Perception & Psychophysics
1997, 59 (2), 165-179

Lexical tone in Cantonese spoken-word processing

ANNE CUTLER
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

and

HSUAN-CHIH CHEN
Chinese University of Hong Kong, Hong Kong

In three experiments, the processing of lexical tone in Cantonese was examined. Cantonese listeners more often accepted a nonword as a word when the only difference between the nonword and the word was in tone, especially when the *F*0 onset difference between correct and erroneous tone was small. *Same–different* judgments by these listeners were also slower and less accurate when the only difference between two syllables was in tone, and this was true whether the *F*0 onset difference between the two tones was large or small. Listeners with no knowledge of Cantonese produced essentially the same *same-different* judgment pattern as that produced by the native listeners, suggesting that the results display the effects of simple perceptual processing rather than of linguistic knowledge. It is argued that the processing of lexical tone distinctions may be slowed, relative to the processing of segmental distinctions, and that, in speeded-response tasks, tone is thus more likely to be misprocessed than is segmental structure.

Sounds produced by the human voice vary along multiple dimensions. Languages use these dimensions in different ways to distinguish utterances. In particular, there are wide differences from one language to another in the "suprasegmental," or prosodic, features: variations in fundamental frequency, amplitude, and duration, which are not a function of intrinsic characteristics of phonetic segments. Lexical stress and lexical tone are the two principal methods by which languages use prosodic features to distinguish one word from another.

In tone languages, a lexically distinctive function is served by the fundamental frequency (*F*0) level or contour realized on a syllable. Thus, the Cantonese consonant–vowel (CV) sequence [si] with the high falling Tone 1 means "poem," with the middle rising Tone 2 means "history," with a low-level Tone 6 means "time," and so on. (There are six tones in Cantonese, of which three have additional abbreviated versions realized only on short syllables with a voiceless stop coda.)

In stress languages, stressed syllables may be distinguished from unstressed syllables in duration, amplitude, *F*0 movement, and segmental structure. In many stress languages, stress position within the word is fixed, and, hence, stress is not lexically distinctive. Where stress is lexically distinctive, such as in English, Dutch, and Russian, it is only infrequently the case that the prosodic features alone accomplish the lexical distinction. Thus, in English, the distinction between *SUBject* and *subJECT* (uppercase representing a stressed syllable) or between *CONtents* and *conTENTS* involves vowel differences in the initial syllable, as well as prosodic intersyllable differences. Pairs such as *FORbear* and *forBEAR* or *FOREgoing* and *forGOing*, in which the vowels do not differ across stress versions, are rare.

Both tone and stress are realized principally on the portion of a syllable that most readily allows variation in *F*0, amplitude, and duration—namely, on the quasi-steady-state portion, the vocalic nucleus of the syllable. Perception of the prosodic features is closely involved with perception of the vowel on which they are realized.

The perceptual question with which the present study is concerned is the processing of tonal and segmental information in the recognition of spoken Cantonese. This is related to the important theoretical question of the role of prosodic features in word recognition. The recognition of spoken words is, above all, a very efficient process. Longer words can often be effectively recognized before their ends (Marslen-Wilson & Welsh, 1978), and coarticulatory information in one segment can be used to predict a following segment (e.g., listeners can tell that they are hearing *can* and not *cat* in the vowel, before the final segment has begun; Ellis, Derbyshire, & Joseph, 1971).

This research was supported by Earmarked Grants from the Research Grants Council of Hong Kong, by a Visiting Scholarship to the first author from the Chinese University of Hong Kong, and by a Visiting Fellowship to the second author from the Max Planck Institute for Psycholinguistics. The authors are grateful to Connie Ho, Kit-Kan Tang, Lai-Hung Au Yeung, Siu-Lam Tang, Him Cheung, Ching Yee Chow, and Arie van der Lugt for their extensive assistance during various phases of these experiments. The authors also thank an anonymous reviewer, Denis Burnham, and Laura Walsh Dickey for useful comments on an earlier version of the paper, and Inge Doehring and Rian Zondervan for graphical assistance. Correspondence should be addressed to A. Cutler, Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands (e-mail: anne.cutler@mpi.nl).

In English, the experimental evidence suggests that—possibly for the very reason that maximum efficiency is aimed for—prosodic stress information is not exploited prelexically. That is, the process of lexical access (achieving contact with an entry or entries in the mental lexicon) operates without the use of purely prosodic information. For instance, word recognition cannot be facilitated by prior information about stress pattern (Cutler & Clifton, 1984). And listeners who hear either *FORbear* or *forBEAR* in a sentence show speeded recognition of words related to either of them (i.e., either *ancestor* or *tolerate*; Cutler, 1986), suggesting that both the lexical entry for *FORbear* and the lexical entry for *forBEAR* have been activated by the input, in just the same way as the lexical entries for a homophone such as *sale/sail* are both activated when either one is heard (Swinney, 1979).

Cutler (1986) argued that recognition efficiency in English could be served by omitting prosodic information from the prelexical access code because, in the case of stress information, the prosodic information is relative: Stressed syllables do not have an absolute level of duration, amplitude, or pitch movement but rather have just more of each than unstressed syllables do. Thus, hearing an initial syllable *for-* is not necessarily sufficient to inform the listener whether or not that syllable bears primary stress; unambiguous recognition is only possible once the second syllable has arrived. Thus, if listeners were to base their access of lexical entries on prosodic information in words such as *FORbear/forBEAR*, they would have to delay initiation of the lexical access procedure until information about the word's second syllable had arrived, and this delay would be inconsistent with maximum efficiency.

In fact, because minimal pairs such as *FORbear/forBEAR* are very rare, and because most pairs of words that vary in stress also vary in segmental structure, the omission of prosodic information from the lexical access code in English would carry remarkably little cost. Specifically, the language would have a few more homophones; *FORbear/forBEAR* and its dozen or so fellow minimal stress pairs would join the huge set of existing homophones such as *sale/sail*. For the recognition of most words, segmental information would suffice to compute a unique code for accessing the appropriate lexical entry. Indeed, studies of the effects of mis-stressing on word recognition in English suggest that listeners are more sensitive to changes in vowels than to changes in stress pattern per se. Thus in "elliptic speech" (speech in which certain speech sounds are systematically distorted), the distortion that most disrupts word recognition is alteration of vowels in stressed syllables (Bond, 1981); word recognition is slowed to a far greater extent by mis-stressing that involves changing vowel quality (e.g., *wallET, DEceit*) than by mis-stressing that involves no vowel quality change (e.g., *nutMEG, TYphoon*; Cutler & Clifton, 1984). Recognition of noise-masked words is not significantly affected by mis-stressing as long as vowel quality is unaltered (Slowiaczek, 1990). Furthermore, English language users appear to prefer to categorize vowels along a vowel quality dimension (full vs. reduced) over a prosodic dimension (stressed vs. unstressed; Fear, Cutler, & Butterfield, 1995).

In a tone language such as Cantonese, however, tonal distinctions between words are pervasive. The situation of the English listener, who can afford simply to ignore prosodic information in computing the prelexical access code, in no way resembles that of the Cantonese listener, for whom prosodic information is constantly decisive in word identification. There is, in fact, very little experimental evidence as yet available on how lexical tone information is processed in spoken-word recognition. There is, of course, clear evidence from standard word recognition paradigms that listeners use tonal information to determine word identity. Fox and Unkefer (1985) conducted a categorization experiment in which a continuum was constructed varying from one tone of Mandarin to another. The crossover point at which listeners switched from reporting one tone to reporting the other shifted as a function of whether the CV syllable upon which the tone was realized formed a real word when combined only with one tone or only with the other tone (in comparison with control conditions in which both tones, or neither tone, formed a real word in combination with the CV). This effect of word/nonword status also appears with ambiguous consonants (e.g., a continuum from [d] to [t]) in CVC syllables, both in word-initial position (Ganong, 1980) and word-final position (McQueen, 1991). Indeed, it also appears when the manipulation determining word versus nonword status is stress pattern (*TIgress* vs. *diGRESS*; Connine, Clifton, & Cutler, 1987). Given the evidence cited above that lexical stress information is not used prelexically, Fox and Unkefer's result cannot be considered evidence of precisely how tone information is processed.

Nevertheless, there are some intriguing suggestions that the processing of tonal information may cause the listener more difficulties than may the processing of segmental information. All of these, as it happens, come from studies with Chinese languages. For instance, in a study by Tsang and Hoosain (1979), Cantonese subjects heard sentences presented at a fast rate and were required to choose between two transcriptions of what they had heard; the transcriptions differed only in one character, representing a single difference of one syllable's tone, vowel, or tone+ vowel. Accuracy was significantly greater for vowel differences than for tone differences, and tone+ vowel differences were not significantly more accurately distinguished than were vowel differences alone. Taft and Chen (1992) found that homophone judgments for written characters in Mandarin were made less rapidly and less accurately when the pronunciation of the two characters differed only in tone, as opposed to in vowel; the response time difference (though not the accuracy difference) was replicated in a second experiment in Cantonese. Repp and

Lin (1990) asked Mandarin listeners to categorize nonword CV syllables according to consonant, vowel, or tone; the tonal categorizations were made less rapidly than were the segmental decisions. All of these results might be regarded as unexpected, given how important tone information is for lexical identification in Chinese languages. (Note that Repp and Lin, in fact, argue that the tonal decisions in their experiment could not be made as rapidly as the segmental decisions because of the way their synthetic materials were constructed; but it is an interesting question whether the materials could have been constructed in a way in which both kinds of decision could have been equally rapid.)

