Working Papers of the Cornell Phonetics Laboratory
1995, vol. 10, pp. 97-144

# Speaker normalization in the perception of Mandarin Chinese tones

Corinne B. Moore

This study investigated speaker normalization in perception of Mandarin Tone 2 (mid-rising) 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 ΔF0 (the difference in F0 between onset and Turning Point) distinguished the two tones. Three tone continua varying in either Turning Point, ΔF0, 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 ΔF0, as compared to stimuli varying in both Turning Point and ΔF0, indicating that normalization effects are reduced for stimuli more closely matching natural speech.

## 1. 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 adjust to this acoustic variability, caused by differences in speaker vocal tract size, in order to identify segments accurately. Consequently, segments that have identical acoustic characteristics may not be perceived identically.

### 1.1 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, in which the formant frequencies of each version were manipulated 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". Subjects were asked to identify the test words in one of the precursor phrases. Results demonstrated that identification of the vowel stimuli changed depending on which version of the precursor phrase preceded them, and the change was in accordance with predictions about the relationship between F1 and F2 in the target words and the precursors. For example, one test word was identified as *bit* by 87% of the subjects when it occurred after one particular version of the precursor phrase, but as *bet* by 90% of subjects when it was preceded by a version of the precursor with a lower F1 value. The change of identification is predictable if the F1 in the stimulus is perceived according to the F1 in the precursor, since a lower F1 would generate the percept of a higher F1 in the target stimulus, and [ɛ] is known to have a higher F1 than [I]. Because identical formant values were perceived differently, according to the formant values in the precursors, this study has been used to support the hypothesis that vowels are perceived according to the vowel space of the speaker.

While Ladefoged and Broadbent's results indicate *what* acoustic information is used in speaker normalization, it remained unclear *how* this information is 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 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. Johnson hypothesized that if acoustic information is used directly, then changing speakers should have no effect on identification of vowel stimuli. On the other hand, if perception shifts as a function of the speaker, these results may be taken to indicate that listeners are using the acoustic information about the speaker in perception of the vowels (as a cue to 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 perception of vowels in isolation and in carrier phrases. The F0 levels of the vowel tokens were 100 Hz 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. The hypothesis of this experiment was that shifts in vowel identification observed for the 150 Hz versus the 100 Hz tokens would be reduced if the carrier phrase signalled that they were produced by the same speaker. This hypothesis was borne out by the results. 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. The hypothesis of this experiment was that there would be no significant difference in shifts for vowel tokens in isolation as compared to those in the carrier phrases, since there was no conflicting information about the speaker. This hypothesis was also confirmed by the results of the experiment. Finally, Johnson examined whether perception of the 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 experiment tested the hypothesis of Ladefoged and Broadbent (1957), which predicted that identical stimuli would be perceived differently if they were produced by different speakers. The results of Johnson's experiment confirmed this hypothesis as well. The three experiments in Johnson's study thus provide evidence that listeners use F0 as a cue to speaker identity, and that listeners normalize for this information in vowel perception.

### 1.2 Context Effects and Mandarin Chinese Tones

While the majority of studies have investigated normalization in the perception of vowels, virtually no experimental work has been done to examine speaker normalization in the perception of other types of speech sounds. The present study attempts to fill this gap by extending 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 mid-rising 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 Figure 1. In particular, the present series of experiments examines whether F0 range, as a cue to speaker identity, influences perception of Tones 2 and 3.

