Speaker normalization in the perception of Mandarin Chinese tones

Corinne B. Moore[a] and Allard Jongman[b]
Cornell Phonetics Laboratory, Department of Modern Languages, Morrill Hall, Cornell University, Ithaca, New York 14853

(Received 12 August 1996; revised 19 March 1997; accepted 16 April 1997)

This study investigated speaker normalization in perception of Mandarin tone 2 (midrising) and tone 3 (low-falling–rising) by examining listeners' use of $F0$ range as a cue to speaker identity. Two speakers were selected such that tone 2 of the low-pitched speaker and tone 3 of the high-pitched speaker occurred at equivalent $F0$ heights. Production and perception experiments determined that turning point (or inflection point of the tone), and $\Delta F0$ (the difference in $F0$ between onset and turning point) distinguished the two tones. Three tone continua varying in either turning point, $\Delta F0$, or both acoustic dimensions, were then appended to a natural precursor phrase from each of the two speakers. Results showed identification shifts such that identical stimuli were identified as low tones for the high precursor condition, but as high tones for the low precursor condition. Stimuli varying in turning point showed no significant shift, suggesting that listeners normalize only when the precursor varies in the same dimension as the stimuli. The magnitude of the shift was greater for stimuli varying only in $\Delta F0$, as compared to stimuli varying in both turning point and $\Delta F0$, indicating that normalization effects are reduced for stimuli more closely matching natural speech.
 [S0001-4966(97)01408-2]

PACS numbers: 43.71.Bp, 43.71.Es [WS]

INTRODUCTION

This study examines the role of speaker-dependent $F0$ information in the perception of lexical tone. It is well known that the perception of segments requires listeners to normalize to reduce overlap among phonetic categories. A classic example of this overlap is illustrated by formant frequency data for vowels in which two phonetic categories from different speakers have similar formant values (Peterson and Barney, 1952). Listeners either actively or passively compensate for this acoustic variability, caused by differences in speaker vocal tract size, in order to identify segments accurately. Consequently, segments that have very similar acoustic characteristics may not be perceived identically.

A. Speaker normalization: The use of speaker-specific acoustic information in vowel perception

Previous work on normalization has shown that listeners use acoustic information outside of the speech sound itself (extrinsic information) about speaker identity in order to classify vowels. For example, Peterson and Barney (1952) found that perception of vowel tokens produced by a wide variety of speakers exhibited confusion within the areas of overlap in the vowel formant data. Ladefoged and Broadbent (1957) provided evidence that listeners refer to extrinsic information specifying the vowel space of a speaker. In this study, six versions of the phrase *please say what this word is* were synthesized, each with a different range of $F1$ and $F2$ to represent different speakers. In addition, four test words of the form, ''*b_t*'' were synthesized with $F1$ and $F2$ values of the vowel approximately corresponding to the vowels in the words ''bit,'' ''bet,'' ''bat,'' and ''but.'' Listeners were asked to identify the final target word. Results showed that perception of a given target word changed as a function of the formant frequency range in the precursor. For example, the target *bit* would be perceived as *bet* when preceded by the precursor in which $F1$ was relatively low. There was thus a contrast effect, whereby a low $F1$ in the precursor would make listeners perceive a given $F1$ in the target as relatively higher, changing the percept from [ɪ] to [ɛ].

While Ladefoged and Broadbent's results suggested *what* acoustic information was used in speaker normalization, it remained unclear *how* this information was used. In particular, acoustic information may either be used to identify the speech sound directly, or it may be used as a cue to speaker identity, establishing a representation against which acoustic characteristics may be calibrated. This question was investigated by Johnson (1990) for vowel continua. In Johnson's study, speaker identity was defined by $F0$, whereas Ladefoged and Broadbent (1957) manipulated formant frequencies to specify speaker identity.

In a series of three experiments, Johnson examined perception of test words in isolation and in carrier phrases whose $F0$ manipulations signaled the same speaker, different speakers, or were ambiguous with respect to speaker identity. In a series of perception pretests, Johnson manipulated $F0$ in a synthesized ''hood-hud'' continuum and also in the synthesized carrier phrase ''this is—.'' These pretests were designed to determine the relationship between $F0$ and speaker identity for both the vowel tokens and the carrier phrases. In the first experiment, Johnson compared percep-

[a]Current address: Corporate Technical Publications, Diebold Incorporated, 5995 Mayfair Road, North Canton, OH 44720.
[b]Electronic mail: aj12@cornell.edu

tion of vowels in isolation and in carrier phrases. The $F0$ levels of the vowel tokens were 100 and 150 Hz, levels which had been shown by the pretests to correspond to different speakers. These vowels were also embedded in carrier phrases which had been attributed to a single speaker. Listeners were asked to label the vowel tokens. The results of the vowels in carrier phrases were then compared to results of those vowels in isolation. Results showed that shifts in vowel identification observed for the 150-Hz versus the 100-Hz tokens were reduced if the carrier phrase signaled that they were produced by the same speaker. The second experiment compared the vowel shifts when both vowel tokens and carrier phrases were ambiguous with respect to whether they had been produced by the same or different speakers. Results showed no significant difference in shifts for vowel tokens in isolation as compared to those in the carrier phrases, presumably since there was no conflicting information about the speaker. Finally, Johnson showed that perception of vowels shifted when the carrier phrases indicated two speakers, but the vowel tokens were at a constant $F0$ level, corresponding to a single speaker. This result supported Ladefoged and Broadbent (1957), in that identical stimuli were perceived differently when perceived as being produced by different speakers. The three experiments in Johnson's study thus provided evidence that listeners use $F0$ as a cue to speaker identity, and that this perceived speaker identity affects vowel perception.

B. Context effects and Mandarin Chinese tones

While the majority of studies have investigated normalization in the perception of vowels, less experimental work has been done to examine speaker normalization in the perception of other types of speech sounds (for examples of normalization for consonants, see Mann and Repp, 1980; Whalen, 1981; Jongman and Miller, 1991; Johnson, 1991). The present study extends work on normalization from the segmental domain to the suprasegmental domain, focusing on lexical tone. The tone language used in this study is Mandarin Chinese, whose four lexical tones include a high-level tone (tone 1), a midrising tone (tone 2), a low-falling–rising tone (tone 3), and a high-falling tone (tone 4). Examples of the four tones for one speaker are shown in Fig. 1. The present series of experiments examines whether $F0$ range, as a cue to speaker identity, influences perception of tones 2 and 3.

As suprasegmentals, tones are perceived relative to other tones, although they are also distinguished by tone-internal (intrinsic) acoustic properties—primarily pitch height and contour (Gandour, 1978; Coster and Kratochvil, 1984). For tones which contrast in both of these dimensions, intrinsic $F0$ information may be sufficient for correct identification. To identify tones differing only in $F0$ height, however, listeners must refer to their knowledge of the speaker's $F0$ range, and where tones occur within that range. For example, a low tone produced by a high-pitched speaker and a high tone produced by a low-pitched speaker may be acoustically very similar. The process by which listeners adjust perception according to speaker-specific acoustic information is referred to as speaker normalization. Few studies have investigated the role of extrinsic $F0$ in tone perception, however, and results from these studies have not provided convincing evidence for speaker normalization.

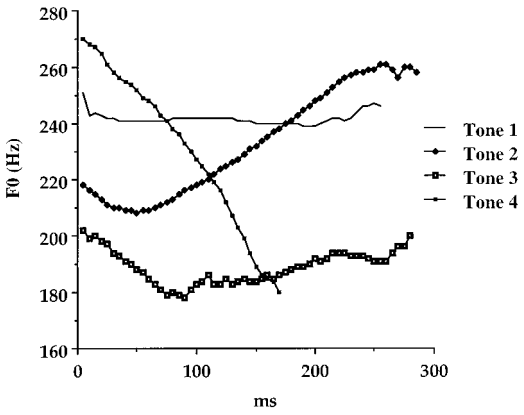

FIG. 1. $F0$ contours for each of the four Mandarin Chinese tones, taken from one token spoken in isolation by one of the speakers in this study (segmental context *ma*).

In a study specifically addressing speaker normalization for tones, Leather (1983) tested perception of Mandarin Chinese tone stimuli in two natural precursor phrases, one representing a low $F0$ range, the other one a higher range. Seven stimuli from two tone 1-tone 2 continua, each continuum representing the $F0$ range of one speaker, were embedded in the precursor phrases. Four steps in the middle of the continuum were identical in $F0$ height and contour, and were included in both continua. Test items (precursor+tone stimulus) were blocked by speaker and presented to five listeners in a labeling task. Individual subject responses were reported for the four pairs of midcontinuum stimuli (paired by speaker condition). Results showed that perception of at least one stimulus pair varied as a function of the speaker.

Unfortunately, Leather did not explicitly predict the type of responses he expected in each condition, nor did the reported data present this information. It is difficult, therefore, to take these data as conclusive evidence for speaker normalization without more detailed information. In particular, it is essential to know the direction of any shift in identification, whether it is consistent across speakers and across stimuli, and whether subject responses conform to predicted results. Leather's use of the tone 1-tone 2 continuum may also have been problematic. These two tones, which vary in both $F0$ height and contour, were synthesized without controlling for confounding acoustic parameters such as onset or offset $F0$. Moreover, both tones occur in the upper region of a speaker's pitch range, making it difficult to compare these tones in terms of $F0$ height.