The phonetic literature does contain a number of studies on the processing of cues to tone identification in Chinese. Lin and Repp (1989), for example, report that identification of Taiwanese tones is based almost solely on the processing of *F*0 (height and movement), although there are, in Taiwanese (as in Cantonese; Kong, 1987), correlations between tone and syllable duration. Gandour (1981) similarly claims that three dimensions of *F*0 are involved in tone identification in Cantonese: *F*0 contour, direction, and height. However, Whalen and Xu (1992), in a study of Mandarin, found that amplitude information could be exploited for tone identification when *F*0 information was removed (only the relatively similar Mandarin Tones 2 and 3 proved difficult to discriminate in this way). In fact, Shen and Lin (1991) studied Mandarin Tones 2 and 3, both of which end in a rise, and report that they are distinguished by the timing of the *F*0 turning point within the syllable. Thus, it is clear that tone identification in Chinese languages normally involves the processing of *F*0, and it is possible that the processing may involve more than one dimension. The processing of tone, it is clear, is certainly no less complex than the processing of segmental information.

In the present study, we used speeded-response tasks to undertake a direct comparison of the perceptual processing of tonal versus segmental information in Cantonese syllables. Our initial experiment employed one of the simplest spoken-word recognition tasks: lexical decision. Listeners were asked to judge whether or not a spoken disyllable was a real word of Cantonese. The crucial items were, however, nonwords (i.e., items that required a "no" response in this task). These items were constructed from real Cantonese words by making some alteration in each case in the second syllable: the onset of the second syllable, its rime, its tone, or any combination of these elements could be altered. If, as the evidence from the experiments of Tsang and Hoosain (1979), Taft and Chen (1992) and Repp and Lin (1990) suggests, listeners process tonal information less rapidly or less accurately than segmental information, then we would expect that nonwords that differ only in tone from a real word would be more likely to elicit a false-positive "yes" response or would be slower to elicit a correct rejection than would nonwords that differ from real words in some aspect of their segmental structure.

EXPERIMENT 1

Method

Materials. Twelve sets of eight disyllabic items were used as the main stimuli in the experiment. They are listed in Appendix A. Each set of items was formed by using one disyllabic word to generate seven disyllabic nonwords. This was done by systematically varying the syllabic components of one syllable of the original word. Syllables in Cantonese are traditionally described as having two parts: initials and finals, corresponding to the linguistic constructs onset and rime. Onsets may be null or may be singleton consonants. Rimes may be V, VV, or VC (where C can only be a nasal or a voiceless stop). We varied the three components onset, vowel, and tone of the second syllable of the original word, such that the second syllable of the resulting items differed from that of the original word in one or more of these components (in fact, in 11 of 12 cases, the rime difference was a vowel difference; in the remaining item set—Item 9 in Appendix A—the rime difference was in the syllabic coda.). Table 1 illustrates the results for one such set of words. All the modified second syllables were existing syllables in Cantonese, but none could go with the first syllable to form a disyllabic word.

These stimuli were formed into four blocks of 42 items each. Each block of stimuli was generated using three sets of disyllabic items (including 3 disyllabic words and 21 disyllabic nonwords). To even up the number of word and nonword items, each of the 3 words was repeated seven times, so that there were 21 word items in each block.

Finally, a further set of 14 disyllabic items—7 words and 7 nonwords—was constructed for use as practice stimuli. All disyllables were recorded by a female native speaker of Cantonese. They were digitized with a sampling rate of 22 kHZ using the SoundEdit program and stored on a Macintosh IIsi computer.

Each disyllable was spoken naturally in the recording, rather than being combined from tokens of the individual component syllables. Such artificially produced combinations would not sound like naturally spoken words (even in the case of the real words), and this could lead to a change in the subjects' lexical decision criterion. However, it was necessary to ensure that the segmental and tonal properties of the nonword disyllables were indeed perceptible as intended. To ascertain this, two control pretests were carried out. In the first (single-syllable identification), the disyllables were edited into their individual component syllables and presented to 10 native speakers of Cantonese, who were asked to write a corresponding character (recall that all syllables were existing syllables of the language). This pretest has the advantage that it is a conceptually simple task for Cantonese listeners; but it has the disadvantage that the edited syllables will have suffered a loss in naturalness, which is likely to lead to errors (in particular, for the initial syllables; second syllables should suffer less since they will be slightly longer due to final lengthening effects). Accordingly, in a second pretest (disyllable recognition), the disyllables were presented to 10 native listeners, who were asked to listen to each spoken item, decide what they had heard, and then judge whether their percept matched the sound of two characters presented subsequently on cards.

Table 1
Sample Stimuli Used in Experiment 1

Mismatch	Example	Correct Response
None	/bok8-si6/	word
Tone	/bok8-si2/	nonword
Vowel	/bok8-sy6/	nonword
Vowel–tone	/bok8-sy2/	nonword
Onset	/bok8-ji6/	nonword
Onset–tone	/bok8-ji2/	nonword
Onset–vowel	/bok8-jy6/	nonword
Onset–vowel–tone	/bok8-jy2/	nonword

Table 2 presents the results of the pretests. In the single-syllable identification test, each character written by the subjects was compared with the spoken item on onset, vowel, and tone. The overall percent correct for first syllables was 68.3% and for second syllables 75.8%. In second syllables, the mean percent correct for onset was 89.4%, for vowel 93.5% and for tone 89.5%; an analysis of variance (ANOVA) across items showed no significant difference between these three properties [$F(2,176) < 1$]. In the disyllable recognition test, the mean percent correct was 82%; an ANOVA across items showed no significant difference between the seven conditions [$F(6,54) = 1.44$]. We concluded that the items were perceptible as intended and, importantly, that there was no asymmetry in the respective perceptibility of the onsets, vowels, and tones.

Subjects. Sixteen subjects were recruited from the introductory psychology subject pool at the Chinese University of Hong Kong. All subjects were native speakers of Cantonese, and none reported a history of hearing loss or speech disorder. None had taken part in either control pretest.

Procedure. The subjects were tested individually in a quiet room. They heard the stimuli at a comfortable level through Sound MD-802A headphones. They were asked to judge whether or not each presented disyllable was a legal word by pressing one of two keys on the keyboard of a Macintosh computer and to respond as quickly and accurately as possible.

The experiment included a practice session followed by four experimental sessions. The practice session consisted of 14 trials, and each experimental session involved 42 trials. Each trial started with the presentation of a short (300-msec) warning tone followed by a 400-msec pause. Immediately after the pause, a disyllabic item was presented. The subjects were allowed 2 sec to respond after the presentation of each item. A new trial would start at the end of this period, unless the subject made a response within this period; in the latter case, a new trial would start after a postresponse pause of 1 sec. The order of the four experimental sessions was counterbalanced across subjects. However, the order of presentation for the trials within each session was randomized for each subject. The whole experiment lasted about 30 min. Timing and response collection was under the control of the Macintosh IIsi computer running the Psyscope experimental control program.

Results and Discussion

Mean response times (RTs), measured from item offset, and mean error frequencies in each condition were calculated for each subject and for each item, and both RT and error measures were subjected to separate ANOVAs with subjects and items as random factors. We, in fact, carried out all analyses in two versions, one including all items sets and another omitting Item Set 9. The pattern of results was identical in both versions, and we will report only the analysis across all items. We will further report only results that were significant in both subjects and items analyses. For the RT analyses, missing data points were replaced by the mean for that subject or that item in the relevant condition.

As expected, "yes" responses (mean RT = 257 msec) were significantly faster than "no" responses (mean RT = 435 msec) [$F1(1,15) = 29.37$, $p < .001$; $F2(1,11) = 119.12$, $p < .001$]. The overall mean error rate was not high (6.8%) and did not differ significantly between "yes" (6.3%) and "no" (7.3%) responses (both Fs < 1).

The mean RTs and error rates for each of the seven mismatch conditions are shown in Table 3. Since the results of interest here concern these seven conditions, all further analyses omitted the real-word items.

An overall ANOVA revealed no significant effect in the RTs but did reveal a significant difference between the seven mismatch conditions in error rate [$F1(6,90) = 5.13$, $p < .001$; $F2(6,66) = 2.41$, $p < .04$]. We conducted multiple post hoc comparisons on each possible pairing of conditions to examine the components of this significant main effect, computing the Studentized range statistic q for each comparison. A conventional way of presenting the results of such multiple comparisons (Winer, 1972, p. 84) is to list the values in ranked order and draw an association line under any set of values between which there are no statistically significant differences. Such a presentation is given in Table 4.

It can be seen that, when only tone differed, the error rate was higher than in any other condition. When only vowel differed, the error rate was higher than in any other condition except tone. When both onset and tone differed, the error rate was lower than in any other condition. The remaining four conditions were statistically indistinguishable.

