, S. R. (**1984**). "Intensity perception. XIII. Perceptual anchor model of context-coding," J. Acoust. Soc. Am. **76**, 722–731.

Burns, E. M., and Sampat, K. S. (**1980**). "A note on possible culture-bound effects in frequency discrimination," J. Acoust. Soc. Am. **68**, 1886–1888.

Ciocca, V., and Lui, J. Y. K. (**2003**). "The development of the perception of Cantonese lexical tones," J. Multiling. Commun. Disord. **1**, 141–147.

Cowan, N., and Morse, P. A. (**1986**). "The use of auditory and phonetic memory in vowel discrimination," J. Acoust. Soc. Am. **79**, 500–507.

Flanagan, J. L., and Saslow, M. G. (**1958**). "Pitch discrimination for synthetic vowels," J. Acoust. Soc. Am. **30**, 435–442.

Francis, A. L., Ciocca, V., and Ng, B. C. K. (to appear). "On the (non)categorical perception of lexical tones," Percept. Psychophys.

Gandour, J. T., and Harshman, R. (**1978**). "Crosslanguage differences in tone perception: A multidimensional scaling investigation," Lang. Speech **21**, 1–33.

Gandour, J. T. (**1981**). "Perceptual dimensions of tone: Evidence from Cantonese," J. Chin. Ling. **9**, 20–36.

Harris, J. D. (**1952**). "Pitch discrimination," J. Acoust. Soc. Am. **24**, 750–755.

Hauser, M. D., and Fowler, C. A. (**1992**). "Fundamental frequency declination is not unique to human speech: Evidence from nonhuman primates," J. Acoust. Soc. Am. **91**, 363–369.

Hellström, A. (**1985**). "The time-order error and its relatives: Mirrors of cognitive processes in comparing," Psychol. Bull. **97**, 35–61.

Huttenlocher, J., Hedges, L. V., and Duncan, S. (**1991**). "Categories and particulars: Prototype effects in estimating spatial location," Psychol. Rev. **98**(3), 352–376.

Huttenlocher, J., Hedges, L. V., and Vevea, J. L. (**2000**). "Why do categories affect stimulus judgment?," J. Exp. Psychol. Gen. **129**(2), 220–241.

IPA (**1999**). *Handbook of the International Phonetic Association* (Cambridge University Press, Cambridge).

Klatt, D. H. (**1973**). ''Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception,'' J. Acoust. Soc. Am. **53**(1), 8–16.
Klatt, D. H., and Klatt, L. C. (**1990**). ''Analysis, synthesis, and perception of voice quality variations among female and male talkers,' J. Acoust. Soc. Am. **87**, 820–857.
Ladefoged, P. (**2001**). *A Course in Phonetics*, 4th ed. (Thompson Learning, Australia), pp. 78–79.
Laver, J. (**1994**). *Principles of Phonetics* (Cambridge University Press, Cambridge), p. 157.
Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (**1957**). ''The discrimination of speech sounds within and across phoneme boundaries,'' J. Exp. Psychol. **54**, 358–368.
Lieberman, P., and Blumstein, S. E. (**1988**). *Speech Physiology, Speech Perception, and Acoustic Phonetics* (Cambridge University Press, Cambridge).
Macmillan, N. A. (**1987**). ''Beyond the categorical/continuous distinction: A psychophysical approach to processing modes,'' in *Categorical Perception*, edited by S. Harnad (Cambridge University Press, Cambridge), pp. 53–85.
Macmillan, N. A., Braida, L. D., and Goldberg, R. F. (**1987**). ''Central and peripheral processes in the perception of speech and nonspeech sounds,'' in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten (Martinus Nijhoff, Dordrecht, The Netherlands), pp. 28–45.
Macmillan N. A., and Creelman, C. D. (**1991**). *Detection Theory: A User's Guide* (Cambridge University Press, Cambridge).
Maddox, W. T., and Estes, W. K. (**1997**). ''Direct and indirect stimulus-frequency effects in recognition,'' J. Exp. Psychol. Learn. Mem. Cogn. **23**(3), 539–559.
Ohala, J. J. (**1978**). ''Production of tone,'' in *Tone: A Linguistic Survey*, edited by V. Fromkin (Academic, New York), pp. 5–39.
Pierrehumbert, J. A. (**1979**). ''The perception of fundamental frequency declination,'' J. Acoust. Soc. Am. **66**, 363–369.
Polka, L., and Bohn, O.-S. (**1996**). ''A cross-language comparison of vowel perception in English-learning and German-learning infants,'' J. Acoust. Soc. Am. **100**(1), 577–592.
Repp, B. H., and Crowder, R. G. (**1990**). ''Stimulus order effects in vowel discrimination,'' J. Acoust. Soc. Am. **88**(5), 2080–2090.
Repp, B. H., Healy, A. F., and Crowder, R. G. (**1979**). ''Categories and context in the perception of isolated steady-state vowels,'' J. Exp. Psychol. Hum. Percept. Perform. **5**(1), 129–145.
Russell, P., and Darwin, C. J. (**1991**). ''Real-time synthesis of complex sounds on a Mac II with 56001 DSP chip,'' Br. J. Audiol. **25**, 59–60.
Sharma, A., and Dorman, M. F. (**2000**). ''Neurophysiologic correlates of cross-language phonetic perception,'' J. Acoust. Soc. Am. **107**(5), 2697–2703.
Stagray, J. R., and Downs, D. (**1993**). ''Differential sensitivity for frequency among speakers of a tone and a non-tone language,'' J. Chin. Ling. **21**, 144–163.
Stevens, K. N. (**1972**). ''The quantal nature of speech: Evidence from articulatory-acoustic data,'' in *Human Communication: A Unified View*, edited by E. E. David, Jr. and P. B. Denes (McGraw-Hill, New York), pp. 51–66.
Stevens, K. N. (**1989**). ''On the quantal nature of speech,'' J. Phonetics **17**, 3–45.
Tanner, W. P., and Rivette, G. L. (**1964**). ''Experimental study of 'tone deafness,' '' J. Acoust. Soc. Am. **36**, 1465–1467.
Tremblay, K., Kraus, N., Carrell, T. D., and McGee, T. (**1997**). ''Central auditory system plasticity: Generalization to novel stimuli following listening training,'' J. Acoust. Soc. Am. **102**(6), 3762–3773.
Umeda, N. (**1982**). '''F_0 declination' is situation dependent,'' J. Phonetics **10**, 279–290.
Vaissiére, J. (**1995**). ''Phonetic explanations for cross-linguistic prosodic similarities,'' Phonetica **52**, 123–130.
Vance, T. J. (**1976**). ''An experimental investigation of tone and intonation in Cantonese,'' Phonetica **33**, 368–392.
Wong, P. C. M. (**1999**). ''The effect of downdrift in the production and perception of Cantonese level tone,'' in Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, Vol. 3, pp. 2395–2398.