Other studies have examined the role of extrinsic $F0$ information in tone perception, though they did not specifically address speaker normalization. Using an *AX* anchoring paradigm, in which the *A* element of the stimuli was constant, Lin and Wang (1985) presented subjects with pairs of Mandarin Chinese tones in which the first tone, representing a high-level tone (tone 1), was held at a constant 115 Hz, while the second tone, representing the high-falling tone (tone 4), varied onset $F0$ from 110 to 140 Hz in 10 Hz steps

with an $F0$ fall of 40 Hz. Subjects were asked to label the first tone in each pair. Their results showed that as the onset $F0$ in the second syllable increased, identification of the first tone as a rising tone (tone 2) increased. Thus the higher onset $F0$ of the second syllable cued a wider pitch range, altering the relative $F0$ height of the first tone 1 syllable to be perceived as low. Without a statistical analysis it is uncertain how robust these results are, but they nevertheless provide some evidence that tones are perceived relative to $F0$ range, such that this information contributes directly to the acoustic characteristics of the tone. While the study more broadly indicates that tone perception is affected by extrinsic $F0$, it does not address whether $F0$ information which serves to distinguish speaker identity may influence perception.

Using a similar anchoring paradigm, Fox and Qi (1990) investigated whether context $F0$ influences tone perception, and whether the influence occurs for both native and nonnative listeners. Tone stimuli were presented in isolation and in pairs. In the isolated-token condition, listeners were asked to rate the stimulus according to how closely it resembled the tone 1 or tone 2 exemplar. In the paired-token condition, the first tone was either a tone 1 or tone 2, while the onset $F0$ of the second tone varied along a continuum from tones 1 to 2; subjects were asked to rate the second tone in the pair, according to the same rating scale as in the isolated-token condition. Results showed no significant difference between perception in isolation and in the context condition for either language group.

Following Leather's study, Fox and Qi (1990) presented chi-square values for individual subject responses to four midcontinuum stimuli, showing highly inconsistent patterns of identification across subjects and stimuli. Fox and Qi interpreted these results as weak support for context effects from $F0$ on tone perception, in contrast to those of Lin and Wang (1985), who showed differences in identification as $F0$ range widened.

The reasons for the inconsistencies in Fox and Qi may be related to the methodology used. In Lin and Wang (1985), manipulating the onset of the second tone had the effect of modifying the pitch range, as in Fox and Qi, but listeners were asked to identify the first tone in the sequence, a tone which was constant throughout the experiment. In comparison, Fox and Qi asked listeners to identify the tone containing the modifications, the second tone. The anchor in Fox and Qi did not shift, but rather it was intended that listeners would use the anchor to identify the onset of the second tone as lower, as in a tone 2, or higher, corresponding to a tone 1. A shift in identification for tone 1 anchors may have been expected, since listeners may not have had enough $F0$ range information against which to calibrate the tone stimuli. However, a tone 2 anchor would provide the listener with adequate pitch range information against which to compare the $F0$ onset of the second tone. Since both anchors were included in one test, listeners had the relevant $F0$ range information throughout the test. Therefore it is not surprising that the results yielded no context effects.

Results from these earlier studies have not provided robust evidence that tone perception is affected by contextual acoustic cues, despite the assumption that tones, as suprasegmentals, are perceived according to surrounding information. In Leather 1983 as well as Fox and Qi (1990), shifts in tone identification did not occur reliably for all subjects, nor for a particular stimulus. Also, the direction of the shift, whether contrastive or assimilatory, was either not specified or was inconsistent across subjects and stimuli.

Some of these problems may be remedied by employing a different methodology. For example, in order to test for speaker normalization, precursors must vary in speaker identity. Precursors in Leather (1983) represented different speakers, but Lin and Wang (1985) and Fox and Qi (1990) limited their investigation to context effects, and so precursors consisted of one syllable which did not represent more than one speaker. Moreover, stimuli should reflect a situation in which normalization would be expected to occur, for example, in perception of different tones occurring within an area of overlap in $F0$ range among speakers. Although Leather (1983) examined tone perception for speakers with overlapping $F0$ ranges, both of the tones occurring in that range were high tones (tones 1 and 2), and so may not have been sufficiently distinguished by $F0$ height. In addition, subject data should be analyzed over the entire continuum, rather than for selected stimuli, to determine whether the identification functions for each subject have shifted reliably, and in what direction. The present study addresses these issues in a new investigation of speaker normalization for Mandarin Chinese tones.

In this study, production and perception tests were used to examine tone 2 and tone 3. These tones were chosen, as opposed to tones 1 and 2 used by Leather (1983) and Fox and Qi (1990), because they occupy distinct registers in a speaker's range; although both tones originate at the midpoint of the range, tone 2 rises to cover the high region of the range, while tone 3 is distinctly low, falling to the low region and ending with a rise (in prepausal position) near the middle of the range. This distinction more clearly demarcates $F0$ height as a perceptual cue. Tones 2 and 3 are also similar in contour when spoken in isolation, which may be the reason they cause the most confusion in perception tests (Kiriloff, 1969; Chuang *et al.*, 1972; Gandour, 1978; Li and Thompson, 1978). While overall $F0$ height may contribute to the distinctive phonetic characteristics of tones 2 and 3, two additional acoustic dimensions are relevant: Timing of the turning point, defined as the duration from the onset of the tone to the point of change in $F0$ direction, and also the decrease in $F0$ from the onset of the tone to the turning point, hereafter called $\Delta F0$. These properties are schematized in Fig. 2. Perception studies of Mandarin tones 2 and 3 have found that both timing of the turning point and $\Delta F0$ are perceptually relevant for identification of the tones (Shen and Lin, 1991; Shen *et al.*, 1993).[1]

Using these two acoustic dimensions, the present experiment examined perception of stimuli in a tone 2-3 continuum whose $F0$ levels fall within an area of overlap in $F0$ range for two speakers. In this scenario, speakers overlap in $F0$ range such that the low region of the high-pitched speaker overlaps with the high region of the low-pitched speaker. Within the area of overlap, tones may occur at equivalent $F0$ heights such that they would be low tones for the high-

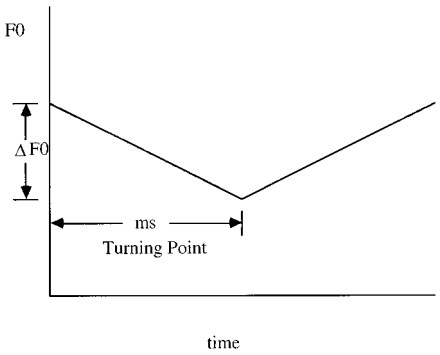

FIG. 2. Turning point and $\Delta F0$ properties schematized for a contour tone.

pitched speaker, but high tones for the low-pitched speaker. Identification of the tone would then be expected to shift (contrastively) depending on the $F0$ range of the precursor. If normalization occurs, stimuli will be identified by using $F0$ range information to ''calibrate'' ambiguous tones. On the other hand, if tone identification does not shift as a function of different $F0$ ranges, speaker normalization will be judged not to have occurred (subjects will not have referred to talker $F0$ range in order to identify tones).

To achieve the scenario conducive to normalization, production data were gathered in order to find two speakers whose $F0$ ranges and tones exhibited areas of overlap. Data from the production study also provided acoustic measurements of tones 2 and 3, which were then used in synthesizing stimuli for the perception experiments.

Stimuli for the perception tests were synthesized to vary in either $\Delta F0$, timing of the turning point, or both acoustic dimensions. These stimuli were presented to listeners in isolation, and then embedded in both high and low precursor phrases. Perception of stimuli in these two conditions was compared to determine whether changes in perceived speaker identity produce changes in tone identification.

I. EXPERIMENT 1: PRODUCTION

This experiment was designed to provide acoustic information about speaker $F0$ ranges and Mandarin tones 2 and 3. The experiment consisted of three reading tasks, the results of which established the mean $F0$ and overall $F0$ range of the speakers.

A. Method

1. Subjects

Four female and three male subjects aged between 19 and 30 years produced the data for this study. The subjects were all native speakers of Mandarin Chinese from Mainland China. They all were graduate or undergraduate students at Cornell University, and competent English speakers as well. None reported any speech disorders. Subjects were paid for their participation.

2. Materials

The data collected were from three reading tasks. The first of these asked subjects to read a long passage from a story, approximately four minutes long, entitled ''*Guo ji da shi he ta de qi zi*'' ''The World Master and His Wife'' by Xiao Fu Xin (Hsu, 1990). For the second task, subjects read minimal sets for each of the four Mandarin tones of the segmental contexts *wu*, *yi*, *bi*, and *ma*. These syllables were randomized and produced in the carrier phrase *Zhe ge zi nian*—(''This word is—''). The third task consisted of subjects reading a randomized list of the minimal sets spoken in isolation. Test items in the carrier phrases and in isolation were produced three times. Both lists also included fillers at the beginning and end of every page to avoid list effects. All reading materials were presented to subjects in Chinese characters.

3. Procedure and analysis

Subjects read the materials in an IAC soundproof booth. They were recorded using an Electrovoice RE20 cardioid microphone and a Carver TD-1700 cassette recorder in the Cornell Phonetics Laboratory. The data were digitized on a Sun Sparcstation 2 computer using a sampling rate of 11 kHz with 16-bit resolution, and were analyzed using Entropics WAVES+/ESPS speech analysis software.

Mean $F0$ and overall $F0$ range were obtained from computer measurements of $F0$ over the long passage. A computer program sampled $F0$ every 5 ms, then filtered out $F0$ values corresponding to a probability of voicing of less than 99%. Mean and modal $F0$ values were then calculated for each speaker.

