Effects of language experience and stimulus complexity on the categorical perception of pitch direction

Yisheng Xu, Jackson T. Gandour,[a] and Alexander L. Francis
Department of Speech, Language, & Hearing Sciences, Purdue University, West Lafayette, Indiana 47907-2038

(Received 16 December 2005; revised 18 May 2006; accepted 19 May 2006)

Whether or not categorical perception results from the operation of a special, language-specific, speech mode remains controversial. In this cross-language (Mandarin Chinese, English) study of the categorical nature of tone perception, we compared native Mandarin and English speakers' perception of a physical continuum of fundamental frequency contours ranging from a level to rising tone in both Mandarin speech and a homologous (nonspeech) harmonic tone. This design permits us to evaluate the effect of language experience by comparing Chinese and English groups; to determine whether categorical perception is speech-specific or domain-general by comparing speech to nonspeech stimuli for both groups; and to examine whether categorical perception involves a separate categorical process, distinct from regions of sensory discontinuity, by comparing speech to nonspeech stimuli for English listeners. Results show evidence of strong categorical perception of speech stimuli for Chinese but not English listeners. Categorical perception of nonspeech stimuli was comparable to that for speech stimuli for Chinese but weaker for English listeners, and perception of nonspeech stimuli was more categorical for English listeners than was perception of speech stimuli. These findings lead us to adopt a memory-based, multistore model of perception in which categorization is domain-general but influenced by long-term categorical representations.
 [DOI: 10.1121/1.2213572]

PACS number(s): 43.71.−k, 43.71.An, 43.71.Hw [KEG] Pages: 1063–1074

I. INTRODUCTION

Categorical perception (CP) has been one of the most extensively studied phenomena in speech perception for nearly 50 years because it is believed to reflect some fundamental aspects of the processing of speech sounds (see Harnad, 1987, for an overview). Up to the present, the bulk of CP research has been directed to segmental features of speech, i.e., consonants and vowels. Though not without controversy, the consensus opinion has been that consonant features are perceived in a mostly categorical fashion whereas vowel features are not (e.g., Liberman *et al.*, 1957; Fry *et al.*, 1962). Recently there has been increasing interest in suprasegmental features of speech (e.g., pitch: Francis *et al.*, 2003; Hallé *et al.*, 2004). In this regard, *tone languages* are especially advantageous because contrasts in pitch height and/or direction are exploited to minimally distinguish lexical items (Gandour, 1978). Mandarin Chinese, for example, has four lexical tones: e.g., ma^1 "mother," ma^2 "hemp," ma^3 "horse," ma^4 "scold". These four tones can be described phonetically as high level, high rising, falling rising, and high falling, respectively (Howie, 1976).

Earlier research on the categorical nature of tone perception reveals the importance of pitch movement (level versus contour) and the influence of language experience (tonal versus nontonal). With respect to pitch movement, a continuum ranging from one (or more) level tone to another is not perceived categorically (Thai—Abramson, 1979; Cantonese—Francis *et al.*, 2003), whereas a continuum ranging from a high level tone to a high rising contour tone is perceived categorically (Mandarin—Wang, 1976; Cantonese—Francis *et al.*, 2003). Using a stimulus continuum ranging from a Mandarin Tone 2 (high rising) to a Tone 1 (high level), cross-language comparisons show that native speakers perceive this tonal contrast in a categorical manner but that speakers of a nontone language (English) do not (Wang, 1976). In a more recent cross-language CP study of lexical tones (Taiwan Mandarin—Hallé *et al.*, 2004), it is argued that there is a gradient in the degree of CP—as measured by the slope of identification curve, the peakedness of identification response time, and the peakedness of discrimination performance—that varies depending on a listener's degree of familiarity with lexical tones. Using three tonal continua (1 versus 2; 2 versus 4; 3 versus 4), their results show that Taiwanese listeners' perception of tones shows a higher degree of CP ("quasicategorical") than that of listeners of a nontone language (French). Despite their lack of familiarity with lexical tones, French listeners are still able to make their identification and discrimination judgments on the basis of psychophysical factors alone. Their theoretical explanation is based on a cross-linguistic model of speech perception, PAM, perceptual assimilation model (Best *et al.*, 2001). According to this account, lexical tones in Taiwanese are not necessarily difficult for French speakers to perceive, but they are perceived in a noncategorical manner because they are *not assimilable* to any phonemic unit of French, and thus do not invoke phonetic perception processes specific to French. This model essentially stems from a view that CP relies on a speech mode of perception based on native phoneme categories (Liberman *et al.*, 1967; Studdert-Kennedy *et al.*, 1970).

[a]Electronic mail: gandour@purdue.edu

Whether CP is restricted to a special speech mode is still a matter of controversy. Earlier data showed that CP occurs in synthetic speech stimuli but not in their spectrally rotated nonspeech correlates (Liberman *et al.*, 1961). Subsequently, CP has been reported for nonspeech continua as well. For example, CP has been observed using a noise-lead time continuum that contrasts the onset of a noise relative to a buzz (Miller *et al.*, 1976) and a tone onset time continuum that contrasts the relative onset time of two component tones (Pisoni, 1977). These findings demonstrate that it is the temporal order of two acoustic events which underlies the CP phenomenon rather than a mechanism unique to speech (e.g., voice onset time). By varying the amplitude rise time of simple sawtooth stimuli, CP has also been demonstrated with a pluck-bow continuum simulating the sounds of music instruments (Cutting and Rosner, 1974; Cutting, 1982). More recently, Mirman *et al.* (2004) compared the degree of CP between a nonspeech continuum employing a rapid-changing cue (amplitude rise time of noise) and another nonspeech continuum employing a steady-state cue (spectral notch center frequency of noise). They concluded that CP is more dependent on rapid-changing than steady-state acoustic cues, which correlates with differences in the degree of CP between stop consonants and static vowels. Another line of evidence supporting a domain-general view comes from the observations of CP in nonhuman animals (e.g., chinchilla: Kuhl and Miller, 1975; monkey: Kuhl and Padden, 1983) when they are presented with speech continua employing temporal-order or rapid-changing cues. Based on these findings, some researchers believe that CP may result from natural sensory discontinuities (Pastore *et al.*, 1977; Kuhl, 1981; Stevens, 1981). According to this hypothesis, the phoneme boundaries in speech simply occur at regions of heightened natural auditory sensitivities.

None of the major explanations of CP appears to be completely adequate on its own (Rosen and Howell, 1987). Some of these shortcomings may be attributed to issues of experimental design. In the extant literature, for example, CP studies have compared native to non-native language listeners and speech to nonspeech stimuli, but as far as we know, no previous study has attempted to incorporate both factors into a single experimental design. We argue that it is important to look at these two intersecting factors of the phenomenon concurrently in order to achieve a better understanding of CP. In this cross-language (Chinese, English) study of the categorical nature of tone perception in Mandarin Chinese, we include both variables: language group (native versus nonnative) and stimulus type (speech versus nonspeech). For the two language groups, one is comprised of native speakers of Mandarin, the other of native speakers of English who are unfamiliar with Mandarin or any other tone language. For the two stimulus types, we employ a physical continuum ranging from a high rising to high level tone in Mandarin in comparison to homologous harmonic tones. In an effort to equalize the stimuli between speech and nonspeech, we opted to use linear instead of curvilinear contours. Linear ramps commonly occur in nonspeech contexts but represent at best a crude approximation of natural speech tonal contours, and therefore are less likely to give any perceptual advantage to the native listeners. This experimental design permits us to evaluate the effect of language experience by comparing Chinese and English groups; to determine whether CP is speech-specific or domain-general by comparing speech to nonspeech stimuli for both Chinese and English listeners; and to examine whether CP involves a separate categorical process, over and above regions of sensory discontinuity, by comparing speech to nonspeech stimuli for English listeners. In so doing, we are led to develop a multistore model of CP. By adapting earlier multistore information processing models (e.g., Atkinson and Shiffrin, 1968; Cowan, 1988) as well as previous memory-based interpretations of CP (e.g., Pisoni, 1975; Macmillan, 1987), this new model enables us to provide a unified account of the CP phenomenon.

II. METHOD

A. Subjects

Thirty native speakers of Mandarin Chinese (13 males; 17 females) and thirty native speakers of American English (15 males; 15 females) participated in this experiment. The two groups were similar in age (Mean±s.d.: Chinese =27.5±2.9; English=23.2±4.3) and years of formal education (Mean±s.d.: Chinese=18.7±2.8; English=15.9±3.3). All participants exhibited normal hearing sensitivity (pure-tone air conduction thresholds of 20 dB HL or better in both ears at frequencies of 0.5, 1, 2, and 4 kHz). All Chinese listeners were from mainland China, and none had received any formal instruction in English until after the age of 11. No English listener had any previous exposure to Chinese or for that matter any other tone language. None of the participants, Chinese or English, had more than five years of formal musical training, and none had any recent musical training (within the past five years). All participants were paid for their participation. They gave informed consent in compliance with a protocol approved by the Institutional Review Board of Purdue University.