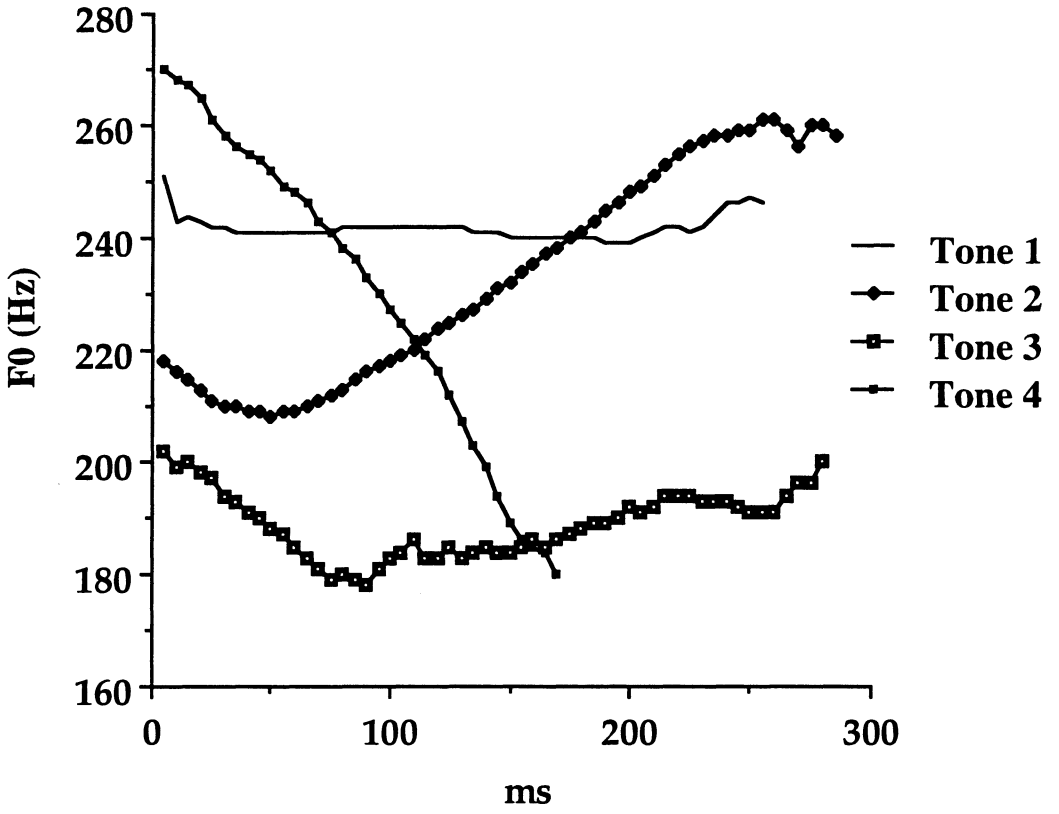



**Figure 1.** F0 contours for each of the four tones, taken from one token spoken in isolation by one of the speakers in this study (segmental context *ma*).

As suprasegmentals, tones are perceived relative to other tones, though 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 identical. 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.

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 mid-continuum stimuli (paired by speaker condition). These results showed at least one significant chi-squared value for each of the subjects, demonstrating that perception of at least one stimulus pair varied as a function of the speaker.

Unfortunately, however, Leather does not explicitly predict the type of responses he expects in each condition, nor do 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. For example, the identification shifts reported by Leather may have been in different directions, such that subjects who identified two stimulus pairs differently depending on the speaker may have classified one in an assimilatory direction (the higher pitch range produced more high tone responses) but contrastive in the other case. In addition, the results show inconsistency across speakers and stimuli. For example, several subjects identified only one stimulus pair among the four according to the speaker condition, while another did so for non-adjacent steps in the continuum. These inconsistencies suggest that if the analysis compared the entire identification functions for each subject rather than for selected stimuli, differences between the two sentence conditions may not have been robust enough to sustain the effect.

Subjects may also have identified the stimuli in a way not predicted by the hypothesis. For instance, the hypothesis may have predicted that the high precursor would trigger more Tone 2 responses for ambiguous stimuli (since the onset F0 of these stimuli is low relative to the precursor F0). It is impossible to determine, based on the chi-squared results, whether there were more Tone 1 or Tone 2 responses in the high

precursor condition. Furthermore, it is not possible to relate subject responses with the F0 information they received; only two stimulus pairs fell within the overlap in F0 range between the two speakers, but only one of those was ambiguous in both F0 height and contour, and for those stimuli, the reported results do not specify how they were identified in the two precursor conditions.

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. Finally, subjects in Leather's study responded to test items blocked by speaker, so it is uncertain if the normalization effect would still obtain in a mixed condition, which corresponds more closely to natural conditions.

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 is 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 non-native 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 presented chi-squared values for individual subject responses to four mid-continuum stimuli, showing inconsistent patterns of identification across subjects and stimuli. Among the 27 chi-squared values (9 subjects x 3 continuum steps), significant shifts were represented in only six of the Chinese subjects, all assimilatory, with five subjects showing no significant values. For the English subjects, five out of 27 chi-squared values were significant among three of the nine subjects, all but one assimilatory. This proportion compares to seven significant values out of 20 in Leather's study which compared responses to four mid-continuum stimuli for five subjects. Fox and Qi interpret 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. The experimental design should also present test items in a mixed condition, as opposed to a blocked condition as in Leather (1983), in order to more closely reflect the natural environment, and to reveal the robustness of any effect. 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 incorporates these issues into a new investigation of speaker normalization for Mandarin Chinese tones.