$F0$ measurements for the minimal sets in isolation and in carrier phrases, as well as the carriers themselves, were taken every 5 ms. Average $F0$ as well as peak and valley $F0$ values for the carriers were calculated for voiced portions. Valley $F0$ values were taken to be the lowest $F0$ value in the tone, peak $F0$ the highest value. $F0$ for tones was measured beginning with the onset of the vowel, or at the first full period of the vowel if the onset of voicing resulted in ''artifact'' $F0$ values which did not appear to be congruous with following $F0$ points. Ending $F0$ values were determined to occur at the offset of voicing (probability of voicing below 99%), or at the offset of the vowel (according to the waveform and spectrogram analysis) if the data showed $F0$ values inconsistent with the path of the tone to that point. Vowel and tone duration was measured from onset to offset of periodicity in the waveform in the *yi* and *wu* syllable types, from the onset of the vowel to its offset as determined by the waveform in the *bi* and *ma* syllables. Spectrograms provided additional help in locating vowel onset and offsets, where vowel onset was marked as the onset of $F1$, and the offset of $F2$ was taken to be vowel offset.

Two acoustic dimensions were measured in isolated tones and in tones embedded in the carriers: timing of the turning point, and the change in $F0$ from the onset of the tone to the turning point ($\Delta F0$). Turning point was defined from the pitch track as the duration between the onset of the tone and the point at which the tone changed $F0$ direction from falling to rising—this point was also the valley for both tones. $\Delta F0$ was calculated to be the difference in $F0$ between the onset of the tone and the turning point. Values were averaged for all instances of the tones, as well as for each syllable type. Tokens repeated in isolation and carrier

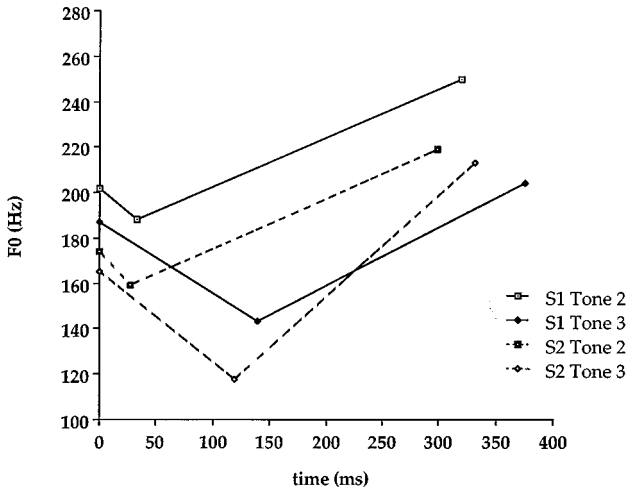

FIG. 3. Averaged $F0$ contours of tones 2 and 3 for the two female speakers (S1, solid lines; S2, dashed lines), across all syllable types.

phrases comprised a corpus of 24 instances of each tone per speaker (4 syllable types×2 reading tasks×3 repetitions). Several instances of the tones contained creak, including four tone 2 and eight tone 3 tokens, which made the relevant measurements impossible, and so these tokens were excluded from analysis.

B. Results

1. F0 range data

Analysis of $F0$ range data was conducted for the long passage, generating roughly 40 000 data points for each speaker. Among seven speakers analyzed, $F0$ range data for two of the four female speakers were found to meet the requirements of the normalization study. Speaker 1 (hereafter S1) had a mean $F0$ of 212 Hz while Speaker 2 (hereafter S2), had a mean $F0$ of 192 Hz. The means reflect that S1 produced more consistently in a higher range than S2. These two female speakers, then, illustrate the $F0$ range characteristics most conducive to testing for speaker normalization. The data show a region of overlap in the $F0$ ranges for S1 and S2. Tones which occur in the overlapping region could conceivably fall in the low region of S1's range, but the high region of S2's range. The tones corresponding to those areas of the speaker ranges are tone 2, the midrising tone, which typically occurs in the upper region of a speaker's range, and tone 3, the low-falling–rising tone, which occupies the low region of a speaker's $F0$ range. If those tones are to be perceived correctly, listeners must adjust tone perception according to which speaker produced the tone.

2. Mandarin tone 2 and tone 3 analysis

Figure 3 shows the $F0$ contours of tones 2 and 3 for both speakers. The $F0$ contours in Fig. 3 represent tone 2 and tone 3 average $F0$ onset, turning point and offset values for all syllable types in the isolation and carrier conditions. Figure 3 shows that the two tones are similar in $F0$ at onset, and have a similar falling–rising contour. Importantly, the $F0$ height of S1's tone 3 falls somewhat below the tone 2 of S2 in a relationship corresponding to the overlap in $F0$ ranges; the tone 2 contour of S2 has an onset of 174 Hz, falling to 159 Hz at the turning point, and ending at 219 Hz, the tone 3 of S1 has an onset of 187 Hz, falling to 143 Hz, and ending at 204 Hz. Tone 2 for S1 is produced outside of the region of tonal overlap; although tone 3 for S1 and S2 overlap, this overlap is irrelevant since it does not involve a change in tonal category. The crucial observation is that S1's tone 3 and S2's tone 2 overlap.

TABLE I. Average duration (ms) of tone 2 for S2 and tone 3 for S1, according to syllable type.

Tone 2 (S2)	ma ''hemp''	yi ''move''	bi ''nose''	wu ''not''
mean	296	304	287	307
Tone 3 (S1)	ma ''horse''	yi ''to lean against''	bi ''pen''	wu ''dance''
mean	338	417	328	418

Recall that the $F0$ range data indicated that the two female speakers shared a region of overlap which encompasses the lower range of S1 and the upper range of S2. The low tone of S1 and the high tone of S2 occurred precisely in this region, a pattern that is predicted by the $F0$ range data. These data attest to the likelihood that tone 2 syllables for S2 and tone 3 syllables for S1 may be confusable in isolation. They were, thus, appropriate to use in subsequent tests for normalization.

Mean duration measurements taken for each tone according to the vowel in each syllable type are shown in Table I. Table I lists average tone durations for each syllable type, including six tokens of each type (three tokens produced in isolation, three in carrier sentences), for a total of 24 tokens possible. Fourteen tokens of S1 (five for *ma*, two for *yi*, five for *bi*, and two for *wu*) and two tokens of S2's *wu* were excluded because the presence of creak made location of vowel offset impossible.[2] S2's tone 2 durations range from 268 to 371 ms; S1's tone 3 range was from 328 to 483 ms. Average duration for the tone 2 (S2) tokens was 299 ms, as compared with 375 ms for tone 3 (S1) tokens. An unpaired, two-tailed t test showed this difference to be significant $[t(30) = 7.98,\ p < 0.001]$. However, a substantial area of overlap was represented in these tone 2 and 3 tokens: from 328 to 371 ms.

Duration was also measured in terms of timing of the turning point for tone 2 (S2) and tone 3 (S1), over all syllable types. These data were analyzed in two ways: in absolute duration (ms), and also as a percentage of tone duration. Tone 2 turning point values averaged 27 ms, occurring at an average of 9% into the tone. Tone 3 showed an average turning point of 139 ms, occurring at 37% into the tone on average. Turning point values for tone 2 ranged from 0 to 91 ms, as compared to 105 to 200 ms for tone 3, demonstrating a significant difference between the two tones $[t(30) = 11.32,\ p < 0.0001]$. Calculated as a percentage of total tone duration, tone 2 ranged from 0% to 30% of the tone, while the tone 3 range was from 28% to 53%. There was, thus, a small area of overlap between the tones, although the difference between the two tones was significant $[t(30) = 13.16,\ p<0.0001]$.

In addition to turning point, the other acoustic parameter observed was the decline in $F0$ from the onset of the tone to

the turning point, or $\Delta F0$. These data showed an average $\Delta F0$ of 15 Hz for tone 2 (S2), and 44 Hz for tone 3 (S1). For tone 2, $\Delta F0$ ranged from 0 to 56 Hz, and from 24 to 106 Hz for tone 3. There was, thus, an area of overlap occurring between 24 and 56 Hz. Unpaired two-tailed t-test results showed that $\Delta F0$ differences between tone 2 (S2) and tone 3 (S1) were significant [$t(30)=4.21$, $p<0.001$].

3. Discussion

The data in experiment 1 show that the two female speakers shared a region of overlap in the $F0$ range, and that tone 2 for the low-pitched speaker and tone 3 for the high-pitched speaker also overlap, two conditions essential to test for normalization effects. The hypothesis of this study was that listeners must normalize for speaker identity ($F0$ range) in order to identify these tones correctly. To test this hypothesis, perception of tone stimuli presented with high and low $F0$ precursors was examined, where tone stimuli formed a continuum from tone 2 to tone 3, varying in $\Delta F0$ and turning point characteristics.

II. EXPERIMENT 2: PERCEPTION OF TURNING POINT AND $\Delta F0$ IN ISOLATION

Although earlier studies using tones 2 and 3 for a single speaker have considered timing of the turning point and $\Delta F0$ to be perceptual cues for these two tones (e.g., Shen and Lin, 1991; Blicher *et al.*, 1990) these studies have not documented a systematic investigation of these parameters which addresses whether $\Delta F0$ and turning point covary, whether perception based on each of these parameters was equally categorical, or what combinations of turning point and $\Delta F0$ trigger shifts in identification from one tone to the other. A perception experiment was therefore devised to determine the relative importance of timing of the turning point and $\Delta F0$. The experiment tests perception of isolated synthetic stimuli in which timing of the turning point and $\Delta F0$ have been systematically manipulated. Subject responses should clarify how these acoustic dimensions are perceived, their relative importance, and any ambiguity created by the combination of parameters. In addition, results of this experiment were used to model more accurately the synthetic stimuli used in subsequent tests.