B. Stimuli

Two sets of stimuli were constructed for this experiment: speech (S) and nonspeech (NS) (Fig. 1). The speech stimuli were derived from the Mandarin syllable [i] with a high level tone (Tone 1: yi^1, "clothing") that was produced by a male native speaker. The nonspeech stimuli were harmonic tones that exhibited the same pitch, amplitude, and duration parameters as the speech stimuli. The speech and nonspeech stimuli differed from one another only in spectral components. Each set consisted of seven stimuli derived from the same pitch continuum ranging in equal steps from a *level* to a *rising* linear ramp.

The fundamental frequency (F_0) contours of the pitch continuum were modeled by seven linear functions:

$$f_i(t) = (f_2 - f_{1,i})t/d + f_{1,i}, \quad i = 1,2,\ldots,7. \tag{1}$$

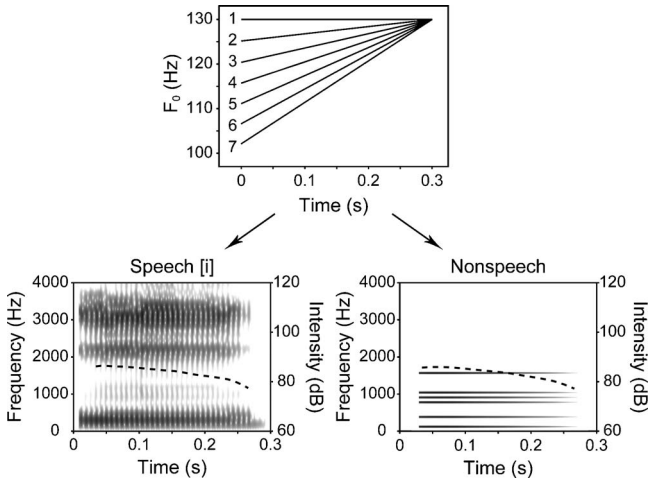

FIG. 1. Synthetic speech (Mandarin [i] syllables) and nonspeech (harmonic tones) stimuli based on the same F_0 continuum (top panel). The bottom-left panel shows a broadband spectrogram of the Step 1 speech stimulus. The bottom-right panel shows a narrowband spectrogram of its nonspeech homologue. Their intensity envelopes (dashed lines) are closely matched.

The constant d was the duration of the stimuli (0.3 s). The offset frequency (f_2) was fixed at 130 Hz. Onset frequencies ($f_{1,i}$) were separated on a psychoacoustic scale by an equal step size of 0.1200 ERB (equivalent rectangular bandwidth) (Table I). The conversions from frequency (f) in Hz to ERB-rate (E) in ERB and vice versa were based on Eq. (2) (Greenwood, 1961):

$$E = 16.7 \log_{10}(f/165.4 + 1),$$
$$f = 165.4(10^{E/16.7} - 1). \tag{2}$$

Using PRAAT software (Boersma and Weenink, 2003), speech stimuli were resynthesized from the Tone 1 natural speech template. Its duration was first scaled to 0.3 s. The F_0 contour of this speech template was then replaced with Eq. (1) by applying PSOLA (pitch-synchronous overlap and add) resynthesis (Valbret *et al.*, 1992). The two ends of the continuum, Step 1 and Step 7, were judged qualitatively to be good exemplars of Mandarin Tones 1 (yi^1, "clothing") and 2 (yi^2, "aunt"), respectively, by three Chinese listeners who did not participate in the experiment.

Nonspeech stimuli were composed of six equal-amplitude harmonics (1, 3, 6, 7, 8, 12) of the F_0 (cf. Francis and Ciocca, 2003). Harmonics 2, 4, 5, 9, 10, and 11 were omitted to increase perceptual dissimilarity with the speech stimuli. These harmonic tones were first created with:

TABLE I. Frequency (f) in Hz and ERB-rate (E) in ERB of the F_0 onset in each stimulus.

Step	f (Hz)	E (ERB)	Step size (Hz)	Step size (ERB)
1	130.00	4.2063	4.85	0.1200
2	125.15	4.0863	4.77	0.1200
3	120.38	3.9663	4.69	0.1200
4	115.70	3.8463	4.61	0.1200
5	111.08	3.7263	4.54	0.1200
6	106.55	3.6063	4.46	0.1200
7	102.08	3.4863		

$$y(t) = \frac{1}{6}\sum_{i=1}^{6} \sin 2\pi h_i[(f_2 - f_1)t/2d + f_1]t,$$
$$h_i = 1,3,6,7,8,12. \tag{3}$$

The resulting wave forms were then multiplied by the amplitude envelope of the speech template sample by sample, and peak normalized to the same intensity level.

For the practice tasks, speech and nonspeech stimuli were similarly constructed using a seven-step pitch continuum ranging from level to rising F_0 contours. However, the practice stimuli differed from the experimental stimuli in terms of both F_0 and spectral properties. F_0 offset frequency was set one octave higher (at 260 Hz) than that of the experimental stimuli. Despite the octave shift, psychophysical step size was maintained (0.1200 ERB) in order to preserve perceptual equivalence of pitch movement (Hermes and van Gestel, 1991). The practice speech stimuli were resynthesized from a Mandarin /a/ syllable produced with a high level tone by a female native speaker. The practice nonspeech stimuli were composed of six harmonics, but most were different (1, 3, 4, 5, 9, 11) from those used in the experimental stimuli.

C. Procedures

Chinese (C) and English (E) listeners were assigned to one of four subgroups according to stimulus set (speech, S; nonspeech, NS). There were 15 participants per subgroup: CS=Chinese group, speech condition; ES=English group, speech condition; CNS=Chinese group, nonspeech condition; ENS=English group, nonspeech condition. Each participant was asked to perform identification and discrimination tasks in either the speech or nonspeech condition. The two tasks were presented in random order across participants.

1. Identification task

In the identification task, participants listened to stimuli from the speech/nonspeech continuum presented in isolation. They were instructed to press the left mouse button upon hearing a "level" pitch, or the right mouse button upon hearing a "rising" pitch. There were 20 occurrences of each of the seven stimuli (140 trials) presented in random order. The rate of presentation was self-paced. Once a response was collected, the next stimulus was presented automatically following a 1-s pause.

2. Discrimination task

In the discrimination task, stimuli were presented in pairs with a 500-ms interstimulus interval (ISI). This ISI duration was selected to maximize differences in the performance of between- versus within-category discrimination (Pisoni, 1973). A total of 170 pairs were presented in random order. Of these pairs, 100 consisted of two different stimuli separated by two steps on the speech/nonspeech continuum (*different* pairs), in either forward (1-3, 2-4, 3-5, 4-6, 5-7) or reverse order (3-1, 4-2, 5-3, 6-4, 7-5). There were 10 occurrences of each of the ten 2-step pairs. The remaining 70 pairs

contained one of the seven stimuli on the speech/nonspeech continuum paired with itself (*same* pairs). There were ten occurrences per stimulus pair. After hearing each pair, participants were instructed to judge whether the two stimuli were the same or different, and to respond by pressing a mouse button (left="same," right="different"). The rate of presentation was self-paced as in the identification task.

All the stimuli were presented binaurally at ~72 dB SPL through a pair of Sony MDR-7506 headphones. Stimulus presentation and response collection were implemented using E-PRIME software (Schneider *et al.*, 2002).

3. Practice tasks and subject prescreening

Prior to each of the two experimental tasks, participants were asked to perform a similar identification or discrimination task using a different set of speech/nonspeech stimuli. These two practice tasks were designed primarily to familiarize them with task requirements in order to stabilize their performance in the actual experiment. Another objective was to minimize differences in task difficulty between language groups that might be attributed to the non-native (English) listeners' unfamiliarity with Chinese sounds. Finally, they allowed us to identify individuals who were unable to reach our threshold criterion for task performance.

After the practice identification task, separate binomial tests were performed on each participant's percentage of "level" responses to Step 1 and the percentage of "rising" responses to Step 7. After the practice discrimination task, a binomial test was performed on the percentage of correct "same" and "different" discriminations. A significance threshold of $p<0.05$ (significantly above chance) was used to determine whether the subject could proceed to the experimental tasks. Only 3 out of 63 participants, one Chinese and two English, failed to pass the prescreening threshold, and thus were excluded from the experiment.