In this study, production and perception tests are 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 pre-pausal 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, Hiki, Sone and Namura 1972; Gandour 1978; Li and Thompson 1977). 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 Figure 2.

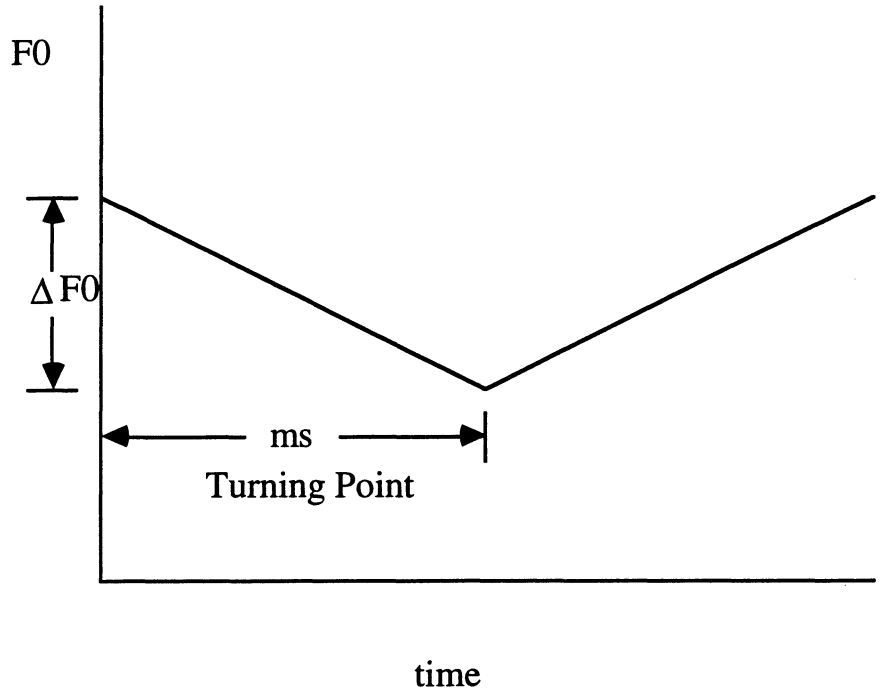



**Figure 2.** Turning Point and ΔF0 properties schematized for a contour tone.

Figure 2 illustrates how Turning Point and ΔF0 values for a tone are defined in this study. Perception studies of Mandarin Tones 2 and 3 have found that both timing of the Turning Point and ΔF0 are perceptually relevant for identification of the tones (Shen and Lin 1991; Shen, Lin and Yan 1993).[1]

Using these two acoustic dimensions, the present experiment examines 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-pitched speaker, but high tones for the low-pitched speaker. This scenario is schematized in Figure 3.

[1] Duration differences between the two tones may also be perceptually relevant (Blicher, Diehl and Cohen, 1990), but will not be investigated in this study. Production data generally show that durations for both Tones 2 and 3 are longerthan 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 non-prepausal form of Tone 3 is shorter than in isolation. An examination of duration differences between Tones 2 and 3 for each of the two speakers in this study found mixed results as well, showing no significant differences between the tone durations for one speaker, but significant differences for the other speaker.