Thus, the results of Experiment 1 did indeed show a difference between the dimensions along which a nonword can deviate from a real word. There was no difference in the speed with which the nonwords could be correctly rejected (which is difficult to interpret since the extensive homophony of the syllables made it impossible to control the degree to which the disyllabic nonwords overlapped with real words and, hence, the number of potential competitors that might have been activated); but there was a difference in the probability that nonwords would be erroneously accepted as a real word. We cannot offer an explanation for the low error rate in the onset–tone condition (especially, why it should be lower than in the onset–vowel–tone condition). But, otherwise, the re-

Table 2
Percent Correct Responses in Pretests of Experiment 1, Separately for Items Presented in Each of the Seven Mismatch Conditions

Match on:	Initial Syllable	Final Syllable, Mismatching in: Onset	Vowel	Tone	Onset–Vowel	Onset–Tone	Vowel–Tone	Onset–Vowel–Tone
				Single-Syllable Identification				
Onset	93.7	85.4	93.7	94.4	80.6	88.9	88.9	93.7
Vowel	81.9	92.4	94.4	90.3	93.7	93.1	97.2	93.7
Tone	88.9	93.1	91.7	91.7	88.9	88.2	84.0	88.9
M	68.3	75.0	80.8	81.7	64.2	72.5	76.7	80.0
				Disyllabic Recognition				
		80.8	89.2	85.0	80.0	79.2	79.2	80.8

Table 3
Mean RTs (in Milliseconds) and Mean Error Rates (%) for Each of the Seven Mismatch Conditions of Experiment 1 (Cantonese Listeners)

Mismatch	RT	Error
Onset	446	6.8
Vowel	410	9.4
Tone	427	15.1
Onset–vowel	443	6.3
Onset–tone	434	2.1
Vowel–tone	425	6.3
Onset–vowel–tone	464	5.2

sults can be simply described: Two dimensions of difference from a real word (or one dimension if it is the syllable onset) suffice to produce accurate rejection of a nonword. Either a vowel difference or a tone difference alone is more likely to be overlooked and result in a nonword being erroneously accepted as a real word; a tone difference alone is significantly more likely to be overlooked than is a vowel difference alone.

The tendency of listeners to overlook a vowel difference is interestingly consistent with recent results from other studies in different languages using different methodologies. In phoneme-monitoring experiments in English and Spanish, Cutler, van Ooijen, Norris, and Sanchez-Casas (1996) found that vowel targets were detected with relatively low accuracy. English listeners also found it easier to detect consonant targets than to detect vowel targets in Japanese despite the fact that the Japanese vowel repertoire is small and relatively distinct (Cutler & Otake, 1994). Van Ooijen (1994, 1996) further found that, when listeners were presented with mispronounced words and asked to restore them to their correctly pronounced form, they found it much easier to alter vowels than to alter consonants. (One way in which this asymmetry manifested itself was in the relative speed of vowel versus consonant changes: Given the input *shevel* and instructed to turn it into a real word by changing only one sound, listeners more rapidly found a word via a vowel change [*shovel*] than via a consonant change [*level*]. Another was in the relative accessibility of each type of change: Listeners were more likely to make an erroneous vowel change when instructed to make a consonant change than vice versa.) Thus, in a word recognition task, listeners are apparently ready to treat vowels as inherently more mutable objects than consonants. Cutler et al. (1996) argued that listeners' speech processing procedures, in fact, are adjusted to take explicit account of the intrinsic variability with which vowel tokens are realized in natural speech. The present finding of a significantly higher error rate when only a vowel was altered than when the onset or any combination of dimensions was altered is consistent with this claim that listeners treat vowels as potentially unreliable evidence.

Most interesting for the present question of interest, however, is the finding that a tone difference alone produced an error rate significantly higher than did a vowel difference alone. A subsequent analysis of the error data also showed a similar difference between vowel and tone manipulations. In this analysis, we attempted to assess the statistical significance of the effects of altering each of the three dimensions separately. To do this, we conducted *t* tests on responses collapsed across the three conditions for each dimension in which that dimension was the same for each pair versus those collapsed across the three comparable conditions in which it was different. Thus, to assess the effect of onset difference, we compared responses in the vowel, tone, and vowel–tone conditions (in all of which onset was the same in the two syllables) with responses in the onset–vowel, onset–tone, and onset–vowel–tone conditions, which differed from the first three just in adding in each case the onset difference. (Note that the onset condition alone cannot be included, since the condition from which its stimuli differ minimally is the base real word!) This analysis revealed that the subjects were significantly more likely to make an error (false-positive response) when onset was the same than when onset was different [$t1(15) = 4.88$, $p < .001$; $t2(11) = 2.27$, $p < .05$].

To assess the effect of vowel difference, we similarly compared onset, tone, and onset–tone with onset–vowel, vowel–tone, and onset–vowel–tone. This comparison revealed that it was, in fact, not significantly more likely that the subjects would err when vowel was the same than when vowel was different ($t1$ and $t2$ were both nonsignificant). However, the comparison to assess the effect of tone difference, between onset, vowel, and onset–vowel, on the one hand, and onset–tone, vowel–tone, and onset–vowel– tone, on the other, showed that it was again more likely that an error would result when tone was the same than when tone was different [$t1(15) = 2.87$, $p < .02$; $t2(11) = 2.55$, $p < .03$].

This pattern of results suggests that the listeners were indeed paying attention to the tone and, in fact, that tone alteration was capable of exercising a more consistent effect than was vowel alteration. Why then was the error rate in general so high in the condition in which only tone was altered, even in comparison with the condition in which only vowel was altered? We decided to examine the effects of manipulating tone more closely by conducting a further analysis in which we took into account the nature of the tone difference. Recall that Cantonese

Table 4
Significant Differences Between Conditions in Multiple Intercondition Comparisons in Experiment 1 (Cantonese Listeners)

Mismatch						
Tone	Vowel	Onset	Vowel–Tone	Onset–Vowel	Onset–Vowel–Tone	Onset–Tone

Note—Conditions linked by an association line do not differ statistically. Conditions not linked by an association line are significantly different at, at least, the .05 level.

has six lexical tones. Figure 1, a reanalysis of production data for a single male speaker and a single female speaker from Fok (1974) reproduced from Gandour (1983), shows that these tones are not highly distinct. Most distinct from the other tones is Tone 1, which begins high and falls; the other five tones each have their onset at a similar point on the *F*0 scale. We did not control which tone differences we used in our materials, since we were constrained by the need to choose possible syllables that in combination made nonexisting words. However, as it happened, some of our tone differences involved Tone 1 against other tones (eight items), whereas some involved two less distinct tones (four items). We predicted that the error rate in the tone condition would be higher for the latter (*hard*) group of items than for the former (*easy*) group. An unequal-*N* ANOVA across items revealed that the easy–hard comparison interacted significantly with the seven-way nonword condition factor [$F2(6,60) = 3.72, p < .005$]. In six of the conditions, the mean error rate for the easy items versus the hard items varied very little (from 4% to less than 1%), but, in the tone condition, there was a very large difference: the mean error rates for the hard and easy items were 31.4% and 7.1%, respectively.

This suggests that some tone distinctions simply cannot be made in the earliest portions of a syllable. In a speeded-response task such as auditory lexical decision, the pressure to respond quickly may encourage listeners to issue their response before the distinguishing tonal information has actually had time to arrive. Evidence from a study of Thai tonal contrasts by Burnham, Kirkwood, Luksaneeyanawin, and Pansottee (1992) is consistent with this suggestion; the order of difficulty of paired Thai tones as judged by English listeners in a *same–different* judgment task was determined by the nominal starting pitch of the tones. When the starting pitch was similar, these listeners' accuracy was little better than chance, whereas, for pairs of tones with very different starting pitch, accuracy was as high as 94%. We therefore decided in our next experiment to move to a simpler task that would allow us to assess the order in which perceptual information becomes available to listeners and the relative speed with which a fairly distinct and a fairly nondistinct tonal difference can be perceived. The task we chose was that used by Burnham et al. (1992)—namely, *same–different* judgment, which in principle requires no linguistic processing at all and certainly requires no lexical access. We asked listeners to judge, as rapidly as possible, whether two auditorily presented open syllables were the same or different. When they were different, the difference could be in any one of the three dimensions of the syllable (consonant, vowel, or tone) or any combination of these. Our materials contained two tonal distinctions. Both were between a falling tone and a rising tone, so that the syllables as a whole should be clearly distinct; but the two pairs differed in how far apart on the *F*0 scale the tones were initiated. One comparison, between Tones 1 (high falling) and 2 (middle rising), we predict to be an easy distinction for listeners to make, since Tone 1 begins at a much higher point than Tone 2 does. The other, between Tones 4 (low falling) and 5 (low rising), we predict to be harder, because the two tones begin at fairly similar points. In a speeded-response task, listeners will be more likely to overlook a difference that is not immediately available.