D. Data analysis

To investigate the effects of language experience (native or non-native) and domain specificity (speech or nonspeech) on identification and discrimination performance, we obtained individual measures for each subject based on three essential characteristics of CP: sharp category boundary, corresponding discrimination peak, and prediction of discrimination from identification (Treisman *et al.*, 1995).

1. Identification function and categorical boundary

Based on the binomial distribution of the identification scores and the sigmoid shape of the response function, a logistic regression [generalized linear model, see Eq. (4)] between the *identification score* (P_I) and a repeated measures predictor, *step number* (x), was adopted to obtain the mean identification function for each subgroup:

$$\log_e\left(\frac{P_I}{1-P_I}\right)=b_0+b_1x. \tag{4}$$

Note that x can be treated as a continuous variable proportional to the ERB-rate scale although we only sampled at discrete steps on the continua. Since the identification functions for "level" and "rising" responses are symmetrical, only the latter was analyzed. The GEE (Generalized Estimating Equations) (Liang and Zeger, 1996) estimated regression coefficient b_1 was used to evaluate the slope of the fitted logistic curve (Kutner *et al.*, 2005, p. 567), which is an indication of the *sharpness* of the categorical boundary. The effects of language group (GROUP=C,E) and stimulus set (STIM=S,NS) on b_1 were estimated by their interactions with step number x in the generalized linear model using GROUP, STIM, and x as predictors. We derived the mean position of the categorical boundary in each subgroup from the value of step number (x_{cb}) corresponding to the 50% identification score,

$$\begin{aligned} &b_0+b_1x_{cb}=\log_e\left(\frac{0.5}{1-0.5}\right)=0 \\ &\Rightarrow x_{cb}=-\frac{b_0}{b_1}. \end{aligned} \tag{5}$$

Similarly, x_{cb} for each subject was derived from individual logistic response functions, and then analyzed by a two-way ANOVA for GROUP and STIM effects. Other dependent measures described in the following sections—P_{bc}, P_{wc}, P_{pk}, z, D—were analyzed by similar ANOVA models.

2. Obtained discrimination scores and related measures

In order to compute the obtained discrimination score (P), we divided the 170 discrimination trials into five 2-step comparison units. Each unit was comprised of all the trials in four types of pairwise comparisons (AB, BA, AA, and BB) for stimuli A and B separated by two steps. There were 40 trials in each unit. Adjacent comparison units contained overlapping AA or BB trials (e.g., the ten 3-3 pairs were included in both 1-3 and 3-5 units). P for each comparison unit was defined by

$$P=P(''S''|S)\cdot P(S)+P(''D''|D)\cdot P(D). \tag{6}$$

The percentages of "same" ("S") and "different" ("D") responses of all the same (S) and different (D) trials (i.e., the correct responses) in each comparison unit were represented by two conditional probabilities, $P(\text{"}S\text{"}|S)$ and $P(\text{"}D\text{"}|D)$, respectively. $P(S)$ and $P(D)$ were the probabilities of S (AA or BB) and D (AB or BA) trials in each unit, which were both equal to 0.5 in this experiment.

The obtained discrimination data for each subject were then examined by three different measures: *between-category* discrimination sensitivity (P_{bc}), measured from the comparison unit corresponding to the categorical boundary (x_{cb}) determined from the subgroup identification functions (for all subjects in this experiment, $P_{bc}=P_{35}$); *within-category* discrimination sensitivity (P_{wc}), which was the average of the two comparison units (P_{13} and P_{57}) at the ends of the continuum (cf. Pisoni, 1973); and *peakedness* of the discrimination function (P_{pk}), estimated by the difference between P_{bc} and P_{wc}.

3. Prediction of discrimination from identification

The predicted discrimination score P^* was computed using Eq. (7) (Pollack and Pisoni, 1971),

$$P^* = [1 + (P_A - P_B)^2]/2, \quad (7)$$

where the identification scores (as one of the two categories, uniformly either "level" or "rising") of the two stimuli A and B in a comparison unit. This equation was drawn from an extreme assumption that same-different discrimination is solely determined by the identification of the two stimuli as the same or different categories [covert identification; the original model was discussed in Liberman *et al.* (1957)].

The predictability of discrimination from identification was examined on two aspects in terms of the degree of CP (Liberman *et al.*, 1957). First, the correlation between the predicted and obtained discrimination scores, i.e., the similarity in the shape of the two discrimination curves, especially the position of the discrimination peaks, was measured by Fisher's z-transformed correlation coefficient (z) to obtain normally distributed data for ANOVA. Second, the distance between the two discrimination curves was measured by the mean difference ($P - P^*$) between the obtained and predicted discrimination scores. The underestimation of discrimination from identification is presumably due to the involvement of continuous perception in the discrimination task (Macmillan, 1987).

III. RESULTS

A. Logistic identification functions

The estimated regression coefficients for the mean logistic response functions of the four subgroups (CS, CNS, ES, ENS) are presented in Table II. These parameters were used to plot the identification functions in Fig. 2.

B. Sharpness of the category boundary

In the generalized linear model for logistic regression, a significant two-way interaction ($Z=3.45$; $p=0.0006$) was observed between GROUP and step number x, indicating that Chinese listeners showed sharper category boundaries (larger b_1; see Table II) than English for both speech and nonspeech stimuli. This difference can also be visualized from the logistic response functions in Fig. 2. The STIM $\times$ x interaction and GROUP $\times$ STIM $\times$ x interaction were not significant.

C. Position of the category boundary

The position of the category boundary (x_{cb}) computed from subgroup and individual logistic regressions is presented in Table II. A two-way ANOVA on individual x_{cb} yielded only a significant STIM main effect [$F(1,56)=4.72$; $p=0.0341$]. For both language groups (Chinese, English), although the mean category boundaries for both speech and nonspeech were approximately centered at the middle of the continua ($x_{cb} \approx 4$), the boundary of the speech continuum was slightly shifted toward the "rising" end ($\approx$0.34 step; Fig. 2) as compared to the nonspeech continuum.

TABLE II. GEE estimates of regression coefficients (b_0, b_1) and the derived categorical boundary (x_{cb}) for each subgroup (C=Chinese, E=English; S=speech; NS=nonspeech).

Subgroup	b_0	b_1	$x_{cb}=-b_0/b_1$
CS	−10.3094	2.4266	4.2485
CNS	−8.1686	2.0855	3.9169
ES	−4.4333	1.0611	4.1780
ENS	−5.0307	1.3146	3.8268

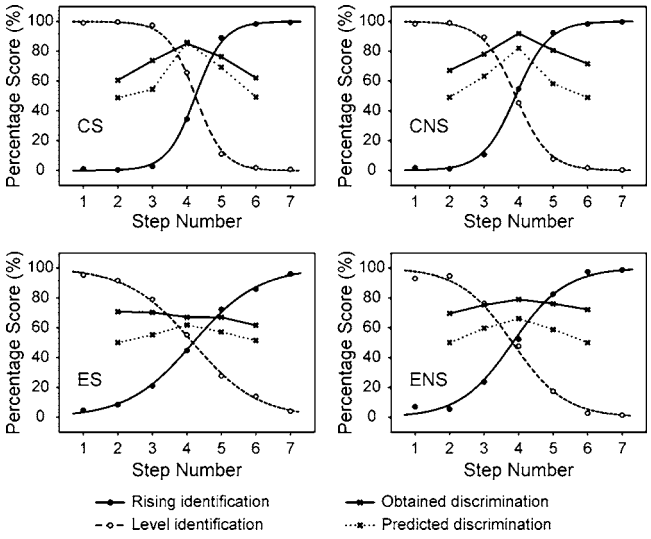

FIG. 2. Logistic identification functions ("level" and "rising") and two-step discrimination curves ("obtained" from the discrimination scores and "predicted" from the identification scores) for each subgroup (C=Chinese; E=English; S=speech; NS=nonspeech). The "level" logistic response functions (dashed lines) were plotted by reflecting the "rising" logistic response functions (solid lines) across the 50% horizontal line ($P_{level}=1-P_{rising}$).

D. Between-category discrimination

A two-way ANOVA revealed significant GROUP [$F(1,56)=69.44$; $p<0.0001$] and STIM [$F(1,56)=23.89$; $p<0.0001$] main effects on between-category discrimination sensitivity (P_{bc}) [Fig. 3(a)]. The GROUP $\times$ STIM interaction effect was not significant, indicating that Chinese listeners showed better between-category discrimination sensitivity than English in both speech (mean: CS=86.5%; ES=68.5%) and nonspeech (CNS=92.2%; ENS=80.3%) stimulus sets. For both language groups, nonspeech stimuli yielded better between-category discrimination sensitivity than speech.