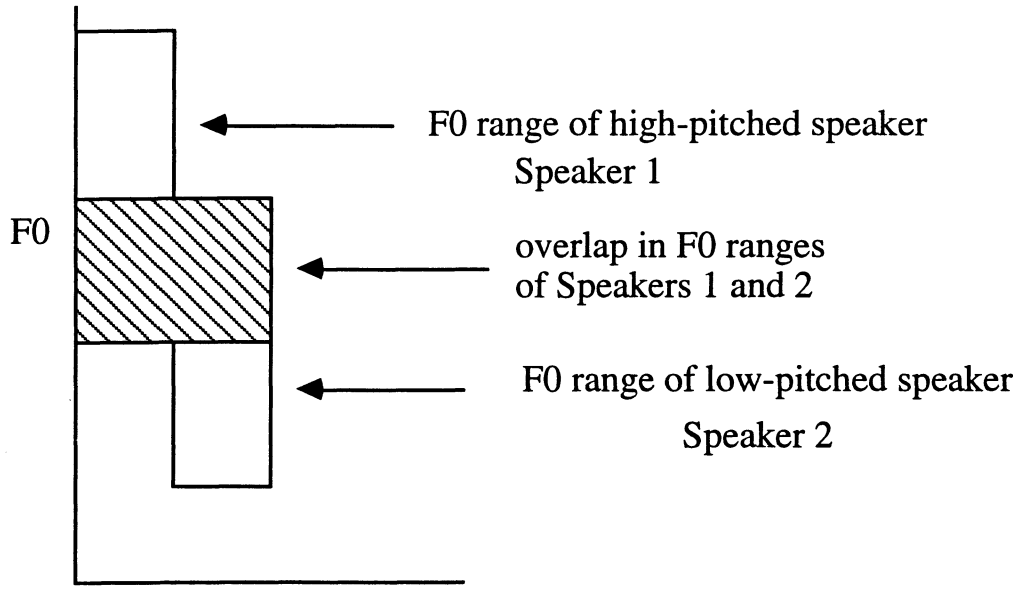



**Figure 3.** F0 ranges for two speakers overlap such that the low region of the range for Speaker 1 overlaps with the high region for Speaker 2. Normalization may occur for acoustically identical tones in the region of overlap.

The scenario in Figure 3 illustrates how two speakers may overlap in F0 range. Within that region of overlap, a high tone for the low-pitched speaker (Speaker 2) may be acoustically identical to a low tone for the high-pitched speaker (Speaker 1). In the example just described, identification of the tone would 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 will be compared to determine whether changes in speaker identity effect changes in tone identification.

The structure of the paper is as follows. Production data for speaker F0 ranges and Mandarin Tones 2 and 3 are given in Section 2. Perception data for tone stimuli varying in Turning Point and ΔF0 are provided in Section 3. In Sections 4-7, methodology and results of the normalization experiments are described, followed by a general summary and conclusion.

## 2. 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.

### 2.1 Method

#### 2.1.1 Subjects

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

#### 2.1.2 Materials

The data collected were from three reading tasks. The first of these asked subjects to read a long passage of text 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.

### 2.1.3 Procedure and Analysis

Subjects read the materials in an IAC sound-proof 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 five milliseconds, then filtered out F0 values corresponding to a probability of voicing of less than 99%. This was done in case the program erroneously calculated F0 points for non-voiced portions of the passage. The F0 values were then organized into histograms showing number of samples as a function of F0. 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 five milliseconds. 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 data for tones in isolation and in carrier phrases were measured at 5 ms intervals, 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. In cases where the valley was constant for longer than 5 ms (exhibited by more than one measurement point on the pitch track), the rightmost point before the increase in F0

values was taken to be the offset of the Turning Point measurement. ΔF0 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 phrases comprised a corpus of 24 instances of each tone per speaker (4 syllable types x 3 isolation repetitions x 3 carrier phrase 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.

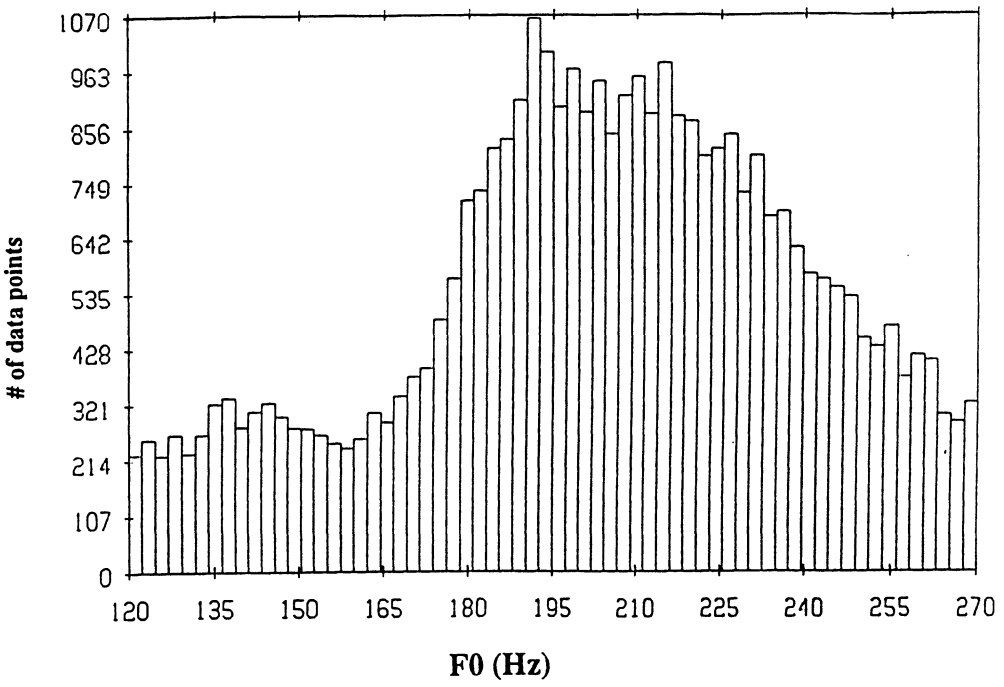
The next section will present results of speaker F0 range analysis, followed by analysis of Tones 2 and 3.