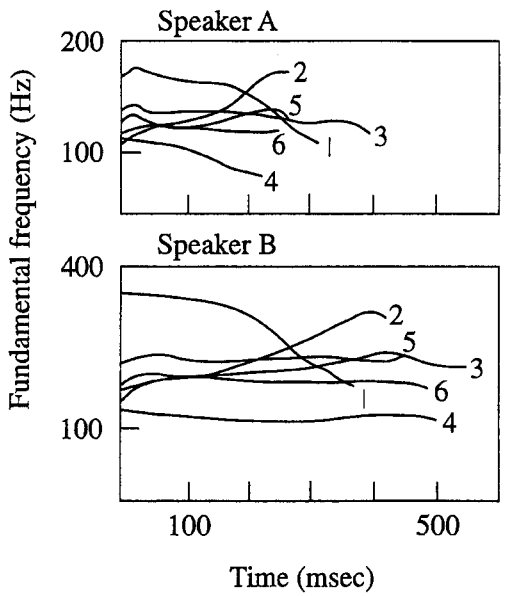

Figure 1. The six tones of Cantonese, spoken by a male (A) and a female (B) speaker. The data are from Fok (1974) as redrawn by Gandour (1983). From "Tone Perception in Far Eastern Languages," by J. Gandour, 1983, *Journal of Phonetics, 11*, p. 152. Copyright 1983 by Academic Press. Reprinted with permission.

EXPERIMENT 2

Method

Materials. Two sets of 8 real-word syllables (Cantonese syllables with corresponding Chinese characters) and two sets of 8 nonword syllables (legal but nonoccurring syllables in Cantonese) served as stimuli (all are listed in Appendix B). Only open syllables were used. Each set of word or nonword syllables was constructed by using two onsets, two vowel rimes, and two tones to compose eight possible combinations. In one each of the real-word set and one each of the nonword set, the two tones were Tones 1 and 2 (the easy distinctions); in the remaining sets, the two tones were Tones 4 and 5 (the hard distinctions). The chosen segmental contrasts also differed in intrinsic difficulty. The easy tone distinction was realized on the syllable pairs *te–gy* (phonetically [tɛ], [ky]) and *ji–sy* (phonetically [ji], [sy]); the onsets of the first pair are two voiceless stops, which (on the perceptual confusion evidence for consonants; see Miller & Nicely, 1955; Wang & Bilger, 1973) should be harder to distinguish than the onsets of the second pair, a glide and a strident fricative, whereas the vowels of the first pair involve a low-central unrounded versus high-front rounded contrast, which should be easier to discriminate than the contrast in the second pair, between two high-front vowels differing only on roundedness (note that relevant perceptual confusion evidence for these vowels is not available in the literature, although studies of American-English vowels—e.g., see Peterson & Barney, 1952, and Hillenbrand, Getty, Clark, & Wheeler, 1995—do suggest that the dimensions back–front, high–low, and rounded–unrounded determine confusion between vowels; the more dimensions of difference, the less likely two

vowels are to be confused). The hard tone distinction was realized on *fou–koe* (phonetically [*fou*], [kʻœ]) and *piu–lei* (phonetically [pʻiu], [lei]); the former onset distinction, between a nonstrident fricative and an aspirated stop, should be harder to distinguish than the latter, between a stop and a lateral, whereas the former vowel distinction, between a back diphthong moving from mid to high and a low-central monophthong, should be easier to distinguish than the latter, between two diphthongs both beginning with front unrounded vowels, one high and the other upper-mid, and both ending with high vowels.

The 32 resulting syllables were recorded by a female native speaker of Cantonese in a quiet room. Each of the syllables was spoken several times at a comfortable rate. Because of the nature of the *same–different* judgment task, we wished to ensure that there were no durational differences between items that could result in responses being issued earlier in some pairs than in others; this could have arisen because of correlations between tone and syllable duration that can occur in Cantonese (Kong, 1987). The following procedure was adopted to control stimulus duration. One token of each item, of maximally similar duration, was chosen. The tokens were then edited, using the SoundEdit program, to a constant length of about 795 msec by compressing or expanding the syllable. In no case did the durational adjustment result in a change greater than 6.25% of the original duration. Care was taken to ensure that good auditory quality of the resulting items was preserved by presenting the stimuli at a rate of one item per 5 sec to a group of 5 pilot subjects, who were asked to repeat the items they had just heard. No subject had any difficulty with any of the 32 items in this pretest. An ANOVA carried out on the measured durations of the tokens used in the experiment revealed no significant difference between tokens as a function of any of the independent variables (word–nonword status, easy vs. hard tone distinction, item identity) alone or in combination. The durations of the syllabic rimes (here, the vowel parts of the syllables; mean duration = 688 msec) were separately analyzed, and this analysis similarly revealed no significant differences as a function of any of the independent variables.

The items were digitized and stored in the same manner as for Experiment 1. Pitch analysis of the syllables was carried out using ESPS speech analysis software. Pitch traces for all 32 syllables are shown in Figure 2. It can be seen that the point of $F0$ onset for Tones 1 and 2 (the easy distinction) differs by approximately 50 Hz, whereas the point of $F0$ onset for Tones 4 and 5 (the hard distinction) is closely comparable. The word and nonword syllables were then used to assemble eight possible types of syllable pairs that involved either two identical items or two items that differed in either one or more syllabic components (i.e., onset, rime, tone), as illustrated in Table 5.

Each set of word or nonword syllables could thus make up 64 different pairs (i.e., 8 identical and 56 different pairs). In order to have equal numbers of positive and negative trials, each pair of identical syllables was repeated seven times. Consequently, for each set of word or nonword syllables, 56 positive pairs and 56 negative pairs were produced, for a total of 112 pairs. These stimulus pairs were then divided into two blocks of 56 pairs each (with 28 positive pairs and 28 negative pairs in each block), so that each unique pair of identical syllables occurred three times in one block and four times in another block, whereas each pair of different syllables appeared only once in the two blocks. In addition, order was counterbalanced across the two blocks, such that each individual syllable occurred 14 times in first position in a stimulus pair (7 times in the positive pairs and 7 times in the negative pairs) and 14 times in second position.

Subjects. Sixteen subjects were recruited from the same subject pool used in Experiment 1. All fulfilled the same criteria as the subjects of Experiment 1.

Procedure. The subjects were tested individually in a quiet room. They heard the stimuli, in pairs, at a comfortable level through Sound MD-802A headphones. All subjects heard both blocks of stimuli for all four stimulus sets. They were asked to judge whether or not the two syllables in each pair were identical, by pressing one of two keys on the keyboard of the Macintosh computer, and to respond as quickly and accurately as possible.

The experiment included a practice session and eight experimental sessions. The first session was always the practice session, which consisted of 28 trials, with 14 identical and 14 different syllable pairs. Half of the practice trials contained word syllables, and the other half contained nonword syllables, but none coincided with those used in the experimental trials. The eight experimental blocks, with 56 trials in each block, were made up of the word and nonword stimuli described above. Thus, four experimental blocks contained word syllables, and the other four contained nonword syllables. The order of blocks was counterbalanced across subjects. However, the order of presentation of trials within each block was randomized for each subject individually. The whole experiment lasted about 1 h.

Each trial started with the presentation of a short (300-msec) warning tone, followed by a 400-msec pause. Immediately after the pause, the first syllable was presented, lasting about 800 msec. At the acoustic offset of the first syllable, a 250-msec pause followed. The second syllable was then presented. The subjects were allowed 2 sec for response after the presentation of the second syllable. A new trial began at the end of this period, unless the subject made a response within this period. In the latter case, a new trial began after a pause of 1 sec. Timing and response collection was controlled as in Experiment 1.

Results and Discussion

Mean RTs and mean error frequencies were calculated, missing data points replaced, and ANOVAs conducted in the same manner as for Experiment 1.

The mean RTs (from the onset of the second member of the pair) and error rates for the seven "different" conditions are shown in Table 6, separately for the easy and hard tone comparisons. Overall ANOVAs first assessed the effects of the word–nonword and easy/hard manipulations. RTs to word stimuli (mean RT = 856 msec) were faster than RTs to nonword stimuli (mean RT = 880 msec) [$F1(1,15) = 4.92$, $p < .05$; $F2(1,28) = 4.25$, $p < .05$], but this factor did not interact with any other factors in our analyses; there was no effect of this factor in the error rates. Since the real-word syllables were also the syllables with the more easily distinguishable onsets, this effect could represent either an effect of lexical status or an effect of onset discriminability, or both (see later discussion); it is in the reverse direction for an effect of vowel discriminability.