E. Within-category discrimination

A two-way ANOVA showed significant GROUP [$F(1,56)=4.75$; $p=0.0335$] and STIM [$F(1,56)=11.09$; $p=0.0015$] main effects on within-category discrimination sensitivity (P_{wc}) [Fig. 3(b)]. The GROUP $\times$ STIM interaction effect was not significant, indicating that English listeners showed better within-category discrimination sensitivity than Chinese in both speech (mean: CS=62.7%; ES=67.6%) and nonspeech (CNS=69.6%; ENS=72.3%) stimulus sets. For both language groups, nonspeech stimuli yielded better within-category discrimination sensitivity than speech.

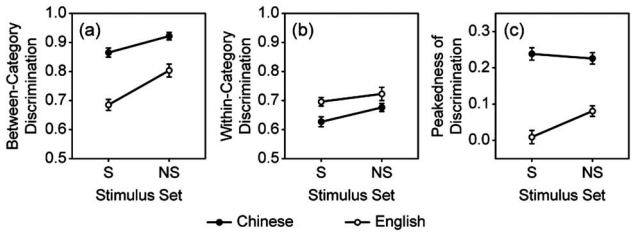

FIG. 3. Discrimination score in percentage correct: (a) between-category; (b) within-category; (c) peakedness of discrimination. (S=speech; NS=nonspeech.)

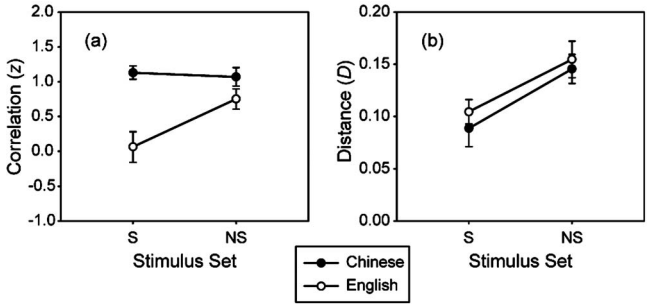

FIG. 4. The relationship between predicted and obtained discrimination curves: (a) Fisher's z-transformed correlation; (b) mean distance. (S=speech; NS=nonspeech.)

F. Peakedness of discrimination

A two-way ANOVA on peakedness of discrimination (P_{pk}) showed a significant GROUP×STIM interaction [$F(1,56)=6.61$; $p=0.0128$; Fig. 3(c)]. Tukey adjusted multiple comparisons indicated that Chinese listeners obtained a higher discrimination peak than English for both speech and nonspeech stimuli ($p_{adjusted}<0.0001$). In addition, English listeners achieved a higher discrimination peak in response to nonspeech than speech stimuli ($p_{adjusted}=0.0157$). Student t-tests for P_{pk} in each of the four subgroups showed that only the mean P_{pk} of the ES subgroup was not significantly greater than zero ($p=0.5777$). A significant discrimination peak was observed in all of the other three subgroups (CS, CNS, ENS; for mean $P_{pk}>0$, $p<0.0001$).

G. Correlation between predicted and obtained discrimination

A two-way ANOVA on the z-transformed correlation between the predicted and obtained discrimination scores revealed a significant GROUP×STIM interaction [$F(1,56)=5.85$; $p=0.0188$; Fig. 4(a)]. Tukey adjusted multiple comparisons indicated that English listeners, when performing the speech discrimination task (ES subgroup), showed significantly lower correlation than the other three subgroups (ES<CS, $p_{adjusted}<0.0001$; ES<CNS, $p_{adjusted}=0.0001$; ES<ENS, $p_{adjusted}=0.0136$). This result is consistent with the shapes of subgroup discrimination curves shown in Fig. 2.

H. Underestimation of discrimination by identification

A two-way ANOVA showed only a significant effect of STIM [$F(1,56)=12.13$; $p=0.0010$] on the mean distance between the obtained and predicted discrimination functions [Fig. 4(b)]. For both language groups, underestimation of discrimination by identification was greater in nonspeech than in speech stimuli.

IV. DISCUSSION

A. Categorical perception of tone

Identification and discrimination tasks reveal classical patterns of CP in Chinese listeners for both speech (CS) and nonspeech (CNS) stimuli varying along a linear F_0 continuum from level to rising. The results for the speech stimuli (CS) are comparable to those for stop consonants (Liberman *et al.*, 1957, 1961) in terms of defined characteristics of CP: a sharp category boundary, a corresponding discrimination peak, and predictability of discrimination from identification. The exhibition of CP for pitch direction is dependent on a listener's experience with a tone language (CS) as shown by the lack of similar CP effects in nontone language listeners (ES). For Chinese listeners, some aspects of CP—sharp identification boundary and discrimination peak—extend to homologous nonspeech stimuli. (CNS). Regardless of language group, discrimination scores in the nonspeech tasks (CNS, ENS) are better overall than those in the speech tasks (CS, ES). This may be related to differences in the complexity between the two sets of stimuli. (See Sec. IV C.)

One explanation for CP claims that it arises from variations in natural auditory sensitivity along some acoustic continuum without any reference to speech-specific mechanisms (Pastore *et al.*, 1977; Stevens, 1981). Interestingly, both speech and nonspeech stimuli in this study yield identification boundaries near the middle of the F_0 continuum regardless of language experience. This finding is inconsistent with a previous claim of separate linguistic and psychophysical boundaries on a synthetic speech continuum varying in pitch direction from level to rising (Wang, 1976). Wang reported two separate boundaries in a comparison between Chinese ($n=2$; linguistic) and English ($n=2$; psychophysical) listeners. The current study, on the other hand, employs two continua, speech and nonspeech, in addition to a much larger sample size ($n=15$) per language group. We argue that the nonspeech continuum is a *sine qua non* for making a direct comparison of the relationship between the two boundary types. Prima facie, this overlap in identification boundaries, irrespective of stimulus type or language group, may be construed to support natural auditory sensitivities as the driving force behind CP. By this account, the presence of a psychophysical boundary due to the heterogeneity of natural auditory sensitivities along the continuum predicts that the identification boundary and discrimination peak would co-occur in the same location for both speech and nonspeech stimuli. The fact that no discrimination peak is observed at the identification boundary in the speech task for the English group (ES) runs counter to the auditory-explanation. We further propose that the relatively small discrimination peak that is observed in response to the nonspeech stimuli for the English group (ENS) is likely due to memory rather than auditory mechanisms. (See Sec. IV B.)

In multidimensional scaling of listeners' perception of

linear F_0 ramps (level, falling, rising), cross-language comparisons show that the relative importance of the pitch height and direction dimensions varies depending on a listeners' familiarity with specific types of pitch patterns that occur in the tone space of their native language (Gandour and Harshman, 1978; Gandour, 1983). Pitch height distinguishes tones, either static or dynamic, that differ according to average F_0, while direction of change distinguishes primarily rising from nonrising tones. The perceptual salience of the direction dimension is greater for native speakers of tone languages, including Mandarin Chinese, than for speakers of nontone languages (English), while English listeners give greater weight to the height dimension than do tone language speakers. Such cross-language differences in perceptual weighting suggest that linguistic experience directs attention to linguistically relevant properties of the auditory signal. In this experiment, the fact that Chinese listeners show as much CP for nonspeech sounds as for speech sounds, whereas English listeners do not, suggests that Chinese listeners' CP for nonspeech sounds does not result simply from the presence of an innate region of heightened sensitivity. Otherwise English listeners would be expected to show the same CP effects in the ENS condition as the Chinese listeners did in the CNS condition. We therefore conclude that Chinese listeners' nonspeech performance must derive at least in part from their experience with listening to Chinese pitch patterns. These data lead us to infer that Chinese listeners' native language experience has changed the way they process pitch patterns regardless of the stimulus context in which these patterns are embedded. Although the basis for cross-language differences in CP may emerge from linguistic experience, the effects of such experience is not specific to speech perception (contra Studdert-Kennedy *et al.*, 1970). More generally, we predict CP effects whenever listeners are asked to judge auditory features that are similar to linguistically relevant speech parameters in their native language no matter whether they are presented in the context of natural speech or not.

B. Memory mechanisms of categorical perception

Instead of attributing CP effects to natural auditory sensitivities or speech-specific processes, we explain the data from this experiment by memory mechanisms involved in the CP tasks and their interactions with different levels of stimulus complexity between speech and nonspeech.