A. Method

1. Subjects

Six subjects from Mainland China, three males and three females between the ages of 19 and 40 years old, participated in experiment 2. All were recruited from the Cornell University community. None reported any hearing disorders. Subjects were paid for their participation.

2. Stimuli

The syllable [u] was chosen for synthesis because this syllable type was also used in the Shen *et al.* (1993) and Shen and Lin (1991) studies. Stimuli were created using the Delta speech synthesis program developed by Hertz (Charif *et al.*, 1992; Zsiga, 1994) and a Klatt synthesizer (1980) in the Cornell Phonetics Laboratory.[3] Formant frequency values for $F1-F5$ were averaged for the two speakers to create an ambiguous voice quality. Formant values for $F1$ and $F2$ included separate measurements for the onset and offset of the formants. Production data for $F3$ showed no substantial difference between onset and offset values for each speaker and thus $F3$ was held constant over the entire vowel. $F4$ and $F5$ were also held constant, based upon measurements from the steady-state portion of the vowel. The resulting composition of formant values is shown in Table II. Duration of the stimuli was constant at 400 ms. Amplitude of voicing began at 55 dB (Klatt parameter) and declined to 53 dB over the duration of the token.

TABLE II. Formant frequency values (Hz) for synthesized stimuli.

	Onset	Offset
$F1$	345	304
$F2$	703	628
$F3$	2940	2940
$F4$	4320	4320
$F5$	4840	4840

Based on the turning point data in experiment 1, stimuli for the perception tests were designed to vary timing of the turning point along a continuum from 20 to 240 ms in 20-ms steps, for a total of 12 stimuli. These stimuli should trigger tone 2 responses when the turning point occurs close to the tone onset, and tone 3 responses when the turning point occurs late in the tone. In addition, $\Delta F0$ was varied from 10 to 70 Hz in steps of 5 Hz, generating 13 stimuli. Because tone 2 typically exhibits a shallower $\Delta F0$, it was expected that tones with a $\Delta F0$ equal to 10 Hz would produce more tone 2 responses than tones with a $\Delta F0$ of 70 Hz.

These two manipulations together allowed for testing of both timing of the turning point and $\Delta F0$, in an effort to understand how these acoustic parameters are used in the perception of tones 2 and 3. Figure 4 represents all combinations of these parameters which were included in experiment 2.

Based on the duration and $F0$ manipulations represented in Fig. 4, predictions can be made about which regions of the graph might be expected to trigger tone 2 and tone 3 responses. According to traditional phonetic descriptions, tone 2 is characterized by a short fall in $F0$ followed by a long rise, while tone 3 has a deeper, longer fall followed by a long rise. The shaded region in fig. 4 which corresponds to the tone 2 characterization contains stimuli with turning points from 20 ms to approximately 140 ms, along with lower $\Delta F0$ values. It might also be expected that a high $\Delta F0$ coupled with an early turning point (20 to 40 ms) would yield a tone 2 percept, since the $F0$ rise of the stimulus is predominant. On the other hand, listeners would be expected to label as tone 3 any stimulus containing a deeper $F0$ fall and a longer duration to turning point, stimuli marked in the lined region of Fig. 4.

3. Procedure

Experiment 2 used a forced-choice labeling paradigm in which subjects heard each [u] stimulus in isolation and were asked to choose from two lexical items. There were 468

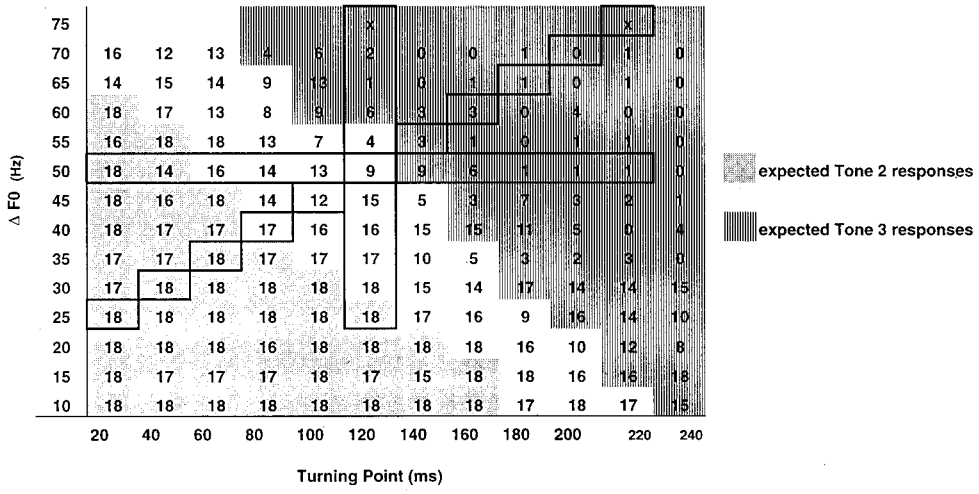

FIG. 4. Combinations of turning point and $\Delta F0$ manipulations for synthesized stimuli. Turning point manipulations are represented along the horizontal axis, $\Delta F0$ manipulations on the vertical axis. The shaded region corresponds to predicted tone 2 responses, and the lined region corresponds to predicted tone 3 responses. Numbers correspond to actual tone 2 responses in experiment 2 for isolated stimuli varying in timing of the turning point and $\Delta F0$. Eighteen responses were possible for each stimulus. These results determined which tone continua to use in experiments 3–5. These continua are enclosed in boxes: The diagonal boxes indicate stimuli varying along both turning point and $\Delta F0$, the horizontal boxes represent stimuli varying only in turning point, and the vertical row of boxes shows stimuli varying only in $\Delta F0$. An ''x'' denotes stimuli added after experiment 2.

tokens in total (12 turning point$\times$13 $\Delta F0$ $\times$ 3 repetitions). Stimuli were low-pass filtered at 5.2 kHz and played out in randomized order on a 12-bit audio system using the BLISS software program (Mertus, 1989) on a Swan 386 PC. Due to the number of stimuli (156 different manipulations), stimuli were presented in three blocks, one set of stimuli per block, with an intertrial interval (ITI) of 2 s.

One to four subjects at a time participated in the test. They were instructed to respond to each item as quickly as possible by pressing the button corresponding to the Chinese character for ''not'' (tone 2) or ''dance'' (tone 3). To avoid misidentification, subjects were asked to pronounce the button labels prior to the experiment. A practice session consisting of 23 test items preceded the test. These practice items provided listeners with end points for each parameter manipulated, as well as stimuli in between. Instructions were given in English, since most of the subjects were undergraduate or graduate students at Cornell and highly proficient English speakers. However, for this and all subsequent tests, the few subjects who were not proficient in English were given instructions in Mandarin in addition to English. There were no differences in responses between the subjects instructed in Mandarin and those instructed in English. Subject responses were collected by computer using the BLISS software system. Responses for each stimulus were added across speakers.

B. Results

Figure 4 also gives the number of tone 2 responses for all stimuli, arranged according to values for timing of the turning point and $\Delta F0$. Figure 4 shows that, as expected, stimuli were clearly identified as tone 2 in the region where turning point and $\Delta F0$ values were low. Along the turning point dimension, unambiguous tone 2 responses spanned virtually the entire continuum, up to a $\Delta F0$ of 30 Hz. Thus tone 2 appears to tolerate substantial delays (up to 240 ms) when the initial $F0$ fall is 30 Hz or less. Although the total duration of the stimuli (400 ms) could have biased the listener toward tone 3 responses, this was clearly not the case.

Along the $\Delta F0$ dimension, decreases of up to 70 Hz were still identified as tone 2. According to Kratochvil (1971), who tested perception of synthetic tones with durations from 90 to 240 ms, a duration of between 50 and 100 ms is required for perception of isolated Mandarin tones. Weber ratios for duration have been reported for 100-ms signals as 0.026 by Ruhm *et al.* (1966), and for 400-ms signals as 0.12 by Stott (1935), corresponding to approximately 10–60 ms for the 400-ms stimuli in experiment 2. If the initial 20–40-ms portion of the tone is imperceptible (below threshold), subjects may hear only a rise, not the initial fall, since $\Delta F0$ is equivalent to 0 at the earliest turning points. $\Delta F0$ began to trigger tone 3 responses when the turning point reached approximately 80 ms into the tone. At this point, tone 3 responses increased as a function of both $\Delta F0$ and timing of the turning point. When the turning point occurred as late as 200 ms into the tone, however, relatively low $\Delta F0$ values (approximately 35 Hz or greater) elicited tone 3 responses. Thus the later the turning point, the easier it is for $\Delta F0$ to effect a change, though even late turning points are resilient to the effects of a $\Delta F0 < 35$ Hz. On the other hand, $\Delta F0$ appeared to trigger more tone 3 responses in stimuli with later turning points rather than early ones.

C. Discussion

The purpose of experiment 2 was to determine how the acoustic dimensions of timing of the turning point and $\Delta F0$ contribute to perception of Mandarin tones 2 and 3. As for which of the two acoustic dimensions might be more important, the data suggested that there is an interdependency between $\Delta F0$ and timing of the turning point. It ap-

pears that $\Delta F0$ was more relevant as turning point increased, while turning point was more relevant for a $\Delta F0$ of more than 30 Hz. Tone 2 perception seems to tolerate more variability overall, while tone 3 requires a late turning point *and* a large fall in $F0$. This may be because the later turning point enhances the perceptual salience of the initial fall.