The seven-way mismatch condition factor was significant in both RTs and error rates [RTs, $F1(6,90) = 16.22$, $p < .001$, and $F2(6,168) = 30.37$, $p < .001$; error rates, $F1(6,90) = 3.08$, $p < .01$, and $F2(6,168) = 18.67$, $p < .001$]. Responses to stimuli involving an easy tone discrimination were significantly faster (mean RT = 820 msec) than responses to stimuli involving a hard tone discrimination (mean RT = 913 msec) [$F1(1,15) = 65.82$, $p < .001$, and $F2(1,28) = 86.37$, $p < .001$], although there was no easy/hard difference in the error rates (mean error rates for easy and hard were 3.9% and 5%, respectively). More importantly, the easy/hard comparison interacted significantly in both RTs and error rates with the seven-way mismatch condition factor [RTs, $F1(6,90) = 6.9$, $p < .001$, and $F2(6,168) = 6.58$, $p < .001$; error rates, $F1(6,90) = 3.37$, $p < .005$, and $F2\,(6,168) = 4.2$, $p < .001$]. Thus, we

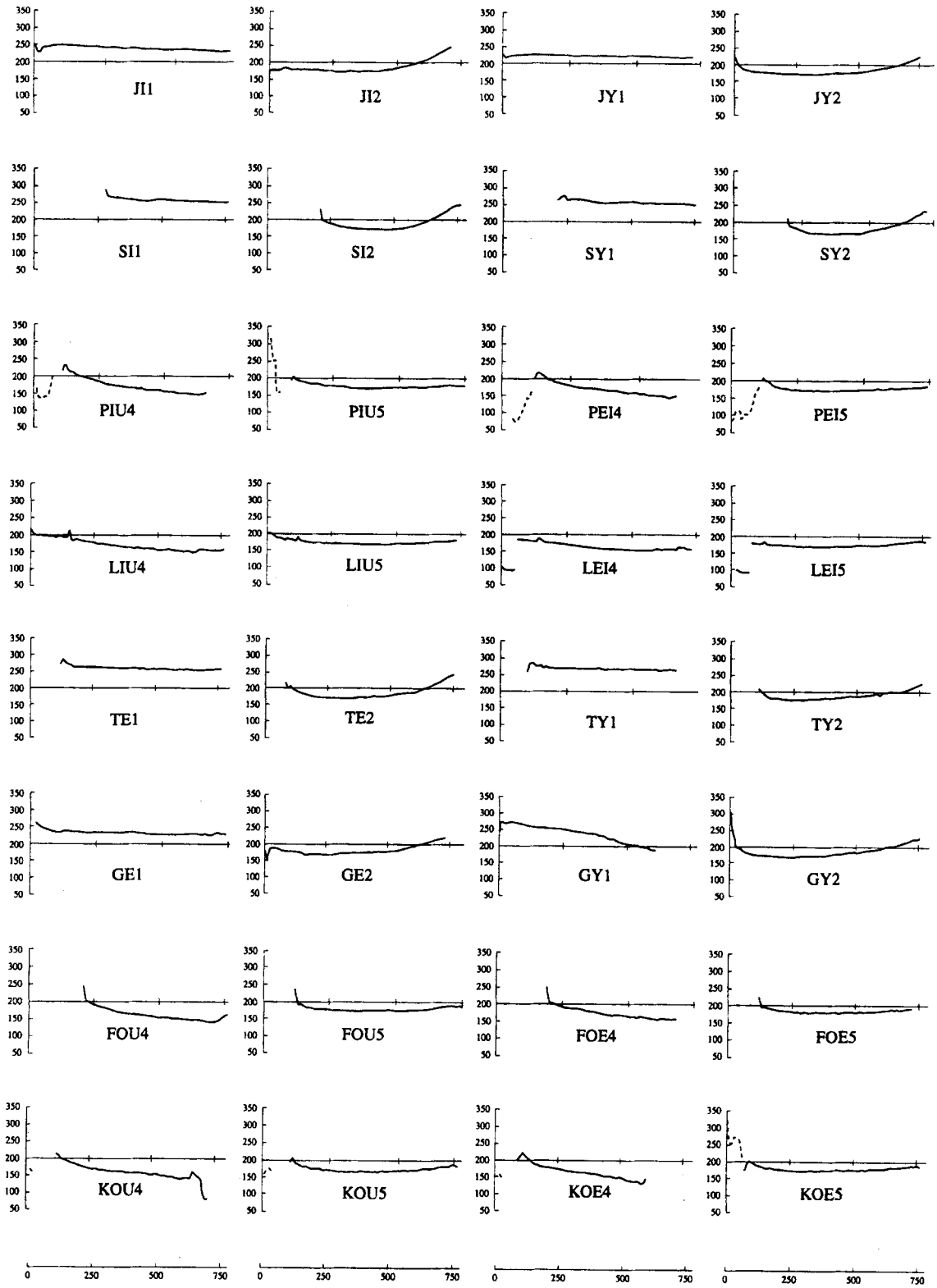

Figure 2. *F*0 traces for the 32 syllables used in Experiments 2 and 3. The vertical axis displays fundamental frequency in hertz, and the horizontal axis displays time in milliseconds. The upper eight panel pairs are real words, the lower eight are nonwords. Tone 1 versus Tone 2 comparisons have more widely separated starting frequencies and were, hence, designated *easy* comparisons. Tone 4 versus Tone 5 comparisons begin at a similar frequency and were, hence, designated *hard* comparisons.

Table 5
Sample Stimuli Used in Experiments 2 and 3

Mismatch	Example	Correct Response
None	/ji1/–/ji1/	same
Tone	/ji1/–/ji2/	different
Vowel	/ji1/–/jy1/	different
Vowel–tone	/ji1/–/jy2/	different
Onset	/ji1/–/si1/	different
Onset–tone	/ji1/–/si2/	different
Onset–vowel	/ji1/–/sy1/	different
Onset–vowel–tone	/ji1/–/sy2/	different

conducted our further analyses on the entire materials set and separately for the easy and hard items.

The further analyses of the mismatch manipulation were in two parts. The first part consisted of multiple individual intercondition comparisons, in the same manner as for Experiment 1. The pattern of all intercondition comparisons (for RTs and for errors, overall and separately for the Easy and Hard tone comparison subsets) is summarized in Table 7.

As Table 6 shows, the six conditions other than tone differed in their ordering; however, in every case, these six conditions did not differ statistically among themselves, whereas the tone condition was always significantly different (slower RTs, higher error rates) from all other conditions. In other words, it was harder to decide that two syllables were different when the only difference between them was in their tone, and this was true whether the tone distinction was easy or hard for listeners to make.

Next, as for Experiment 1, we assessed the statistical significance of the effects of altering each of the three dimensions separately, by conducting analyses in which responses were collapsed across the three conditions for each dimension in which that dimension was the same for each pair versus the three comparable conditions in which it was different. Because the four item sets differed in onset discriminability, vowel discriminability, and tone discriminability, we conducted these analyses separately for each item set. The comparison to assess the RT effect of onset difference revealed that "different" responses were significantly faster when onset differed than when onset was the same [$t1(15) = 4.49, p < .001$; $t2(7) = 7.21, p < .001$; t tests across both subjects and trials were also separately significant for all four item sets]. The same comparison for error rates was marginally significant [$t1(15) = 2.03, p < .06$; $t2(7) = 5.02, p < .002$; t tests across both subjects and trials were separately significant for two of the four item sets, *fou-koe* and *piu-lei*].

The comparison to assess the effect of vowel difference showed, similarly, that "different" responses were significantly faster when vowel was different than when vowel was the same [$t1(15) = 5.27, p < .001$; $t2(7) = 8.15$, $p < .001$; both subjects and items t tests were separately significant for the sets *ji-sy*, *fou-koe*, and *piu-lei*). Furthermore, the same comparison for error rates was marginally significant [$t1(15) = 1.82, p < .09$; $t2(7) = 7.55, p < .001$; here, t tests across both subjects and trials were separately significant only for *piu-lei*, whereas the trials t test was significant for *te-gy* and *fou-koe*).

The overall comparison to assess the effect of tone difference revealed no significant differences in either overall RT or error rate as a function of whether the tone was the same or different; t tests across both subjects and trials revealed significantly faster RTs when tone was different than when tone was the same for the *ji-sy* set only, and they revealed no other significant effects either across subjects or across trials in RTs or in errors.

Thus, although the item sets differed in the relative discriminability of the onset contrast, this variation had little effect: The onset always contributed to the responses, as measured by either RTs or errors. Similarly, despite variability in the discriminability of the vowel contrast, the vowel also contributed in either RTs or errors for each item set. The tone contrast, however, made a noticeably smaller contribution (significant only in the case of the RTs to one easy tone pair).

In conclusion, then, this simple *same–different* judgment task has revealed that differences of tone—as the earlier experiments by Tsang and Hoosain (1979), and Taft and Chen (1992) indicated—have less robust effects on processing than do segmental differences. The effects of manipulating onset and vowel in our experiment were very similar. "Different" responses were, overall, faster and more accurate when onset differed than when it did not and when vowel differed than when it did not. In contrast, tone difference alone did not lead to an increase in speed or accuracy of response; instead, the reverse was true. When the subjects were presented with a pair of syllables differing only in tone, their responses were slow, and they made a relatively high number of errors.

The effects on response accuracy have perhaps more far-reaching implications than do the effects on RT. An error in the tone condition consisted of the subject's responding "same" when the stimuli, in fact, differed; the fact that the proportion of errors in this condition was the highest of all suggests that the listeners were sometimes simply not processing the tonal information effectively. Nor was this only the case when the tone comparison was hard; in the easy subset, the error rate was highest in this condition.