1. The dual-process model

The interpretation of CP by the contribution of distinct memory codes was first proposed by Fujisaki and Kawashima (1969, 1970, 1971), often referred to as the dual-process model (Macmillan, 1987, p. 55), and extensively discussed by Pisoni and colleagues in a series of experiments on phoneme discrimination (Pisoni, 1973; Pisoni and Lazarus, 1974; Pisoni, 1975). In this model, short-term memory (STM) recruited in CP tasks is divided into a continuous auditory short-term store and a categorical phonetic short-term store (Pisoni, 1975, pp. 8–9). The auditory memory code is subject to rapid decay and is dominant in within-category discrimination. The phonetic memory code is more stable due to "contact" with representations residing in long-term memory, and is dominant in between-category discrimination. Overall discrimination sensitivity is determined by the sum of these two types of memory available in the decision-making stage after decay. Both auditory and phonetic modes-coexist during the discrimination task but subjects operate exclusively in phonetic mode during the identification task. The dual-process model can successfully account for the underestimation of discrimination by identification observed in applications of the classical Haskins model (Liberman *et al.*, 1957), including the present experiment. Pisoni (1973) also employed the model to interpret differences in the degree of CP between consonants and vowels (Fry *et al.*, 1962; Stevens *et al.*, 1969), arguing that auditory STM appears to make a greater contribution to the discrimination of vowels as compared to stop consonants.

The dual-process model can account for our finding that Chinese listeners showed a sharper identification boundary and higher discrimination peak than English listeners. These language-dependent effects presumably indicate the role of phonetic memory in the CP tasks. However, because this model claims that CP effects emerge from a speech-specific phonetic mode, it cannot explain the robust categorical effects for nonspeech discrimination in the Chinese listeners (CNS) or the quasicategorical effects in the English listeners (ENS), as well as those well-documented observations of CP for certain nonspeech continua (Cutting and Rosner, 1974; Miller *et al.*, 1976; Pisoni, 1977; Mirman *et al.*, 2004). Moreover, although a dual-process model can account for the observed increase of between-category discrimination by the existence of phonetic memory, it cannot account for the observed decrease of within-category discrimination in Chinese listeners as compared to English. According to the dual-process model, both Chinese and English listeners would be expected to show similar within-category discrimination unless we admit a culture-bound difference in auditory sensitivity (Tanner and Rivette, 1964; Stagray and Downs, 1993; for contra evidence, see Burns and Sampat, 1980).

2. Signal-detection models

Another approach to CP utilizing memory mechanisms was based on Durlach and Braida's (1969) quantitative model for intensity perception. This approach employs a continuous theory of perception, i.e., signal detection theory (SDT: Green and Swets, 1966; Macmillan and Creelman, 1991). Again, two modes of memory operation are assumed: a *trace* mode and a *context-coding* mode. In the trace mode, the sensory memory trace of a previously heard stimulus is maintained for comparison with the sensation of a following stimulus; the trace is subject to temporal decay and interference, called trace variance. In the context-coding mode, the sensation of each stimulus is compared to a general stimulus context; the variance of this mode increases with the width of the context (i.e., overall stimulus range). The total variance is the sum of the two sources of memory variances plus the basic sensory variance, which is presumed to be constant. These two memory operations reflect domain-general processes that exist in both speech and nonspeech perception. Applying this model to CP explains the discrepancy between

fixed/roving discrimination and identification sensitivity as a function of differential contributions of trace and context-coding modes (Macmillan *et al.*, 1977; Macmillan, 1987). The vowel-consonant difference in CP can also be explained by the differential amount of context-coding variance (Ades, 1977; Macmillan *et al.*, 1988). The notion of context-coding now allows us to examine CP not only on the basis of stimulus sensitivity, but also in terms of task context (Macmillan *et al.*, 1977; Gerrits and Schouten, 2004).

A major limitation of SDT models is that they do not distinguish between short-term and long-term components in the context-coding mode, as pointed out by Macmillan *et al.* (1988, p. 1277). According to Durlach and Braida's (1969, p. 374) original definition, context may be derived either from the stimuli presented in the actual experiment (experimental context) or from a much larger set of stimuli from earlier experience (permanent context). Unfortunately, only factors related to the experimental context were considered in these models (Durlach and Braida, 1969; Braida *et al.*, 1984), thus making it impossible to measure any experience-dependent effects on CP. In the present experiment, this distinction is crucial. Cross-language differences in the sharpness of the identification boundary and discrimination peak can only be explained in terms of differences in the two groups' long-term experience with their respective native languages, and *not* in terms of differences in experimental context (which was identical for Chinese and English listeners). The current formulation of SDT models, however, does not incorporate a formal distinction between the effects of experimental and permanent context. Such models are therefore unable to give a full account of our data.

A more recent SDT model (van Hessen and Schouten, 1992) has incorporated a *labeling* process. The total variance consists of sensory variance and three sources of memory variance: trace, context-coding (temporary context), and labeling (permanent context). In this model, however, the categorical nature of the labeling process is not clearly distinguished from continuous sensory encoding. Moreover, CP reflects access to *permanent* phoneme labels only. The quasicategorical effects observed in our experiment are left unexplained in the case of nonspeech discrimination for English listeners (ENS) since they do not have access to permanent phoneme labels for such pitch stimuli.

3. A multistore model of CP

In our view, both dual-process and SDT models recognize CP as a complex phenomenon that emerges from at least two memory processes. Yet neither of these two models recognizes *categorization* as a process inherent to perception.

The SDT models attempted to explain CP by adapting a theory for continuous sensory processes (Macmillan *et al.*, 1977; Macmillan, 1987). We argue that the internal responses generated by the categorization process obey the rule of *discrete* probability distributions, fundamentally different from the sensory-level continuous Gaussian distributions assumed by SDT. In the case of binary internal responses, the probability for each response as a function of the stimulus value can be modeled by sigmoidal functions [e.g., a logistic function; see Eq. (4) and related commentary in Sec. II]. This discrete distribution generates a local maximum of perceptual sensitivity at the category boundary [cf. Eq. (7)]. Its mechanism represents essentially the same labeling process as described in the Haskins model. The dual-process model incorporated the categorization process as a separate memory mechanism parallel to sensory processing. But this process was treated as a special mode for speech perception only, whereas we claim that it is independent of domain or modality. This argument is supported not only by CP of melodic music intervals (Burns and Ward, 1978) and color hues (Bornstein, 1987), but also by observations from visual and auditory processing in other species (Herrnstein and Loveland, 1964; Kuhl and Miller, 1975; Freedman *et al.*, 2001; Ohl *et al.*, 2001). In addition, although categorization is often observed to be related to long-term memory representations, it can nevertheless operate automatically on certain novel stimulus features, resulting in temporary representations stored in short-term categorical memory. (See Sec. IV B 4.)

This short-term categorical store is to be distinguished from (continuous) sensory memory. A model that contains only sensory memory would fail to predict the differential experimental outcomes between speech (ES) and nonspeech (ENS) stimuli for the English listeners. In the former case, we observe no CP effects; in the latter, we observe a quasi-categorical effect. Thus, separate memory processes must be recruited for encoding continuous and categorical information. Furthermore, this short-term categorical store is to be distinguished from categorical representations stored in long-term memory. A model that contains only a long-term categorical store cannot account for the quasicategorical effect of ENS since English listeners have no previous exposure to Mandarin tones. More important, our proposed short-term categorical memory differs from the traditional view of this store that serves only as an intermediate buffer between sensory and long-term memory (Atkinson and Shiffrin, 1968). We claim that it can operate in parallel with fine-grain sensory memory and produce a CP effect independent of long-term memory. To account for our data, we introduce a new multistore model including sensory, short-term, and long-term memory components (Fig. 5).

a. Sensory memory. Two separate sensory stores have been proposed for auditory memory (Cowan, 1984, 1987). According to Cowan (1984), the *sensory memory trace* is derived from unanalyzed raw sensory data with possible temporal integration. Its lifetime has been estimated to be about 300 ms (i.e, cannot last beyond the duration of one trial in delayed discrimination tasks with parameters that are comparable to those used in this experiment). The *analyzed sensory memory* contains fine-grain analyzed sensory codes including steady-state (e.g., pitch height), time-varying (e.g., pitch slope), and event-timing (e.g., onset time or duration) information. Its lifetime is on the order of seconds. This relatively longer sensory store is necessarily required for the context-coding in the Durlach-Braida model (1969).

b. Short-term categorical memory. As hypothesized, this memory store captures only those critical features of the stimuli that are used for perceptual categorization. The strategy of such a separate memory process is to improve computational efficiency and to reduce working memory load by omitting most of the irrelevant details in sensory inputs. The likelihood that categorical encoding runs parallel to fine-