While turning point and $\Delta F0$ appear to be interdependent, there were places where either dimension alone was sufficient to produce categorical functions. For example, stimuli with a constant turning point of 140 ms showed all tone 2 responses for a $\Delta F0$ of 20 Hz, moving to 50% responses for a $\Delta F0$ of 50 Hz, and all tone 3 responses for the largest $\Delta F0$. Along the turning point dimension, tone 2 responses moved categorically from 100% to 0% for the continuum of stimuli with a constant 50-Hz $\Delta F0$. While either of the two acoustic parameters were robust enough to trigger categorical identification functions, it is clear from both the production and perception data that timing of the turning point and $\Delta F0$ operate in tandem as perceptual cues to tones 2 and 3.

In summary, the results of this perception test showed how tone stimuli which vary in $\Delta F0$ and timing of the turning point are perceived. Subjects made categorical responses based on these two acoustic dimensions, such that identification functions were obtained for either $F0$, turning point, or both parameters. Experiments described in the following sections test normalization effects using tone 2 to tone 3 continua based on the acoustic parameters of $\Delta F0$ and turning point examined above.

III. EXPERIMENTS 3–5: PERCEPTION OF STIMULI IN PRECURSOR PHRASES

The following three experiments tested the hypothesis that listeners perceive tones in part by normalizing for speaker $F0$ range. Experiments 3–5 employed a design which compared how identical stimuli were identified when presented in two contexts differing in $F0$ range. To ensure that the effect was caused by normalization of different talker characteristics, the test used naturally spoken carrier phrases from different speakers as precursors. Normalization effects in this experiment would cause a shift in identification of stimuli as a function of which precursor phrase was heard, high or low $F0$.

As previously noted, earlier experiments on tones have provided evidence that extrinsic acoustic information may influence perception, although only Leather (1983) examined normalization effects due to perceived speaker identity. Experiments 3–5 of the present study expanded on Leather's work by testing two different Mandarin tones, tones 2 and 3, and sought to provide more robust evidence of normalization. This was done in several ways: by examining the direction of any shift in identification relative to the precursor, and by measuring shifts in identification over the entire function, rather than arbitrarily selected points in the middle.

The three experiments were conducted based on the perception data from experiment 2. Each experiment is distinguished according to the stimuli used: Experiment 3 employed a continuum of stimuli containing cues about both timing of the turning point and $\Delta F0$; stimuli for experiment 4 included a continuum of stimuli varying only $\Delta F0$, and experiment 5 used a continuum varying only timing of the turning point. The continua marked by boxes in Fig. 4 were used in these experiments. All three continua share a common midpoint which has a turning point of 120 ms, and a $\Delta F0$ of 50 Hz. This midpoint stimulus received 50% tone 2 responses for the corresponding identification functions resulting from experiment 2.

TABLE III. $F0$ and duration information for precursors used in experiments 3–5.

Speaker	Duration	Average $F0$	Peak	Valley
high $F0$	718 ms	226 Hz	272 Hz	192 Hz
low $F0$	722 ms	187 Hz	229 Hz	170 Hz

A. Experiment 3: Perception of stimuli varying in $\Delta F0$ and turning point

1. Method

a. Subjects. Eleven subjects, seven male and four female, aged between 19 and 40 years, participated in this experiment. All were native speakers of Mandarin Chinese, eight from Mainland China and three from Taiwan, with no known hearing disorders. Because there are many dialects spoken in Mainland China and Taiwan, the subject population in this study was restricted to those speaking only one of the Mandarin dialects according to Norman (1988, p. 191). This restriction provided a more homogeneous subject group, although not as strict as if they had been limited to Beijing Mandarin only. Examples of Chinese languages not represented by subjects included in the study were Shanghai, Cantonese, and Taiwanese.

b. Stimuli. Stimuli for this experiment were synthesized [u] syllables which formed a continuum from tone 2 to tone 3, varying in both timing of the turning point and $\Delta F0$. This continuum is the diagonal set of stimuli shown in Fig. 4. One additional step was created on the tone 3 end of the continuum to provide an equal number of stimuli on either end of the crossover stimulus. Timing of the turning point varied from 20 to 220 ms, in steps of 20 ms. $\Delta F0$ ranged from 25 to 75 Hz, in steps of 5 Hz.

Two natural precursor phrases spoken at a normal speaking rate were chosen from the production data discussed in experiment 1, one from each the high-pitched speaker (S1) and the low-pitched speaker (S2). Table III presents duration and $F0$ information for each precursor, high and low. The peak and valley $F0$ points represent boundaries of the $F0$ range for each speaker, showing a shared region of 192 to 229 Hz. The two phrases differed by 39 Hz in average $F0$, but were further distinguished by the range; the high precursor spanned 192–272 Hz, as compared to 170–229 Hz for the low precursor. In order to visualize how the stimuli were situated with respect to these $F0$ ranges, recall that the synthesized stimuli had a fixed onset and offset of 188 and 212 Hz, respectively, levels which were based upon production data. The $\Delta F0$ value decreased from 163 to 113 Hz in the continuum containing both $\Delta F0$ and turning point cues, and also in the $\Delta F0$ continuum.

Because they were naturally produced, the phrases differed in voice quality as well as $F0$ range. Although formant frequencies of the target stimuli were synthesized to be ambiguous relative to the precursors to reduce any speaker bias (see Table II), it could still be the case that one of the two precursors provided a better match with the target than the other.[4] Each phrase contained the segmental context *Zheige zi nian*— (''This word is—''), each had preceded a high-tone syllable (tone 1 or tone 4) in the production task,[5] and each matched in duration. Synthesized stimuli from the pretest were appended to the precursors, leaving a 50-ms silence between the precursor and the test word.

As additional controls, the carrier phrase contained no instances of tones 2 or 3 or [u]. There were two advantages of limiting the phrases this way. One advantage was to eliminate the environment for tone sandhi effects, which particularly affect tones 2 and 3 (Chao, 1968). The other advantage is that subjects only heard one instance of the test tones, rather than possibly comparing precursor examples of the test tones with the stimuli.

2. Procedure and analysis

The experiment was conducted in the Cornell Phonetics Laboratory. Test items were presented to subjects by way of the PC-based software program BLISS, which randomized and played the stimuli via a D/A converter (12-bit resolution, 11-kHz sampling rate, low-pass filtered at 5.2 kHz). Eleven stimuli were preceded by each of the high and low precursor phrases, creating a total of 22 sentences. Subjects first heard 12 test items in a practice session. The test consisted of a total of 220 trials (22 sentences$\times$10 repetitions), with an intertrial interval of 2250 ms. One to four subjects at a time listened to test items over headphones in separate booths. Subjects were instructed to respond by pressing one of two buttons, which were labeled using the Chinese characters for either ''not,'' or ''dance,'' corresponding to the tone 2 or tone 3 lexical item, respectively.

Responses were recorded and tabulated by computer. For each subject, the crossover points for stimuli in both the high and low precursor conditions were determined by using a probit statistical analysis (Finney, 1971). This statistical method takes into account the subject's responses over the entire stimulus continuum.

3. Results

Results of experiment 3 are summarized in figure and table form below. Figure 5 (top panel) shows the percentage of tone 2 responses for stimuli in the two presentation types (high and low precursor conditions), averaged across subjects. Table IV lists probit values for each subject as a function of the precursor conditions.

The probit values in Table IV represent category boundaries for the listeners who participated in experiment 3. As shown in Fig. 5, subjects perceived more tone 3 (low tone) responses when stimuli were preceded by a high precursor than when stimuli were preceded by a low precursor. The category boundary for the high precursor was earlier for eight of the eleven subjects, and averaged 5.60, as compared

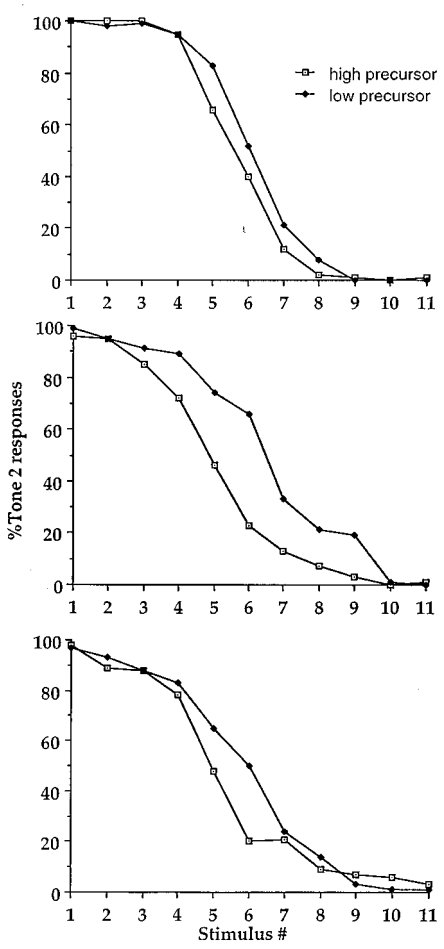

FIG. 5. Top panel: Experiment 3 turning point/$\Delta F0$ continuum identification functions for high and low precursor conditions, averaged across subjects. Stimulus 1 corresponds to predicted tone 2 responses. Middle panel: Experiment 4 $\Delta F0$ identification functions for high and low precursor conditions. Bottom panel: Experiment 5 turning point continuum identification functions for high and low precursor conditions.

to 5.99 for the low precursor. A paired two-tailed t test shows this difference between boundaries to be significant [$t(10) = -2.57$; $p < 0.03$]. Subjects thus appear to refer to the $F0$ range of the precursor in perception of the tones. Moreover, the normalization effect was robust enough to be obtained in a mixed block condition, in comparison to

TABLE IV. Experiment 3 probit values for turning point/$\Delta F0$ stimuli in high and low precursor conditions.