Our interpretation would be that these results are evidence of the limits of tone processing and are acoustic in

Table 6
Mean RTs (in Milliseconds) and Mean Error Rates (%) for Each of the Seven Mismatch Conditions, Separately for the Easy and Hard Tonal Comparisons in Experiment 2 (Cantonese Listeners)

	Easy		Hard	
Mismatch	RT	Error	RT	Error
Onset	830	3.5	893	2.7
Vowel	844	3.1	899	3.9
Tone	864	8.2	1,062	16.8
Onset–vowel	798	2.7	869	2.7
Onset–tone	787	2.7	920	3.5
Vowel–tone	815	3.1	879	3.5
Onset–vowel–tone	800	3.5	868	2.0

Table 7
Significant Differences Between Conditions in All Multiple Intercondition Comparisons in Experiment 2 (Cantonese Listeners)

Mismatch						
Tone	Vowel	Onset	Onset–Vowel	Onset–Tone	Vowel–Tone	Onset–Vowel–Tone

Note—Conditions linked by an association line do not differ statistically. Conditions not linked by an association line are significantly different at, at least, the .05 level.

nature rather than linguistic. A simple way to test whether an effect in a language perception experiment is acoustic or linguistc in nature is to present the same input to listeners who do not know the language in question (see Cutler, Mehler, Norris, & Segui, 1987). Linguistic effects should disappear with such a subject population. Since initial auditory processing of acoustic stimuli should reflect characteristics of the human auditory system rather than effects of linguistic knowledge, however, it should be constant across listener groups, so that effects that are due to this level of processing should be maintained. In Experiment 3, therefore, we presented the materials of Experiment 2 to listeners who knew no Cantonese and were native speakers of Dutch.

EXPERIMENT 3

Method

Subjects. Seventeen subjects were recruited from the subject pool of the Max Planck Institute for Psycholinguistics. All subjects were undergraduate students at Nijmegen University and were native speakers of Dutch, and none had any knowledge of Cantonese. No subject reported a history of hearing loss or speech disorder.

Materials and Procedure. The materials were the same as those for Experiment 2. The subjects were tested in groups of up to 4 in a quiet room. Presentation and instructions were (except for the language of instruction) as in Experiment 2. Timing and response collection was under the control of a Hermac PC running the NESU experimental control program.

Results and Discussion

Mean RTs and mean error frequencies were calculated and analyzed as in Experiment 2. The mean RTs and error rates for the seven mismatch conditions are shown in Table 8, separately for the easy and hard tone comparison subsets. The word–nonword comparison was insignificant in analyses of both RTs and error rates (all Fs < 1).

There was a main effect of mismatch conditions in both RTs and errors [for RTs, $F1(6,96) = 51.87, p < .001$, and $F2(6,168) = 34.25, p < .001$; for error rates, $F1(6,96) = 29.86, p < .001$, and $F2(6,168) = 190.34, p < .001$]. Responses to easy stimuli were again significantly faster and more accurate than to hard stimuli [for RTs, $F1(1,16) = 29.01, p < .001$, and $F2(1,28) = 32.25, p < .001$; for error rates, $F1(1,16) = 42.5, p < .001$, and $F2(1,28) = 87.68, p < .001$]. Again, there was an interaction of the easy/hard factor with the seven-way mismatch factor [for RTs, $F1(6,96) = 6.86, p < .001$, and $F2(6,168) = 6.88, p < .001$; for error rates, $F1(6,96) = 31.68, p < .001$, and $F2(6,168) = 119.1, p < .001$]. As Table 8 shows, the Dutch listeners found the hard discrimination in the tone-alone condition very difficult indeed.

Further analyses of the mismatch manipulation were conducted as for Experiments 1 and 2. Multiple intercondition comparisons to examine the significant effect of the seven-way mismatch factor revealed in this case different patterns for RTs and for errors and different patterns for the easy and hard tone comparison subsets. For errors, the overall pattern for these Dutch listeners was the same as for the Cantonese listeners of Experiment 2 (i.e., the tone condition was significantly more error-prone than were any of the six other conditions, which did not differ statistically among themselves). However, the easy and hard subsets differed. For the hard subset, the overall pattern of significance was repeated (despite some difference of ordering within the six statistically equivalent conditions); however, in the easy subset, there was no significant difference between any pair of the seven conditions. Thus, for Dutch listeners, the easy tone discrimination allowed a reduction in error rate for the condition in which only tone differed across the two syllables, such that this condition was not significantly different from the other mismatch conditions.

In the RTs, the tone condition was again statistically different (slower RTs) from each of the other conditions, and, again, this was separately true for both the easy subset and the hard subset. However, the other six conditions were not in this case statistically indistinguishable. The statistical patterns of association between conditions are shown in Table 9.

In the analyses collapsing across conditions in which a given component was the same versus different, it proved significantly easier to make a "different" response when onset was different than when onset was the same [RTs, $t1(16) = 9.97, p < .001$, and $t2(7) = 14.52$,

Table 8
Mean RTs (in Milliseconds) and Mean Error Rates (%) for Each of the Seven Mismatch Conditions, Separately for the Easy and Hard Tonal Comparisons in Experiment 3 (Dutch Listeners)

	Easy		Hard	
Mismatch	RT	Error	RT	Error
Onset	712	2.2	724	1.5
Vowel	729	3.3	761	0.7
Tone	845	7.4	1,008	49.6
Onset–vowel	665	1.1	681	1.5
Onset–tone	657	1.1	730	0.7
Vowel–tone	693	0.7	750	0.7
Onset–vowel–tone	659	1.5	703	0.4

Table 9
Significant Differences Between Conditions in Multiple Intercondition Comparisons in Experiment 3 (Dutch Listeners)

Errors, Overall and Hard Tone Contrasts						
Tone	Vowel	Onset	Onset–Vowel	Onset–Tone	Vowel–Tone	Onset–Vowel–Tone
RT, Overall						
Tone	Vowel	Vowel–Tone	Onset	Onset–Tone	Onset–Vowel–Tone	Onset–Vowel
RT, Easy Tone Contrasts						
Tone	Vowel	Onset	Vowel–Tone	Onset–Vowel	Onset–Vowel–Tone	Onset–Tone
RT, Hard Tone Contrasts						
Tone	Vowel	Vowel–Tone	Onset–Tone	Onset	Onset–Vowel–Tone	Onset–Vowel

Note—Conditions linked by an association line do not differ statistically. Conditions not linked by an association line are significantly different at, at least, the .05 level.

$p < .001$; errors, $t1(16) = 5.63$, $p < .001$, and $t2(7) = 9.5$, $p < .001$]. As in Experiment 2, this effect was separately significant across both subjects and trials in the RTs for each of the item sets and in errors for the sets *fou-koe* and *piu-lei*; here, it was further significant across subjects in the errors made to the remaining two item sets.

It was also significantly easier to make a "different" response when vowel was different than when vowel was the same [RTs, $t1(16) = 7.12$, $p < .001$, and $t2(7) = 16.11$, $p < .001$; errors, $t1(16) = 5.64$, $p < .001$, and $t2(7) = 11.33$, $p < .001$]. This effect also was separately significant across both subjects and trials in the RTs for each of the item sets and in the errors for the three item sets *fou-koe*, *piu-lei*, and *te-gy*.

For the comparable analysis of tone effects, neither for RTs nor for errors did subjects or trials overall comparisons reach the .05 level of significance. The item set *ji-sy* again produced significantly faster responses when tone differed than when tone was the same, in both subjects and trials analyses, and this effect also appeared in RTs in subjects analyses only for *te-gy* and in errors in the trials analysis only for *piu-lei*.

One obvious, and perhaps unexpected, finding in Experiment 3 is that the Dutch listeners in fact performed the judgment task more rapidly than did the native Cantonese listeners of Experiment 2. An analysis combining the results of Experiments 2 and 3 revealed no significant difference between the two subject groups in error rate but did reveal an RT advantage for the Dutch listeners [$F1(1,31) = 5.87$, $p < .025$; $F2(1,28) = 426.24$, $p < .001$]. This could simply reflect the greater facility of the Dutch subject group (experienced members of the MPI subject pool) with RT experiments; or it could result from the fact that, for the Dutch listeners, all the stimuli were nonsense items, which might have encouraged them to focus attention at a relatively low processing level. However, the most important feature of the results of Experiment 3 is actually their broad similarity to the results of Experiment 2. Just as the native Cantonese speakers had responded significantly less rapidly and significantly less accurately to the stimuli in which only tone differed than to any other set of stimuli, so too did the Dutch listeners. For both subject groups, "different" responses were faster and more accurate when onset differed than when it did not and when vowel differed than when it did not. For neither subject group did the comparison of conditions in which tone was different with conditions in which tone was the same produce any overall difference in RTs or error rates. Again, all item sets, regardless of the relative discriminability of the contrasts involved, showed clear effects of onset and vowel difference, whereas comparable effects of Tone difference hardly ever appeared. Across the 48 separate such individual comparisons (4 items sets × 3 dimensions [onset, vowel, tone] × 2 dependent variables [RTs, errors] × 2 random factors [subjects, trials]), 40 patterned the same in Experiment 3 as in Experiment 2.

The subjects in Experiment 3 were not native speakers of Cantonese; indeed, they knew nothing of this language. Note that the one effect that appeared in Experiment 2 but not in Experiment 3 was the main effect of the word–nonword comparison shown only by the native speaker subjects of Experiment 2. Although this effect could have been interpreted as an effect of onset discriminability, the otherwise closely parallel results of the onset comparisons across the two experiments suggest that this asymmetry is better ascribed to the lexical knowledge of the Cantonese listeners. No word recognition was, in fact, required in the *same–different* judgment task (and, indeed, the effects of the mismatch manipulations were the same for word and nonword items for both subject groups, and the word–nonword comparison did not interact with other factors). The principal results of both experiments may therefore be presumed to owe nothing to lexical knowledge. Instead, we propose that these experiments tell us about constraints on the perceptual processing of tonal information, irrespective of whether or not the listener is accustomed to using such information in the course of lexical access.