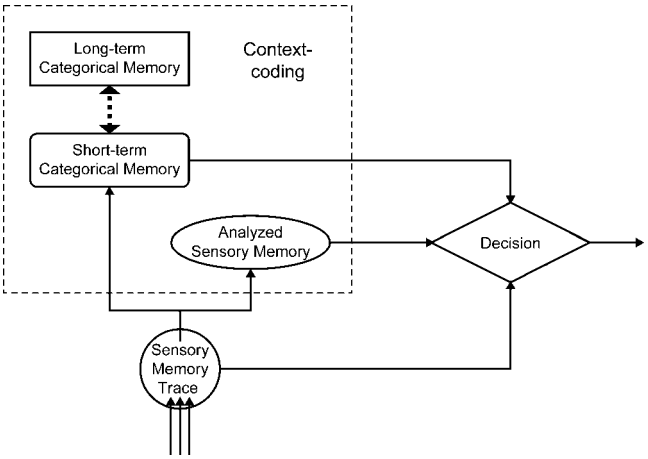

FIG. 5. A multistore model of CP. It includes four memory stores: *sensory memory trace, analyzed sensory memory, short-term categorical memory, and long-term categorical memory*. Information is encoded in a hierarchical order but short-term categorical memory and analyzed sensory memory can be processed in parallel. All the sensory and short-term categorical components are subject to memory decay. The available memory traces after decay are input for *decision-making*. If long-term categorical memory is also available, it will interact with short-term categorical memory via both top-down and bottom-up mechanisms. All the memory components with relatively longer lifetime are involved in context-coding.

grain sensory encoding is supported by duplex perception for certain speech or music stimuli (Liberman *et al.*, 1981; Pastore *et al.*, 1983). In the present experiment, we argue that this memory component is responsible for the quasicategorical effect observed in the English listeners' response to the nonspeech stimuli (ENS). Beyond our data, it also provides an explanation for other CP phenomena generated by nonspeech stimuli (Cutting and Rosner, 1974; Miller *et al.*, 1976; Pisoni, 1977). (See Sec. IV B 4)

c. Long-term categorical memory. Short-term categorical representations can be permanently preserved in long-term memory as a result of perceptual learning. These long-term categorical representations may serve as *templates* to be activated later by *bottom-up matching* of similar features. They also provide *top-down expectations* in the encoding of short-term categorical memory that allow listeners to better direct selective attention to a critical stimulus feature or dimension (Grossberg, 1980, 1999; Francis and Nusbaum, 2002). Moreover, this conversion from temporary to permanent representations increases the automation of categorical processing which, in turn, leads to a reduction of working memory load and an enhancement of categorization accuracy (Shiffrin and Schneider, 1977; Johnson and Ralston, 1994). In this experiment, we argue that this memory component is responsible for the stronger CP effects observed in Chinese listeners in both speech (CS) and nonspeech (CNS) conditions as compared to English listeners.

d. Memory variance and decision-making. Similar to the "resistance network" proposed by SDT models (van Hessen and Schouten, 1992), we hypothesize that the relative dominance of each memory resource for the input of decision-making (Fig. 5) is determined by the variance of that memory component. Memory variances may come from encoding, decay, or limited memory capacity. Since this experiment is not aimed at studying the effects of interstimulus interval (measure of *decay*) or number of categories (measure of *capacity*), our discussion is limited to the *encoding* variance related to stimulus complexity. (See Sec. IV C.)

e. Context-coding. Our model also extends the concept of context-coding (Fig. 5) that originated in the Durlach-Braida model (1969). In a CP task, subjects not only process sensory and categorical information of the target stimuli, but also integrate residual information from previous stimuli. Thus, context-coding may be involved with those memory components that span multiple stimuli and trials, i.e., analyzed sensory memory and short- and long-term categorical memory (Fig. 5). For example, in analyzed sensory memory, sensitivity decreases as a function of the width of the stimulus continuum in roving discrimination (Berliner and Durlach, 1973). In categorical memory, various factors (e.g., speaker normalization effects) may influence the location of category boundaries on a physical stimulus continuum (for a review, see Repp and Liberman, 1987). In addition, if both sensory and categorical memory resources are available in a roving discrimination task, the discrimination peak can be sharpened by the experimental context (see Sec. IV B 4 c).

4. Applications of the multistore model

Whereas this model does provide a full account of our CP data on tonal perception, it more importantly yields fresh perspectives on long-standing issues of controversy related to CP.

a. Nonspeech CP. By our model, we interpret varying degrees of CP that have been observed in some nonspeech continua, but not in others, to be attributable to stimulus-dependent categorization mechanisms based on *intrinsic* and *extrinsic* references. Both mechanisms may operate in the short-term categorical memory. An intrinsic reference is defined by the comparison of two acoustic levels changing across time (e.g., direction of pitch movement or formant transition) or the temporal order of two events inside a stimulus (e.g., temporal order of acoustic cues for voicing relative to the release of stop consonants). Categorization of intrinsic features is less demanding on working memory because computation can be carried out within each stimulus as in judgments of pitch direction herein. Most of the nonspeech continua exhibiting a CP effect have been based on acoustic features with intrinsic references (Cutting and Rosner, 1974; Miller *et al.*, 1976; Pisoni, 1977). In contrast, steady-state features such as the formant contrast of static vowels or the pitch contrast of level tones lack intrinsic references. Categorical encoding of these acoustic features is dependent on extrinsic references that are based on either a normalized acoustic level derived from other stimuli in the context or the best matched exemplar in memory. In either case, working memory load may be increased by computations required to integrate across a sequence of stimuli or to evaluate the fitness among multiple exemplars. This explains why steady-sate nonspeech continua were not categorically perceived if presented in isolation (Mirman *et al.*, 2004).

b. Degree of CP. We argue that the distinction between intrinsic and extrinsic references also determines the degrees of CP for speech continua. Macmillan *et al.* (1988), for example, found that the context variance of vowels (extrinsic reference) is up to three times larger than that for consonants (intrinsic reference). Similarly, it also explains the greater degree of CP in judgments of pitch direction (intrinsic) of linear ramps in this experiment (cf. contour tones: Wang, 1976; Francis *et al.*, 2003; Hallé *et al.*, 2004) relative to judgments of pitch height (extrinsic) of level tones (Abramson, 1979; Francis *et al.*, 2003). In agreement with the CP

data, it has also been shown that contour tones are less context-dependent than level tones (Fox and Qi, 1990; Francis *et al.*, 2003; Wong and Diehl, 2003).

c. Acquired similarity versus acquired distinctiveness. A peak in roving discrimination can emerge from simultaneously available sensory and categorical information depending on the task context (Macmillan, 1987; Kewley-Port *et al.*, 1988). If within-category trials are presented in a fixed discrimination task, a categorical distinction is not available because the stimuli always belong to the same category. In this fixed task context, discrimination relies only on sensory information. However, if such trials are presented in a roving discrimination task, they are also affected by the categorical distinction of between-category trials. If both types of information draw subjects' attention, conflicting sensory ("different") and categorical ("same") labels may be generated for two different stimuli within the same category. This conflicting information may result in less accurate within-category discrimination. In this sense, a discrimination peak is sharpened by the interdependency between the two types of trials in a roving task context. Thus, in our roving discrimination task, cross-language differences reveal that *both* an enhancement of between-category *and* a reduction of within-category discrimination contribute to the peakedness of discrimination. Such an explanation based on task context may help to resolve the persistent controversy about whether CP stems from "acquired similarity" (Kuhl, 1991) or "acquired distinctiveness" (Liberman *et al.*, 1961). Iverson and Kuhl (2000) have investigated this question by directly comparing roving and fixed discrimination tasks, and conclude that these two aspects of CP derive from independent mechanisms. However, their results are also consistent with the model proposed herein. The ratios between roving and fixed discrimination sensitivities across a vowel continuum with 60-mel formant frequency intervals indicate that within-category discrimination is significantly lower in the roving discrimination task (roving/fixed ≈ 0.6) but between-category discrimination is very close between the two tasks (roving/fixed ≈ 0.9). This result follows directly from our account that invokes a unified mechanism to explain these two aspects of CP.

C. Effects of stimulus complexity

The speech stimuli exhibit a greater complexity than the nonspeech stimuli according to at least three operational criteria. First, the existence of high-order unresolved harmonics in the speech stimuli reduces the ratio of spectral energy distribution in low-order resolved harmonics. Given the equivalent overall spectral energy between speech and nonspeech, it yields a lower pitch salience for the speech stimuli because resolved harmonics contribute more to pitch perception than do unresolved harmonics (Stagray *et al.*, 1992; Shackleton and Carlyon, 1994). Second, pitch judgments of the voice fundamental frequency (F_0) in the speech stimuli may be influenced by a pitch percept ("sibilant pitch") evoked by spectral energy allocation, which is determined by F_2 and higher formants (Traunmüller, 1987). Third, the fact that a *vowel* is perceived in the speech stimuli may interfere with pitch judgments due to a perceptual integration between segmental and suprasegmental dimensions (Carrell *et al.*, 1981; Repp and Lin, 1990).