Subject	High precursor	Low precursor
1	5.58	5.21
2	4.64	4.51
3	5.71	6.2
4	6.34	7.11
5	5.88	5.81
6	4.36	5.65
7	5.75	6.54
8	5.3	5.44
9	6.18	6.99
10	6.39	6.71
11	5.44	5.63
mean	5.60	5.99

TABLE V. Experiment 4 probit values by subject for $\Delta F0$ stimuli in high and low precursor conditions.

Subject	High precursor	Low precursor
1	4.47	6.18
2	4.74	6.08
3	6.02	7.76
4	4.29	4.09
5	3.33	4.75
6	4.49	4.56
7	3.32	6.28
8	4.79	6.57
9	5.06	7.24
10	4.10	5.59
mean	4.46	5.91

Leather (1983) in which stimuli were blocked by speaker.

The shift away from tone 2 responses in the high precursor condition demonstrates a contrast effect; the high $F0$ context caused a shift toward low tone (tone 3) responses. While this result is to be expected given the assumption that $F0$ height of the tone is interpreted relative to a speaker's $F0$ range, it differs from earlier findings by Fox and Qi (1990), who instead found primarily assimilatory shifts for paired-token identification tasks.

B. Experiment 4: Perception of stimuli varying in $\Delta F0$

1. Method

a. Subjects. Twenty-two native speakers of Mandarin Chinese participated in this experiment. Twelve of these were subsequently excluded from the results on the basis of criteria outlined below. The ten remaining subjects included five males and five females. One of the subjects was from Taiwan, and nine were from Mainland China. None reported any hearing disorders.

b. Stimuli. Test items were sentences composed of the two precursors used in experiment 3, followed by a test word taken from the $\Delta F0$ continuum in the pretest. This continuum varied only $\Delta F0$ in 11 steps of 5 Hz from 163 Hz. The timing of the turning point was fixed at 120 ms.

c. Procedure and analysis. The test procedure and analysis of data were identical to those used in experiment 3. However, this experiment seemed to be more difficult for subjects than experiment 3, judging both by number of missed trials and failure to achieve categorical identifications at continuum end points. Because of these two problems, it was decided that a subject's responses would be included in the results only if they met the following criteria: (1) they responded to more than 90% of the total trials for each continuum, and (2) they achieved at least an 80% correct response rate on continuum end points. Failure to meet these criteria led to the disqualification of 12 subjects.

2. Results

The averaged identification functions for the $\Delta F0$ continuum in the high and low precursor conditions are presented in Fig. 5 (middle panel) and the probit values are listed in Table V.

The data in Table V show that for the tone 2 to tone 3 continuum varying only in $\Delta F0$, there was an earlier shift to tone 3 responses in the high precursor condition for nine of the ten subjects; average crossover points were 4.46 as compared to 5.91 in the low precursor condition. This difference was significant [$t(9) = -4.69$, $p < 0.001$]. The shift was one of contrast—the high precursor prompted more low-tone responses and vice versa. This result supports the hypothesis that subjects refer to extrinsic $F0$ as a frame of reference for tone perception.

The magnitude of the shift in this experiment was much greater than that in experiment 3. These differences, computed as the difference between the low and high precursor crossover points, were shown to be significant in a two-tailed t test for independent means [$t(19) = -3.33$, $p < 0.003$]. These results indicate that listeners relied more on speaker $F0$ range to disambiguate the tones when the stimuli provided less intrinsic acoustic information about tone category. The implications of this finding are discussed in Sec. IV.

C. Experiment 5: Perception of stimuli varying in turning point

1. Method

Twenty native speakers of Mandarin Chinese participated in experiment 5. Eight subjects were disqualified according to the criteria outlined in Sec. III B. 1 c. The twelve remaining subjects included six males and six females. Four subjects were from Taiwan, and eight were from Mainland China. None reported any hearing disorders.

Again, stimuli used in this experiment were part of the set used in the isolation pretest, appended to the end of the natural precursors used in experiments 3 and 4. The continuum from tone 2 to tone 3 varied only timing of the turning point, from 20 to 220 ms into the tone. The decrease in $\Delta F0$ was constant at 50 Hz.

Test procedures and analysis of results were identical to those of experiments 3 and 4.

2. Results

Figure 5 (bottom panel) displays average percent tone 2 responses for the turning point continuum in the high and low precursor conditions. Table VI lists probit values for each subject in both the high and low precursor conditions. The average boundary in the high precursor condition as compared to the low was 4.78 versus 5.20. Only eight of the 12 subjects showed a shift in the direction predicted by the normalization hypothesis, and the difference between the probits in the two conditions was not statistically significant [$t(11) = -1.55$, $p > 0.15$]. These results suggest that stimuli varying only in a temporal cue did not induce a normalization effect for contexts that vary in an $F0$ dimension.

IV. GENERAL SUMMARY AND DISCUSSION

This study tested the hypothesis that listeners use acoustic information about the speaker in the perception of lexical tones. In particular, the study investigated whether listeners used extrinsic information about speaker $F0$ range in perception of intrinsic acoustic properties of Mandarin tones 2 and

TABLE VI. Experiment 5 probit values for turning point stimuli in high and low precursor conditions.

Subject	High precursor	Low precursor
1	5.83	5.15
2	5.54	6.12
3	5.25	4.73
4	4.55	4.86
5	4.21	5.39
6	4.0	5.08
7	4.56	4.44
8	3.24	5.63
9	4.13	5.12
10	3.61	4.01
11	6.08	6.34
12	6.4	5.49
mean	4.78	5.20

3. The hypothesis predicts that tone identification is affected by changes in perceived speaker identity. If speaker information is not relevant in tone perception, on the other hand, changes in perceived speaker identity should cause no significant shift in identification of tone categories.

To examine the hypothesis, a series of production and perception experiments were conducted. First, production analyses from experiment 1 located two speakers who overlapped in $F0$ range. The analysis revealed that within the area of $F0$ range overlap, a low tone for a high-pitched speaker and a high tone for a low-pitched speaker occurred at equivalent $F0$ heights. Experiment 2 demonstrated that while both $\Delta F0$ and turning point are used in production, either cue alone was sufficient to distinguish the two tones in perception.

The study then investigated whether changes in perceived speaker identity affected tone perception by presenting tone continua in precursor phrases from two different speakers. Results of experiments 3 and 4, which examined perception of both $F0$ and temporal properties of tones 2 and 3 in high-$F0$ and low-$F0$ precursor phrases, showed a small but significant shift in tone identification, in the direction expected if tone stimuli were perceived according to the $F0$ range of the precursor; that is, identical stimuli were perceived as high tones in the low $F0$ precursor phrase, but as low tones in the high $F0$ precursor phrase. These findings thus support the hypothesis that tone identification is influenced by changes in $F0$ range, demonstrating that this information is used as a frame of reference according to which ambiguous tones may be interpreted.

No significant shift was observed for the tone continuum in experiment 5, however, which varied only the temporal dimension of turning point. The stimuli in experiment 5 differed in only one aspect from the stimuli in experiments 3 and 4: they did not vary in $\Delta F0$. These results suggest that normalization is triggered only when both stimuli and precursors vary along the same acoustic dimension. Findings from Moore (1995) for tone stimuli identical to those used in the present study indicate that when precursors vary in a temporal dimension (speaking rate), listeners normalize for rate by shifting category boundaries for the temporal cue (turning point).

If the temporal dimension was not relevant for normalization for $F0$ range in the turning point stimuli, it is tempting to assume that temporal information may not have contributed to the normalization effect in experiment 3, where stimuli varied along both dimensions. However, the larger magnitude of the effect in experiment 4 as compared to experiment 3 contradicts this assumption. This difference in the magnitude of the effect for stimuli varying only in the $F0$ dimension as compared to stimuli varying in both the $F0$ *and* temporal dimensions supports the hypothesis that listeners utilize contextual information to a greater degree when intrinsic acoustic information for tone contrasts is degraded. Such differences between effects have been observed in rate normalization work on vowel perception by Gottfried *et al.* (1990), as well as rate effects in the perception of [b]-[w] continua in Shinn *et al.* (1985). Both of these studies show reductions in normalization effects as stimuli resemble natural speech more closely. Thus it is possible that when listeners are given accompanying temporal information for tone contrasts as in experiment 3, they do not refer to $F0$ range as much as when intrinsic tonal cues are restricted, as in experiment 4. Further work is needed to understand the relative contribution of temporal and $F0$ cues in contexts that also vary in both of these dimensions.

Although this investigation contributes additional data and addresses several inadequacies of Leather's (1983) study, findings of this study support the conclusions of Leather (1983). First of all, the present study observes $F0$ range normalization effects for tones which differ in $F0$ height; tone 2 is an upper register tone, compared to the lower register tone 3. Leather used two upper register tones whose contours are more dissimilar than tones 2 and 3. Second, this study shows normalization effects robust enough to be obtained in a mixed block condition; Leather's subjects were trained on one speaker's voice before hearing stimuli embedded in precursors for that particular speaker. Third, analysis methods for the present study compared crossover boundaries based on the entire identification function, so that reliable shifts were observable based on responses to all stimuli in each condition. The analysis of responses to only selected stimulus pairs rather than analysis of crossover boundaries may have led to the appearance of inconsistent results reported in Leather (1983) as well as Fox and Qi (1990). Fourth, while Leather did not report whether changes in perception were assimilatory or contrastive, or whether changes were consistent for all speakers, the present study provides conclusive evidence that shifts in identification were contrastive—in a direction opposite to the precursor $F0$—and that this shift was consistent across subjects in experiments 3 and 4.