GENERAL DISCUSSION

In three experiments, we have examined listeners' processing of lexical tone information in Cantonese. In

an auditory lexical decision task, Cantonese listeners were significantly more likely to erroneously accept a nonword as a real word when the only difference between the nonword and a real word was in the tonal value of the second syllable. Such an error was particularly probable when the *F*0 onset difference between the correct tone of the real word and the erroneous tone on the nonword was small, so that the tone distinction was, in effect, perceptually hard to make. In a *same–different* judgment task, Cantonese listeners were slower and less accurate in their responses when the only difference between two syllables was in their tonal value, and this was true whether the *F*0 onset difference rendered the distinction between the two tones perceptually easy or perceptually hard; only one perceptually easy tonal distinction produced an effect on RTs such that responses were faster when the tone differed than when it was the same. Since both the syllable onset and the rime (here, a vowel only) had clear effects of this kind in this task, it appears that only a perceptually easy tonal distinction can be as effective a discriminator as segmental distinctions. In a final experiment, the *same–different* judgment task was repeated with non-native listeners who had no knowledge of Cantonese and no experience in making lexical tone distinctions. These listeners produced a pattern of results highly similar to that produced by the native listeners: Responses were, in general, slower and less accurate when the only difference between two syllables was in their tonal value, and only the same perceptually easy tonal distinction produced a reliable effect on RTs such that responses were faster when the tone differed than when it was the same.

This pattern of results offers some clarification of the apparent puzzle provided by the findings summarized in the introduction—namely, that although tonal distinctions in a language such as Cantonese are pervasive and are necessary for successful word recognition, listeners are slower and more error-prone in utilizing tonal information than in utilizing segmental information. Our results suggest that many tonal discriminations are simply quite hard to make. In speeded-response tasks, the pressure to respond quickly shows the advantage of segmental over tonal information: In some cases, the subjects issued their response before the tonal information had been adequately processed.

In Cantonese, as Figure 1 shows, a high proportion of tonal discriminations are hard to make. This is not necessarily true for every tone language; Mandarin, for example, has four lexical tones that, at least in comparison with Cantonese, must be considered to be relatively distinct. It would be interesting to ascertain whether our present results would be replicated in full in a language such as Mandarin. The results of Taft and Chen (1992) and of Repp and Lin (1990), however, suggest that the perception of tones in Mandarin and in Cantonese is, in fact, not greatly different; recall that the former study found that tone differences alone led to difficulty in a homophone judgment task in both these languages, whereas the latter study, in which tone judgments were made less rapidly than were segmental judgments, was carried out with Mandarin listeners. Shen and Lin (1991), in fact, describe the discrimination of Mandarin tones as involving perception of the timing of the *F*0 turning point within the syllable, so that it is clear that at least some Mandarin discriminations, like Cantonese discriminations, cannot be made without a certain accumulation of information across the syllable.

What, then, do our results tell us about the role of tonal information in speech perception in this tone language? First, it is absolutely clear that the listeners in our experiments were paying full attention to the tonal information and were processing it where and as soon as they could; the very highest error rate, for lexical decisions in which the difference from a real word involved only a perceptually hard tone discrimination, was still only around 30%. When distinctive tonal information arrived early (e.g., the *F*0 onset points of Tone 1 vs. 2 in Experiments 2 and 3), it was processed more efficiently (i.e., exercised a clear effect on response patterns) than when it arrived late (Tones 4 vs. 5 in Experiments 2 and 3). However, our overall pattern of results suggests that tonal information simply does not usually arrive early. Tones are primarily realized upon vowels; therefore, they cannot be processed until the vowel information is available. Tonal information conveyed on a vowel and the vowel information itself are unlikely to be processed fully independently; classification of vowels in CV syllables is slower when the pitch of the syllable varies than when it is held constant, and, likewise, classification of pitch is slower when the vowel on which it is realized varies than when it is constant (Lee & Nusbaum, 1993; Miller, 1978; Repp & Lin, 1990). Vowels, however, can in principle be identified very early; in a CV sequence, the transition from the consonant into the vowel is enough for listeners to achieve vowel identification (Strange, 1989). In a given syllable, then, the order of arrival of the components of the syllable (as manipulated in our experiments) must be onset, then vowel, then tone.

In fact, we now know that the processing of vowels is also undertaken with some caution by listeners and that vowels in naturally spoken words are regarded as inherently mutable information sources (Cutler et al., 1996; van Ooijen, 1994, 1996). The underlying reason for this behavior on the part of listeners is taken to be the fact that the realization of vowels in natural phonetic contexts is highly variable; as a result, in computation of the lexical access code, listeners assign a lower priority to vocalic information than to consonantal information. The realization of lexical tone in natural phonetic contexts is, however, also subject to considerable variability. Contextual effects of the tone of one syllable upon the tone of an adjacent syllable (tone sandhi) may result in intertone distinctions that are quite clear in citation-form pronunciations being lost or greatly reduced in context. Thus, we may reasonably expect that listeners would exercise caution in processing natural tone information and would make contextually dependent tone identifications where required (see Speer, Shih, & Slowiaczek, 1989, for evidence that this is indeed necessary in Man-

darin). Nevertheless, we believe that our results are symptomatic of a real perceptual disadvantage for the processing of tonal information in comparison with segmental information.

Other evidence shows that the kind of perceptual decision involved in tone processing, even in its simplest form, requires a certain accumulation of evidence and may be more difficult than perceptual decisions about vowels. Ritsma, Cardozo, Domburg, and Neelen (1965) reported a direct improvement in accuracy of pitch matching as a function of increasing duration of complex tone stimuli. More recently, Robinson and Patterson (1995) asked English listeners to classify vowel segments on one of three dimensions: vowel quality, tone height, or tone chroma. Performance was measured as a function of stimulus duration. Vowel quality could be reliably reported for segments too short for reliable categorization of either of the tone dimensions. At all stimulus durations, moreover, classification of vowel quality was significantly superior to classification of either of the tone dimensions. Robinson and Patterson argue that an interactive relationship between pitch identification and vowel quality cannot consist of the use of pitch information to guide vowel quality identification; if such an interactive relationship exists, it is more likely to be in the reverse direction. Although the brief synthetic stimuli used in these experiments are an imperfect analogue of natural speech, the results do suggest that, in a simple perceptual task, decisions on the segmental dimension of vowel quality can be made far more rapidly than can tonal decisions.

In the simple perceptual task we used in Experiments 2 and 3, the results certainly accord with this account. When the discrimination to be made was between two syllables differing only in tone, the subjects responded more slowly and less accurately. Only perceptually easy tonal distinctions had any significant effect such that responses were facilitated when tone differed as opposed to when it was the same; in contrast, syllable onset and vowel consistently exercised such facilitatory effects. Thus, in this task, segmental information about vowel quality was clearly more salient than was tonal information.

Interestingly, the picture was not quite so clear-cut with the other methodology we used, the auditory lexical decision task of Experiment 1. Recall that, in that task, although again the condition in which the input differed (from a real word) only in tone was clearly harder than the other conditions, there was a greater facilitatory effect on responses of different tone relative to same tone than there was of different vowel relative to same vowel. (Onset, again, had a consistent facilitatory effect.) This relative lack of an effect of vowel is, as we pointed out in discussing the results of Experiment 1, consistent with listener caution in the processing of vowel information, as revealed by other recent studies (Cutler et al., 1996; van Ooijen, 1994, 1996). Our later results, from the simpler task of Experiments 2 and 3, suggest that the disadvantage for vowels is specific to word recognition and indeed supports van Ooijen's (in press) argument that the disadvantage reflects, in computation of the lexical access code in spoken-word recognition, a lower priority for vocalic information than for consonantal information.

Lexical tone, however, as we suggested in the introduction, does participate fully in the lexical access code. Responses in Experiment 1 were more accurate when tone differed (from a real word) that when tone was the same. In contrast to the situation with lexical stress, speakers of a tone language cannot ignore prosodic (suprasegmental) information about lexical identity. In lexical stress languages, the prosodic information per se is usually redundant, since there are also segmental correlates of nearly all stress distinctions; thus, the speaker of a stress language incurs remarkably little cost, and possibly a considerable benefit in simplification of processing, by omitting prosodic considerations entirely from computation of the lexical access code. In lexical tone languages, the cost of ignoring the prosodic dimension in word recognition would be inordinately high.

Nevertheless, processing the prosodic dimension in a language like Cantonese, as the results from all three of our experiments attest, is not a simple matter: tonal information often arrives later than does information about the vowel that bears the tone, and, hence, the processing of tone can be at a disadvantage in comparison with the processing of the very segment (the vowel) upon which it is realized. In speeded-response tasks such as we used in the present study, this temporal delay in the availability of the information shows up in a significantly greater probability that tone will be misprocessed than that the segmental dimensions onset and vowel will be misprocessed. In this respect, the situation in lexical tone languages is indeed similar to the situation in lexical stress languages: Information is processed as soon as it becomes usable, but prosodic information may reach this state relatively slowly. In lexical stress languages, prosodic information may not become usable until more than one syllable of a word is heard; in tone languages, it may become usable only when more of the vowel that carries it is available than is needed for identification of the vowel itself. In either case, limitations on the usability of prosodic information arise simply and necessarily from the acoustic characteristics of speech.