Stimulus complexity affects the encoding variance of the memory components differentially in our multistore model. The short-term (sensory and categorical) memory components involve real-time encoding processes, and thus are subject to this type of encoding variance. With respect to sensory memory, encoding variance caused by stimulus complexity decreases overall pitch sensitivity, affecting both within- and between-category discrimination. This effect is demonstrated by lower discrimination scores and a smaller distance between obtained and predicted discrimination curves when comparing speech to nonspeech across the two language groups. With respect to short-term categorical memory, increased stimulus complexity makes it more difficult to extract categorical features and to form robust memory representations. It selectively affects between-category discrimination resulting in a reduced peakedness of discrimination. The absence of a significant peak for English listeners in the speech discrimination task (ES) may indicate that their discrimination judgments are based exclusively on continuous sensory encoding.

In contrast, the activation of permanent categorical representations does not involve any real-time encoding since they are stored *a priori* in long-term memory. Thus, this long-term component involved in CP tasks is less affected by stimulus complexity, especially when it operates dominantly in a top-down manner (Grossberg, 1999). In the current experiment, we observe that adult native Chinese listeners easily recognize pitch patterns associated with Mandarin tonal categories irrespective of stimulus set. This finding is presumably due to overlearned pitch representations that result from long-term exposure to their native language. It is therefore not surprising that we observe no effect on the peakedness of discrimination between speech and nonspeech stimuli for the Chinese listeners.

If an increase in stimulus complexity reduces the overall pitch sensitivity, we can also explain why the location of the identification boundary differs as a function of the stimulus continuum for both language groups. Although the physical step size of the speech continuum is equal to that of the nonspeech continuum, the perceptual step size of the speech continuum must be smaller due to this lowered sensitivity. We further assume that the mean decision-making criterion is constant for both stimulus sets in the identification task. Because the level end of the continuum is represented by a flat contour, there is no F_0 movement during the course of its trajectory. This stimulus is therefore likely to serve as an anchor point. As a consequence, subjects need more steps to make a "rising" response in the speech stimuli as compared to nonspeech, which results in a small boundary shift toward the rising end of the speech continuum.

D. Neurophysiological evidence for multistore memory processing

The putative memory components in our model appear to be consistent with recent findings from the brain imaging literature. A magnetoencephalography (MEG) study (Lu *et al.*, 1992) shows that the physiological lifetime of auditory sensory memory is significantly longer for auditory association cortex than for primary auditory cortex. This evidence

provides an anatomical basis for the distinction between *analyzed sensory memory* and the *sensory memory trace* proposed in our model. In another MEG experiment Luo *et al.* (2005) compared categorical versus simple auditory discrimination. They demonstrated that alpha-band activities are enhanced in auditory areas for nonspeech stimuli, and in frontal areas for both speech and nonspeech stimuli. Moreover, the alpha-band brain activity in auditory areas was stronger when directly comparing the categorical discrimination of newly learned nonspeech categories to long-term speech categories. These findings not only suggest a distinction between categorical and continuous auditory processing, but also point to different neural networks for the activation of *short-term* and *long-term categorical memory*.

V. CONCLUSION

A multistore model consisting of *unanalyzed* and *analyzed sensory memory, short-term* and *long-term categorical memory*, and parallel processing of sensory and categorical information offers a unified account of CP that explains not only the data herein but also a wide range of disparate data from the extant literature. Short-term categorical memory is hypothesized to be domain-general, inherent to the perceptual system, and separate from continuous sensory processes. Differential effects of stimulus complexity on these memory stores further support their distinctive contributions in CP tasks. Although this model is at an early stage of development, it offers promise of illuminating some of the topics of controversy surrounding CP over the past half-century.

ACKNOWLEDGMENTS

Research supported in part from the National Institutes of Health and the Purdue Research Foundation (J.G.). This article is based on part of a doctoral dissertation completed by Y.X. at Purdue University in December 2005.

Abramson, A. S. (**1979**). "The noncategorical perception of tone categories in Thai," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhman (Academic, London), pp. 127–134.

Ades, A. E. (**1977**). "Theoretical notes. Vowels, consonants, speech, and nonspeech," Psychol. Rev. **84**, 524–530.

Atkinson, R. C., and Shiffrin, R. M. (**1968**). "Human memory: A proposed system and its control processes," in *The Psychology of Learning and Motivation: Advances in Research and Theory*, edited by K. W. Spence and J. T. Spence (Academic, New York), pp. 89–195.

Berliner, J. E., and Durlach, N. I. (**1973**). "Intensity perception. IV. Resolution in roving-level discrimination," J. Acoust. Soc. Am. **53**, 1270–1287.

Best, C. T., McRoberts, G. W., and Goodell, E. (**2001**). "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," J. Acoust. Soc. Am. **109**, 775–794.

Boersma, P., and Weenink, D. (**2003**). "PRAAT: Doing phonetics by computer," (Version 4.1.12) [Computer program]. Retrieved May 8, 2003, from http://www.praat.org/.

Bornstein, M. H. (**1987**). "Perceptual categories in vision and audition," in *Categorical Perception: The Groundwork of Congnition*, edited by S. R. Harnad (Cambridge University Press, New York), pp. 287–300.

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.

Burns, E. M., and Ward, W. D. (**1978**). "Categorical perception—phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals," J. Acoust. Soc. Am. **63**, 456–468.

Carrell, T. D., Smith, L. B., and Pisoni, D. B. (**1981**). "Some perceptual dependencies in speeded classification of vowel color and pitch," Percept. Psychophys. **29**, 1–10.

Cowan, N. (**1984**). "On short and long auditory stores," Psychol. Bull. **96**, 341–370.

Cowan, N. (**1987**). "Auditory sensory storage in relation to the growth of sensation and acoustic information extraction," J. Exp. Psychol. Hum. Percept. Perform. **13**, 204–215.

Cowan, N. (**1988**). "Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system," Psychol. Bull. **104**, 163–191.

Cutting, J. E. (**1982**). "Plucks and bows are categorically perceived, sometimes," Percept. Psychophys. **31**, 462–476.

Cutting, J. E., and Rosner, B. S. (**1974**). "Categories and boundaries in speech and music," Percept. Psychophys. **16**, 564–570.

Durlach, N. I., and Braida, L. D. (**1969**). "Intensity perception. I. Preliminary theory of intensity resolution," J. Acoust. Soc. Am. **46**, 372–383.

Fox, R., and Qi, Y. Y. (**1990**). "Context effects in the perception of lexical tone," J. Chin. Linguist. **18**, 261–283.

Francis, A. L., and Ciocca, V. (**2003**). "Stimulus presentation order and the perception of lexical tones in Cantonese," J. Acoust. Soc. Am. **114**, 1611–1621.

Francis, A. L., Ciocca, V., and Ng, B. K. (**2003**). "On the (non)categorical perception of lexical tones," Percept. Psychophys. **65**, 1029–1044.

Francis, A. L., and Nusbaum, H. C. (**2002**). "Selective attention and the acquisition of new phonetic categories," J. Exp. Psychol. Hum. Percept. Perform. **28**, 349–366.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (**2001**). "Categorical representation of visual stimuli in the primate prefrontal cortex," Science **291**, 312–316.

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. (**1962**). "The identification and discrimination of synthetic vowels," Lang Speech **5**, 171–189.

Fujisaki, H., and Kawashima, T. (**1969**). "On the models and mechanisms of speech perception," Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo, Vol. **28**, pp. 67–73.

Fujisaki, H., and Kawashima, T. (**1970**). "Some experiments on speech perception and a model for the perceptual mechanism," Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo, Vol. **29**, pp. 207–214.

Fujisaki, H., and Kawashima, T. (**1971**). "A model of the mechanisms for speech perception-quantitative analysis of categorical effects in discrimination," Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo, Vol. **30**, pp. 59–68.

Gandour, J. (**1978**). "The Perception of Tone," in *Tone: A linguistic survey*, edited by V. Fromkin (Academic, New York), pp. 41–76.

Gandour, J. (**1983**). "Tone perception in Far Eastern languages," J. Phonetics **11**, 149–175.

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

Gerrits, E., and Schouten, M. E. (**2004**). "Categorical perception depends on the discrimination task," Percept. Psychophys. **66**, 363–376.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Greenwood, D. D. (**1961**). "Critical bandwidth and the frequency coordinates of the basilar membrane," J. Acoust. Soc. Am. **33**, 1344–1356.

Grossberg, S. (**1980**). "How does a brain build a cognitive code?," Psychol. Rev. **87**, 1–51.

Grossberg, S. (**1999**). "The link between brain learning, attention, and consciousness," Conscious Cogn **8**, 1–44.

Hallé, P. A., Chang, Y.-C., and Best, C. T. (**2004**). "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," J. Phonetics **32**, 291–453.

Harnad, S. R. (ed). (**1987**). *Categorical Perception: The Groundwork of Cognition* (Cambridge University Press, New York).