The contrastive context effects shown here differ from those reported in Fox and Qi (1990). Their findings, for paired-token identification tasks, instead showed assimilatory shifts in all but one case. Their study focused on context effects from one preceding tone, rather than on speaker normalization, however. Fox and Qi further argued that assimilatory shifts are evidence for auditory, rather than phonetic, processing of the acoustic signal, based on experimental work by Fujisaki and Kawashima (1971), Pisoni (1975),

Shigeno and Fujisaki (1979), and Shigeno (1986). In these models of perception, assimilatory shifts occur for stimuli which do not undergo category-level perceptual identification, such as for continua whose end points do not represent different phonemes, or for nonspeech stimuli. For continua whose end points represent phonemic distinctions, or for complex tone continua, a categorical memory process is employed, generating contrastive shifts in identification. Shigeno (1991), however, provides evidence that both assimilatory and contrastive effects may occur within the process of phonetic judgment. From the standpoint of these two-stage perceptual processing models, results of the current study would suggest that higher-level phonetic processes are involved in speaker normalization for tones. Notwithstanding the different methods employed in Fox and Qi (1990) as compared to the present study, the opposite shifts in identification raise the question of whether contextual $F0$ information is processed differently depending upon whether it was used as a cue to tone identity, as in Fox and Qi, versus as a cue to speaker identity, as in the present study.

It could be argued that the present results arose from an auditory level of processing or from a simple response bias.[6] However, a comparison of the present results with those obtained from English listeners under the same experimental conditions argues against such explanations (Moore and Jongman, forthcoming). While the Mandarin listeners showed the greatest effect of speaker $F0$ range for stimuli varying only in $F0$, English listeners showed effects only when the foreign distinction was perceptually salient, namely when it was cued by both fundamental frequency and temporal information simultaneously. The fact that Mandarin and English listeners responded differently to our experimental conditions argues against a simple response bias. In addition, these findings suggest that normalization occurs for phonemic contrasts for native listeners, but that it is a function of auditory discriminability for non-native listeners.

Results of this study are consistent with those of Johnson (1990), who found that both intrinsic and extrinsic $F0$ contributes to vowel perception. As the results of experiments 3 and 4 from the present study demonstrate, extrinsic $F0$ significantly influences tone perception by serving as a cue to speaker identity, causing intrinsic $F0$ cues ($\Delta F0$) to be perceived relative to the extrinsic cues ($F0$ range). In other words, extrinsic $F0$ enabled listeners to construct a representation of $F0$ range, against which intrinsic acoustic characteristics of the tones were calibrated.

V. CONCLUSION

The results of this series of experiments suggest that perception of tones is a talker-contingent process. Evidence was provided to show that listeners use extrinsic $F0$ information in perception of lexical tones. Since no explicit tests were conducted to verify that the precursors were perceived as having been produced by different speakers, the present results cannot definitively demonstrate that listeners established a representation of speaker identity. However, the fact that the precursors in this study were natural, intact sentences produced by two different speakers, one with a high $F0$ and one with a low $F0$, suggests that intrinsic acoustic information is mediated through a representation of speaker identity, rather than contributing to tone identification independent of speaker information. These results suggest that the same normalization processes participate in perception of suprasegmentals as well as segments.

Speaker normalization has been assumed to occur as a response to acoustic variability which derives from vocal tract differences among speakers. This variability is exhibited when different speech sounds are acoustically very similar, as illustrated in this study, or when the same speech sound exhibits different acoustic characteristics. Additional research on normalization in perception in the latter instances of variability would further clarify the relationship between acoustic variability and normalization.

Other research on the effects of speaker variability on perception indicates that speech perception is more difficult, and not as accurate, in multiple-talker conditions as compared to single-talker conditions (Mullenix *et al.*, 1989; Sommers *et al.*, 1992), blocked conditions (Strange *et al.*, 1976; Assmann *et al.*, 1982) or when listeners have increased familiarity with the talkers' voices (Verbrugge *et al.*, 1976; Nygaard *et al.*, 1994). These studies suggest that there is a ''cost'' associated with the process of normalizing for speaker differences. While the costs of normalizing for contextual information may be expected for segments, which are perceived highly accurately given only intrinsic cues (Verbrugge *et al.*, 1976), it is not as straightforward in the case of suprasegmentals, where context is assumed to be more intimately connected with identification. In the case of Mandarin Chinese, contour differences between the tones also yield high identification rates in isolation (Howie, 1976). The more relevant case for establishing differing degrees of interdependence on context may be to examine normalization in perception of tones which contrast only in $F0$ height, such as the level tones in Cantonese (Fok, 1974). To the extent that tone perception uses identical perceptual processes as segments, the observation in the present study that normalization effects obtained in a mixed block condition suggests that speaker normalization is a robust process, even for Mandarin tones.

This study has illuminated the dual nature of tones as suprasegmentals in that both extrinsic and intrinsic acoustic information contribute to the description of a tone. Tones do not depend on absolute acoustic values to gain their identity. Rather, they contrast with other tones in the utterance as well as speaker $F0$ range to attain a relative identity. These assumptions are consistent with the results of this study showing that listeners use speaker $F0$ range in tone identification. Despite their intimate relationship with context, however, lexical tones also exist as independent phonological units, contrasting intrinsic acoustic characteristics such as turning point and $\Delta F0$ for Mandarin tones 2 and 3. Thus it is possible that in addition to contextual information specifying speaker $F0$ range, intrinsic $F0$ may also enable listeners to establish a representation of speaker identity. This hypothesis is consistent with findings by Slawson (1967) and Johnson (1990) for vowel perception, and Mullenix *et al.* (1989) for word recognition. The use of both extrinsic and intrinsic acoustic information in identifying speaker $F0$

range also avoids the ''bootstrap'' problem (Nearey, 1989), which confronts the issue of how listeners are able to establish a representation of speaker identity without precursor acoustic information.

ACKNOWLEDGMENTS

This research was conducted as part of a doctoral dissertation at Cornell University by the first author under the direction of the second author. The work was supported in part by a grant from Sigma Xi. We thank Scott Gargash for technical assistance, Abby Cohn, Joan A. Sereno, Ratree Wayland, and editor Winifred Strange for valuable comments on earlier versions of the manuscript, and the Cornell Chinese Students Association for getting us in contact with native speakers. Parts of the research were presented at the spring meeting of the Acoustical Society of America, Washington DC, 1995 and the annual meeting of the Linguistic Society of America, New Orleans, 1995.

[1]Duration differences between the two tones may also be perceptually relevant (Blicher *et al.*, 1990), but will not be investigated in this study. Production data generally show that durations for both tones 2 and 3 are longer than for other tones and that tone 3 is longer than tone 2 (Dreher and Lee, 1966; Ting, 1971; Chuang *et al.*, 1972; Rumjancev, 1972; Lyovin, 1978; Nordenhake and Svantesson, 1983), perhaps because the nonprepausal form of tone 3 is shorter than in isolation.

[2]Other tokens of both tones 2 and 3 exhibited some degree of creak as well, although vowel onset and offset points were undisturbed. In these cases the creak was located in the middle of the vowel, and the expected formant structure returned before vowel offset.

[3]There has been much concern about whether female voices can be successfully synthesized given the current design of the Klatt synthesizer. Parameters now considered to improve the naturalness of synthesized female voices include breathiness, open quotient, and glottal waveform [see Klatt and Klatt (1990) for summary and experimental data]. Stimuli synthesized for the present experiment relied largely on manipulating traditional parameters of fundamental frequency and formant frequencies. In addition, a more breathy quality was modeled by setting a Delta parameter which filters the upper frequencies relative to the lower frequencies. Subjects reported hearing a female speaking, and were often surprised to learn the stimuli were not produced naturally.

[4]Although we have no direct indications as to whether our listeners perceived the precursor+target sequences as being produced by the same speaker, our instructions encouraged listeners to process the sequences as unitary. Specifically, listeners were told that they would hear sentences produced by two different speakers and that they were to listen to the entire sentence. After the experiments, the vast majority of listeners did not comment on any perceived mismatch between precursor and target, while only a few listeners remarked that one precursor made a more unitary sequence with the target than the other. Of course, in the latter case, it is unkown whether that impression was based on $F0$ or voice quality.

[5]This restriction to a high-tone context served as a control for tonal coarticulation cues which may have been present in carriers preceding tones 2 or 3. However, a study on coarticulation in Mandarin tones by Shen (1990) shows that there is no anticipatory effect on $F0$ height or direction from tones 2 or 3, and particularly no effect from those tones on a preceding tone 4. The similar $F0$ onset of tones 2 and 3 probably obviates coarticulation, since it would be in anticipation of the onset $F0$ height that anticipatory coarticulation would occur (Shen, 1990).

[6]We thank reviewer Rob Fox for raising these alternative interpretations.

Assmann, P., Nearey, T., and Hogan, J. (**1982**). ''Vowel identification: Orthographic, perceptual and acoustic aspects,'' J. Acoust. Soc. Am. **71,** 975–989.

Blicher, D. L., Diehl, R., and Cohen, L. B. (**1990**). ''Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: evidence of auditory enhancement,'' J. Phon. **18**, 37–49.