REFERENCES

BOND, Z. S. (1981). Listening to elliptic speech: Pay attention to stressed vowels. *Journal of Phonetics*, **9**, 89-96.

BURNHAM, D., KIRKWOOD, K., LUKSANEEYANAWIN, S., & PANSOTTEE, S. (1992). Perception of Central Thai tones and segments by Thai and Australian adults. In *Pan-Asiatic linguistics: Proceedings of the Third International Symposium of Language and Linguistics* (pp. 546-560). Bangkok: Chulalongkorn University Press.

CONNINE, C. M., CLIFTON, C. E., & CUTLER, A. (1987). Lexical stress effects on phonetic categorization. *Phonetica*, **44**, 133-146.

CUTLER, A. (1986). *Forbear* is a homophone: Lexical prosody does not constrain lexical access. *Language & Speech*, **29**, 201-220.

CUTLER, A., & CLIFTON, C. E. (1984). The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 183-196). Hillsdale, NJ: Erlbaum.

CUTLER, A., MEHLER, J., NORRIS, D. G., & SEGUI, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, **19**, 141-177.

CUTLER, A., & OTAKE, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory & Language*, **33**, 824-844.

CUTLER, A., VAN OOIJEN, B., NORRIS, D., & SANCHEZ-CASAS, R. (1996). Speeded detection of vowels: A cross-linguistic study. *Perception & Psychophysics*, **58**, 807-822.

ELLIS, L., DERBYSHIRE, A. J., & JOSEPH, M. E. (1971). Perception of electronically gated speech. *Language & Speech*, **14**, 229-240.

FEAR, B. D., CUTLER, A., & BUTTERFIELD, S. (1995). The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America*, **97**, 1893-1904.

FOK, C. Y.-Y. (1974). A perceptual study of tones in Cantonese. *Centre of Asian Studies: Occasional Papers and Monographs* (No. 18). Hong Kong: Centre of Asian Studies, University of Hong Kong.

FOX, R. A., & UNKEFER, J. (1985). The effect of lexical status on the perception of tone. *Journal of Chinese Linguistics*, **13**, 69-90.

GANDOUR, J. (1981). Perceptual dimensions of tone: Evidence from Cantonese. *Journal of Chinese Linguistics*, **9**, 20-36.

GANDOUR, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, **11**, 149-175.

GANONG, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance*, **6**, 110-125.

HILLENBRAND, J., GETTY, L. A., CLARK, M. J., & WHEELER, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, **97**, 3099-3111.

INSTITUTE OF LANGUAGE IN EDUCATION, HONG KONG EDUCATION DEPARTMENT (1992). *Common Chinese characters pronounced according to Cantonese*. Hong Kong: Hong Kong Government.

KONG, Q. M. (1987). Influence of tones upon vowel duration in Cantonese. *Language & Speech*, **30**, 387-399.

LEE, L., & NUSBAUM, H. C. (1993). Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese. *Perception & Psychophysics*, **53**, 157-165.

LIN, H.-B., & REPP, B. H. (1989). Cues to the perception of Taiwanese tones. *Language & Speech*, **32**, 25-44.

MARSLEN-WILSON, W. D., & WELSH, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

MCQUEEN, J. M. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 433-443.

MILLER, G. A., & NICELY, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338-352.

MILLER, J. L. (1978). Interactions in processing segmental and suprasegmental features of speech. *Perception & Psychophysics*, **24**, 175-180.

PETERSON, G. E., & BARNEY, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

REPP, B. H., & LIN, H.-B. (1990). Integration of segmental and tonal information in speech perception. *Journal of Phonetics*, **18**, 481-495.

RITSMA, R. J., CARDOZO, B. L., DOMBURG, G., & NEELEN, J. J. M. (1965). The buildup of the pitch percept. *IPO Progress Report*, **1**, 12-15.

ROBINSON, K., & PATTERSON, R. D. (1995). The stimulus duration required to identify vowels, their octave, and their pitch chroma. *Journal of the Acoustical Society of America*, **98**, 1858-1865.

SHEN, X. S., & LIN, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language & Speech*, **34**, 145-156.

SLOWIACZEK, L. M. (1990). Effects of lexical stress in auditory word recognition. *Language & Speech*, **33**, 47-68.

SPEER, S. R., SHIH, C.-L., & SLOWIACZEK, M. L. (1989). Prosodic structure in language understanding: Evidence from tone sandhi in Mandarin. *Language & Speech*, **32**, 337-354.

STRANGE, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America*, **85**, 2135-2153.

SWINNEY, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning & Verbal Behavior*, **18**, 645-659.

TAFT, M., & CHEN, H.-C. (1992). Judging homophony in Chinese: The influence of tones. In H.-C. Chen & O. J. L. Tzeng (Eds.), *Language processing in Chinese* (pp. 151-172). Amsterdam: Elsevier.

TSANG, K. K., & HOOSAIN, R. (1979). Segmental phonemes and tonal phonemes in comprehension of Cantonese. *Psychologia*, **22**, 222-224.

VAN OOIJEN, B. (1994). *The processing of vowels and consonants*. Unpublished doctoral dissertation, University of Leiden.

VAN OOIJEN, B. (1996). Vowel mutability and lexical selection in English: Evidence from a word reconstruction task. *Memory & Cognition*, **24**, 573-583.

WANG, M. D., & BILGER, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America*, **54**, 1248-1266.

WHALEN, D. H., & XU, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, **49**, 25-47.

WINER, B. J. (1972). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

APPENDIX A
Word and Nonword Items Used in Experiment 1

1	2	3	4	5	6
博士	說話	回憶	業務	茶杯	諸侯
/bok8-si6/	/syt8-wa6/	/wui4-jik7/	/jip9-mou6/	/tsa4-bui1/	/dzy1-hau4/
/bok8-si2/	/syt8-wa4/	/wui4-jik9/	/jip9-mou5/	/tsa4-bui3/	/dzy1-hau6/
/bok8-sy6/	/syt8-wo6/	/wui4-juk7/	/jip9-miu6/	/tsa4-bei1/	/dzy1-hai4/
/bok8-sy2/	/syt8-wo4/	/wui4-juk9/	/jip9-miu5/	/tsa4-bei3/	/dzy1-hai6/
/bok8-ji6/	/syt8-ha6/	/wui4-dik7/	/jip9-lou6/	/tsa4-pui1/	/dzy1-lau4/
/bok8-ji2/	/syt8-ha4/	/wui4-dik9/	/jip9-lou5/	/tsa4-pui3/	/dzy1-lau6/
/bok8-jy6/	/syt8-ho6/	/wui4-duk7/	/jip9-liu6/	/tsa4-pei1/	/dzy1-lai4/
/bok8-jy2/	/syt8-ho4/	/wui4-duk9/	/jip9-liu5/	/tsa4-pei3/	/dzy1-lai6/

7	8	9	10	11	12
資金	表演	事實	笑容	電燈	腦筋
/dzi1-gam1/	/biu2-jin2/	/si6-sat9/	/siu3-jung4/	/din6-dang1/	/nou5-gan1/
/dzi1-gam2/	/biu2-jin1/	/si6-sat7/	/siu3-jung1/	/din6-dang6/	/nou5-gan2/
/dzi1-gon1/	/biu2-jyn2/	/si6-sap9/	/siu3-jing4/	/din6-dong1/	/nou5-gun1/
/dzi1-gon2/	/biu2-jyn1/	/si6-sap7/	/siu3-jing1/	/din6-dong6/	/nou5-gun2/
/dzi1-ham1/	/biu2-sin2/	/si6-hat9/	/siu3-tung4/	/din6-hang1/	/nou5-ban1/
/dzi1-ham2/	/biu2-sin1/	/si6-hat7/	/siu3-tung1/	/din6-hang6/	/nou5-ban2/
/dzi1-hon1/	/biu2-syn2/	/si6-hap9/	/siu3-ting4/	/din6-hong1/	/nou5-bun1/
/dzi1-hon2/	/biu2-syn1/	/si6-hap7/	/siu3-ting1/	/din6-hong6/	/nou5-bun2/

Note—All syllable markings are from *Common Chinese Characters Pronounced According to Cantonese* (Institute of Language in Education, Hong Kong Education Department, 1992).

APPENDIX B
Word and Nonword Syllables Used in Experiments 2 and 3

Word Syllables

/ji1/, /ji2/, /jy1/, /jy2/, /si1/, /si2/, /sy1/, /sy2/,
/piu4/, /piu5/, /pei4/, /pei5/, /liu4/, /liu5/, /lei4/, /lei5/

Nonword Syllables

/te1/, /te2/, /ty1/, /ty2/, /ge1/, /ge2/, /gy1/, /gy2/,
/fou4/, /fou5/, /foe4/, /foe5/, /kou4/, /kou5/, /koe4/, /koe5/

Note—All syllable markings are from *Common Chinese Characters Pronounced According to Cantonese* (Institute of Language in Education, Hong Kong Education Department, 1992).

(Manuscript received July 27, 1995;
revision accepted for publication April 8, 1996.)