Hermes, D. J., and van Gestel, J. C. (**1991**). "The frequency scale of speech intonation," J. Acoust. Soc. Am. **90**, 97–102.

Herrnstein, R. J., and Loveland, D. H. (**1964**). "Complex visual concept in the pigeon," Science **146**, 549–551.

Howie, J. (**1976**). *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge University Press, Cambridge).

Iverson, P., and Kuhl, P. K. (**2000**). "Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common

mechanism?," Percept. Psychophys. **62**, 874–886.
Johnson, K., and Ralston, J. V. (**1994**). "Automaticity in speech perception: Some speech/nonspeech comparisons," Phonetica **51**, 195–209.
Kewley-Port, D., Watson, C. S., and Foyle, D. C. (**1988**). "Auditory temporal acuity in relation to category boundaries; Speech and nonspeech stimuli," J. Acoust. Soc. Am. **83**, 1133–1145.
Kuhl, P. K. (**1981**). "Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories," J. Acoust. Soc. Am. **70**, 340–349.
Kuhl, P. K. (**1991**). "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not," Percept. Psychophys. **50**, 93–107.
Kuhl, P. K., and Miller, J. D. (**1975**). "Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants," Science **190**, 69–72.
Kuhl, P. K., and Padden, D. M. (**1983**). "Enhanced discriminability at the phonetic boundaries for the place feature in macaques," J. Acoust. Soc. Am. **73**, 1003–1010.
Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (**2005**). *Applied Linear Statistical Models* (McGraw-Hill Irwin, Boston).
Liang, K. Y., and Zeger, S. L. (**1996**). "Longitudinal data analysis using general linear models," Biometrika **73**, 13–22.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**, 431–461.
Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (**1957**). "The discrimination of speech sounds within and across phonemic boundaries," J. Exp. Psychol. **54**, 358–368.
Liberman, A. M., Harris, K. S., Kinney, J. A., and Lane, H. (**1961**). "The discrimination of relative onset-time of the components of certain speech and non-speech patterns," J. Exp. Psychol. **61**, 379–388.
Liberman, A. M., Isenberg, D., and Rakerd, B. (**1981**). "Duplex perception of cues for stop consonants: Evidence for a phonetic mode," Percept. Psychophys. **30**, 133–143.
Lu, Z. L., Williamson, S. J., and Kaufman, L. (**1992**). "Human auditory primary and association cortex have differing lifetimes for activation traces," Brain Res. **572**, 236–241.
Luo, H., Husain, F. T., Horwitz, B., and Poeppel, D. (**2005**). "Discrimination and categorization of speech and non-speech sounds in an MEG delayed-match-to-sample study," Neuroimage **28**, 59–71.
Macmillan, N. A. (**1987**). "Beyond the categorical/continuous distinction: A psychophysical approach to processing modes," in *Categorical Perception: The Groundwork of Cognition*, edited by S. R. Harnad (Cambridge University Press, New York), pp. 53–85.
Macmillan, N. A., and Creelman, C. D. (**1991**). *Detection Theory: A User's Guide* (Cambridge University Press, Cambridge, UK).
Macmillan, N. A., Goldberg, R. F., and Braida, L. D. (**1988**). "Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua," J. Acoust. Soc. Am. **84**, 1262–1280.
Macmillan, N. A., Kaplan, H. L., and Creelman, C. D. (**1977**). "The psychophysics of categorical perception," Psychol. Rev. **84**, 452–471.
Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., and Dooling, R. J. (**1976**). "Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception," J. Acoust. Soc. Am. **60**, 410–417.
Mirman, D., Holt, L. L., and McClelland, J. L. (**2004**). "Categorization and discrimination of nonspeech sounds: Differences between steady-state and rapidly-changing acoustic cues," J. Acoust. Soc. Am. **116**, 1198–1207.
Ohl, F. W., Scheich, H., and Freeman, W. J. (**2001**). "Change in pattern of ongoing cortical activity with auditory category learning," Nature (London) **412**, 733–736.
Pastore, R. E., Ahroon, W. A., Baffuto, K. J., Friedman, C., Puleo, J. S., and Fink, E. A. (**1977**). "Common-factor model of categorical perception," J. Exp. Psychol. Hum. Percept. Perform. **3**, 686–696.
Pastore, R. E., Schmuckler, M. A., Rosenblum, L., and Szczesiul, R. (**1983**). "Duplex perception with musical stimuli," Percept. Psychophys. **33**, 469–474.
Pisoni, D. B. (**1973**). "Auditory and phonetic memory codes in the discrimination of consonants and vowels," Percept. Psychophys. **13**, 253–260.
Pisoni, D. B. (**1975**). "Auditory short-term memory and vowel perception," Mem. Cognit. **3**, 7–18.
Pisoni, D. B. (**1977**). "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops," J. Acoust. Soc. Am. **61**, 1352–1361.
Pisoni, D. B., and Lazarus, J. H. (**1974**). "Categorical and noncategorical modes of speech perception along the voicing continuum," J. Acoust. Soc. Am. **55**, 328–333.
Pollack, I., and Pisoni, D. B. (**1971**). "On the comparison between identification and discrimination tests in speech perception," Psychonomic Sci. **24**, 299–300.
Repp, B. H., and Liberman, A. M. (**1987**). "Phonetic category boundaries are flexible," in *Categorical Perception: The Groundwork of Cognition*, edited by S. R. Harnad (Cambridge University Press, New York), pp. 89–112.
Repp, B. H., and Lin, H. B. (**1990**). "Integration of segmental and tonal information in speech perception: A cross-linguistic study," J. Phonetics **18**, 481–495.
Rosen, S., and Howell, P. (**1987**). "Auditory, articulatory and learning explanations of categorical perception in speech," in *Categorical Perception: The Groundwork of Cognition*, edited by S. R. Harnad (Cambridge University Press, New York), pp. 113–160.
Schneider, W., Eschman, A., and Zuccolotto, A. (**2002**). *E-Prime Reference Guide* (Psychology Software Tools Inc., Pittsburgh).
Shackleton, T. M., and Carlyon, R. P. (**1994**). "The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination," J. Acoust. Soc. Am. **95**, 3529–3540.
Shiffrin, R. M., and Schneider, W. (**1977**). "Controlled and automatic human information processing. II. Perceptual learning, automatic attending, and a general theory," Psychol. Rev. **84**, 127–190.
Stagray, J. R., and Downs, D. (**1993**). "Differential sensitivity for frequency among speakers of a tone and a nontone language," J. Chin. Linguist. **21**, 144–162.
Stagray, J. R., Downs, D., and Sommers, R. K. (**1992**). "Contributions of the fundamental, resolved harmonics, and unresolved harmonics in tone-phoneme identification," J. Speech Hear. Res. **35**, 1406–1409.
Stevens, K. N. (**1981**). "Constraints imposed by the auditory system on the properties used to classify speech sounds," in *The Cognitive Representation of Speech*, edited by T. F. Myers, J. Laver, and J. Anderson (North-Holland, Amsterdam).
Stevens, K. N., Liberman, A. M., Studdert-Kennedy, M., and Ohman, S. E., (**1969**). "Crosslanguage study of vowel perception," Lang Speech **12**, 1–23.
Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., and Cooper, F. S. (**1970**). "Theoretical notes. Motor theory of speech perception: A reply to Lane's critical review," Psychol. Rev. **77**, 234–249.
Tanner, W. P. Jr., and Rivette, G. L. (**1964**). "Experimental study of "tone deafness"," J. Acoust. Soc. Am. **36**, 1465–1467.
Traunmüller, H. (**1987**). "Some aspects of the sound of speech sounds," in *The Psychophysics of Speech Perception*, edited by M. E. Schouten (Martinus Nijhoff, Dordrecht), pp. 293–305.
Treisman, M., Faulkner, A., Naish, P. L., and Rosner, B. S. (**1995**). "Voice-onset time and tone-onset time: The role of criterion-setting mechanisms in categorical perception," Q. J. Exp. Psychol. A **48**, 334–366.
Valbret, H., Moulines, E., and Tubach, J. P. (**1992**). "Voice transformation using PSOLA," Speech Commun. **11**, 513–546.
van Hessen, A. J., and Schouten, M. E. (**1992**). "Modeling phoneme perception. II. A model of stop consonant discrimination," J. Acoust. Soc. Am. **92**, 1856–1868.
Wang, W. S.-Y. (**1976**). "Language change," Ann. N.Y. Acad. Sci. **208**, 61–72.
Wong, P. C., and Diehl, R. L. (**2003**). "Perceptual normalization for inter- and intratalker variation in Cantonese level tones," J. Speech Lang. Hear. Res. **46**, 413–421.