Fok, Chan Yuen-Yuen. (**1974**). ''A Perceptual Study of Tones in Cantonese,'' Occasional Papers and Monographs, Centre of Asian Studies (University of Hong Kong, Hong Kong), Vol. 18.

Chao, Y-R. (**1968**). *A Grammar of Spoken Chinese* (University of California, Berkeley).

Charif, R. A., Hertz, S. R., and Weber, T. J. (**1992**). *Delta System User's Guide* (Eloquent Technology, Ithaca, NY).

Chuang, C. K., Hiki, S., Sone, T., and Nimura, T. (**1972**). ''The acoustical features and perceptual cues of the four tones of Standard Colloquial Chinese,'' Proceedings of the Seventh International Congress on Acoustics, Budapest, p. 297–300.

Coster, D. C., and Kratochvil, P. (**1984**). ''Tone and stress discrimination in normal Beijing dialect speech,'' New Papers on Chinese Language Use (Canberra), p. 119–132.

Dreher, J., and Lee, P. C. (**1966**). ''Instrumental investigation of single and paired Mandarin tonemes,'' Res. Commun. **13**, Douglas Advanced Research Laboratories.

Finney, D. J. (**1971**). *Probit Analysis* (Cambridge U.P., Cambridge, England).

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

Fujisaki, H., and Kawashima, T. (**1971**). ''A model of the mechanisms for speech perception: Quantitative analysis of category 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. A. Fromkin (Academic, New York), pp. 41–76.

Gårding, E., Kratochvil, P., Svantesson, J.-O., and Zhang, J. (**1986**). ''Tone 4 and Tone 3 Discrimination in Modern Standard Chinese,'' Language Speech **29**, 281–293.

Gottfried, T. L., Miller, J. L., and Payton, P. E. (**1990**). ''Effect of speaking rate on the perception of vowels,'' Phonetica **47**, 155–172.

Howie, J. M. (**1976**). *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge U.P., Cambridge, England).

Hsu, V. L. (**1990**). *A Reader in Post-Cultural Revolution Chinese Literature* (The Chinese University of Hong Kong, Hong Kong), pp. 344–381.

Johnson, K. (**1990**). ''The role of perceived speaker identity in $F0$ normalization of vowels,'' J. Acoust. Soc. Am. **88,** 642–654.

Johnson, K. (**1991**). ''Differential effects of speaker and vowel variability on fricative perception.'' Language Speech **34**, 265–279.

Jongman, A., and Miller, J. D. (**1991**). ''Method for the location of burst-onset spectra in the auditory-perceptual space: A study of place of articulation in voiceless stop consonants'', J. Acoust. Soc. Am. **89,** 867–873.

Kiriloff, C. (**1969**). ''On the auditory perception of tones in Mandarin,'' Phonetica **20**, 63–67.

Klatt, D. (**1980**). ''Software for a cascade/parallel formant synthesizer,'' J. Acoust. Soc. Am. **67,** 971–995.

Klatt, D., and Klatt, L. C. (**1990**). ''Analysis, synthesis and perception of voice quality variations among female and male talkers,'' J. Acoust. Soc. Am. **87,** 820–857.

Kratochvil, P. (**1971**). ''An experiment in the perception of Peking dialect,'' in *A Symposium on Chinese Grammar* (Scandanavian Institute of Asian Studies), pp. 7–31.

Ladefoged, P., and Broadbent, D. E. (**1957**). ''Information conveyed by vowels,'' J. Acoust. Soc. Am. **29,** 98–104.

Leather, J. (**1983**). ''Speaker normalization in perception of lexical tone,'' J. Phonetics **11**, 373–382.

Li, C., and Thompson, S. (**1981**). *Mandarin Chinese: A Functional Reference Grammar* (University of California, Berkeley).

Li, C. N., and Thompson, S. (**1978**). ''The acquisition of tone in Mandarin-speaking children,'' in *Tone: A Linguistic Survey*, edited by V. A. Fromkin (Academic, New York).

Lin, M. C. (**1988**). ''Putong hua sheng diao de sheng xue texing he zhi jue zhengzhao, [Standard Mandarin tone characteristics and percepts],'' Zhongguo Yuyan **3**, 182–193.

Lin, T., and Wang, W. Y.-S. (**1985**). ''Shengdiao ganzhi wenti [tone perception],'' Zhongguo Yuyan Xuebao **2**, 59–69.

Lyovin, A. (**1978**). ''Review of Tone and Intonation in Modern Chinese by M. K. Rumjancev,'' J. Chinese Ling. **6**, 120–168.

Mann, V. A., and Repp, B. H. (**1980**). ''Influence of vocalic context on the perception of the [ʃ]-[s] distinction,'' Percept. Psychophys. **23,** 213–228.

Mertus, J. (**1989**). *BLISS Manual* (Brown University, Providence, RI).

Moore, C. B. (**1993**). ''Some observations on tones and stress in Mandarin

Chinese,'' Working Papers of the Cornell Phonetics Laboratory **8**, 82–117.
Moore, C. B. (**1995**). ''Speaker and rate normalization in the perception of lexical tone by Mandarin and English listeners,'' Ph.D. dissertation, Cornell University.
Moore, C. B., and Jongman, A. (forthcoming). ''Cross-language effects in the perception of Mandarin tone.''
Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). ''Some effects of talker variability on spoken word recognition,'' J. Acoust. Soc. Am. **85,** 365–378.
Nearey, T. M. (**1989**). ''Static, dynamic, and relational properties in vowel perception,'' J. Acoust. Soc. Am. **85,** 2088–2113.
Nordenhake, M., and Svantesson, J.-O. (**1983**). ''Duration of Standard Chinese word tones in different sentence environments,'' Working Papers **25** (Lund, Sweden), 105–111.
Norman, J. (**1988**). *Chinese* (Cambridge U.P., Cambridge, England).
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). ''Speech perception as a talker-contingent process,'' Psychol. Sci. **5**, 42–46.
Peterson, G. E., and Barney, H. L. (**1952**). ''Control methods used in a study of the vowels,'' J. Acoust. Soc. Am. **24,** 175–184.
Pisoni, D. B. (**1975**). ''Auditory short-term memory and vowel perception,'' Mem. Cogn. **3**, 7–18.
Ruhm, H. B., Mencke, E. O., Milburn, B., Cooper, Jr., W. A., and Rose, D. E. (**1966**). ''Differential sensitivity to duration of acoustic stimuli,'' J. Speech Hear. Res. **9,** 371–384.
Rumjancev, M. K. (**1972**). ''Ton i Intonacija v Sovremennon Kitajskom Jazyke [Tone and Intonation in Modern Chinese] (Izdatel'stvo Moskovskogo Universiteta, Moscow),'' reviewed by A. V. Lyovin (1978). J. Chinese Ling. **6**, 120–168.
Shen, X. (**1990**). ''Tonal coarticulation in Mandarin,'' J. Phonetics **18**, 281–285.
Shen, X., and Lin, M. (**1991**). ''A perceptual study of Mandarin Tones 2 and 3,'' Language Speech **34**, 145–156.
Shen, X., Lin, M., and Yan, J. (**1993**). ''*F*0 turning point as an *F*0 cue to tonal contrast: A case study of Mandarin tones 2 and 3,'' J. Acoust. Soc. Am. **93,** 2241–2243.
Shigeno, S. (**1986**). ''The auditory tau and kappa effects for speech and nonspeech stimuli,'' Percept. Psychophys. **40,** 9–19.
Shigeno, S. (**1991**). ''Assimilation and contrast in the phonetic perception of vowels,'' J. Acoust. Soc. Am. **90,** 103–111.
Shigeno, S., and Fujisaki, H. (**1979**). ''Effect of a preceding anchor upon the categorical judgment of speech and nonspeech stimuli,'' Jpn. Psychol. Res. **21**, 165–173.
Shinn, P. C., Blumstein, S. E., and Jongman, A. (**1985**). ''Limitations of context conditioned effects in the perception of [b] and [w],'' Percept. Psychophys. **38,** 397–407.
Slawson, A. W. (**1967**). ''Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency,'' J. Acoust. Soc. Am. **43,** 87–101.
Sommers, M. S., Nygaard, L. C., and Pisoni, D. B. (**1992**). ''Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude,'' in *ICSLP 92 Proceedings: 1992 International Conference on Spoken Language Processing*, edited by J. J. Ohala, T M. Nearey, B. L. Derwing, M. M. Hodge, and G. E. Wiebe (Priority Printing, Edmonton, Canada), Vol. 1, pp. 217–220.
Stott, L. H. (**1935**). ''Time-order errors in the discrimination of short tonal durations,'' J. Exp. Psychol. **18,** 741–766.
Strange, W., Verbrugge, R., Shankweiler, D., and Edman, T. (**1976**). ''Consonant environment specifies vowel identity,'' J. Acoust. Soc. Am. **60,** 213–224.
Ting, A. C. (**1971**). ''Mandarin tones in selected sentence environments: An acoustic study,'' Ph.D. dissertation, University of Wisconsin.
Verbrugge, R. R., Strange, W., Shankweiler, D. P., and Edman, T. R. (**1976**). ''What information enables a listener to map a talker's vowel space?,'' J. Acoust. Soc. Am. **60,** 198–212.
Whalen, D. H. (**1981**). ''Effects of vocalic formant transitions and vowel quality on the English [s]-[ʃ] boundary,'' J. Acoust. Soc. Am. **69,** 275–282.
Zsiga, E. (**1994**). *Syllt: The Delta Syllable Tool* (Eloquent Technology, Ithaca, NY).