## 2.2 Results

### 2.2.1 F0 range data

Analysis of F0 range data was conducted for the long passage, generating roughly 40,000 data points for each speaker. These data were organized in the histograms.

Among seven speakers analyzed, F0 range data for two of the four female speakers were found to meet the requirements of the normalization study. Figure 4 reports the F0 range data for these two speakers.



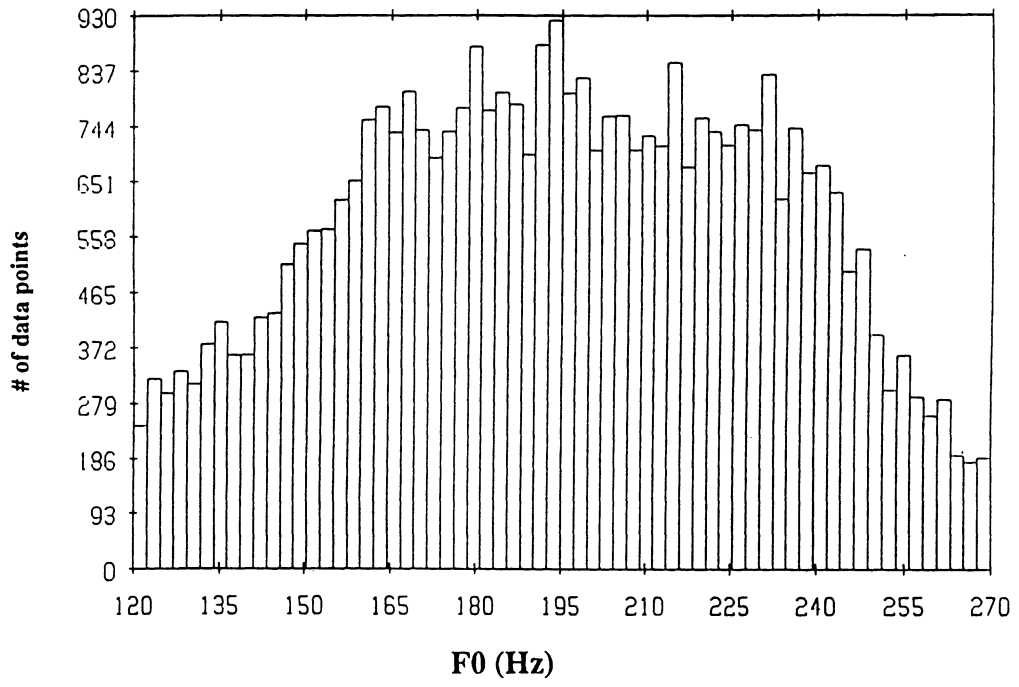




**Figure 4.** F0 range data from the long reading task for Speaker 1 (top panel), and Speaker 2 (bottom panel). The horizontal axis denotes F0 range; the vertical axis denotes number of data points for each F0 level. Mean F0 for Speaker 1 is 212 Hz, mode is 190 Hz. Mean F0 for Speaker 2 is 186 Hz, mode is 194 Hz.

Figure 4 shows mean, mode and F0 range data calculated between 120 Hz and 270 Hz for each speaker. F0 points above 270 Hz decreased gradually in number until 300 Hz; below 120 Hz a relatively small number of F0 points were recorded and the amount remained relatively stable. The histograms show a large peak at the mode F0 value, decreasing at the upper and lower regions. Speaker 1 (hereafter S1) shows a mean of 212 Hz and a mode of 190 Hz. For Speaker 2 (hereafter S2), the mean is 186 Hz and the mode 194 Hz. These data show a large area of overlap in F0 range, with a non-overlapping region at the low end of the distribution, corresponding to S2. Although the overlap extends to the high end of the region, the means reflect that S1 produces more consistently in a higher range than S2.

### 2.3 Discussion

Among the seven speakers tested, these two female speakers 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 mid-rising 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 produces the tone.

To verify that tones for the two speakers also overlap in F0 level, the next section presents Tone 2 and Tone 3 data for the two speakers, and describes the acoustic properties for the tones that are used in subsequent experiments.

### 2.4 Mandarin Tone 2 and Tone 3 analysis

Figure 5 shows the F0 contours of Tones 2 and 3 for both speakers.

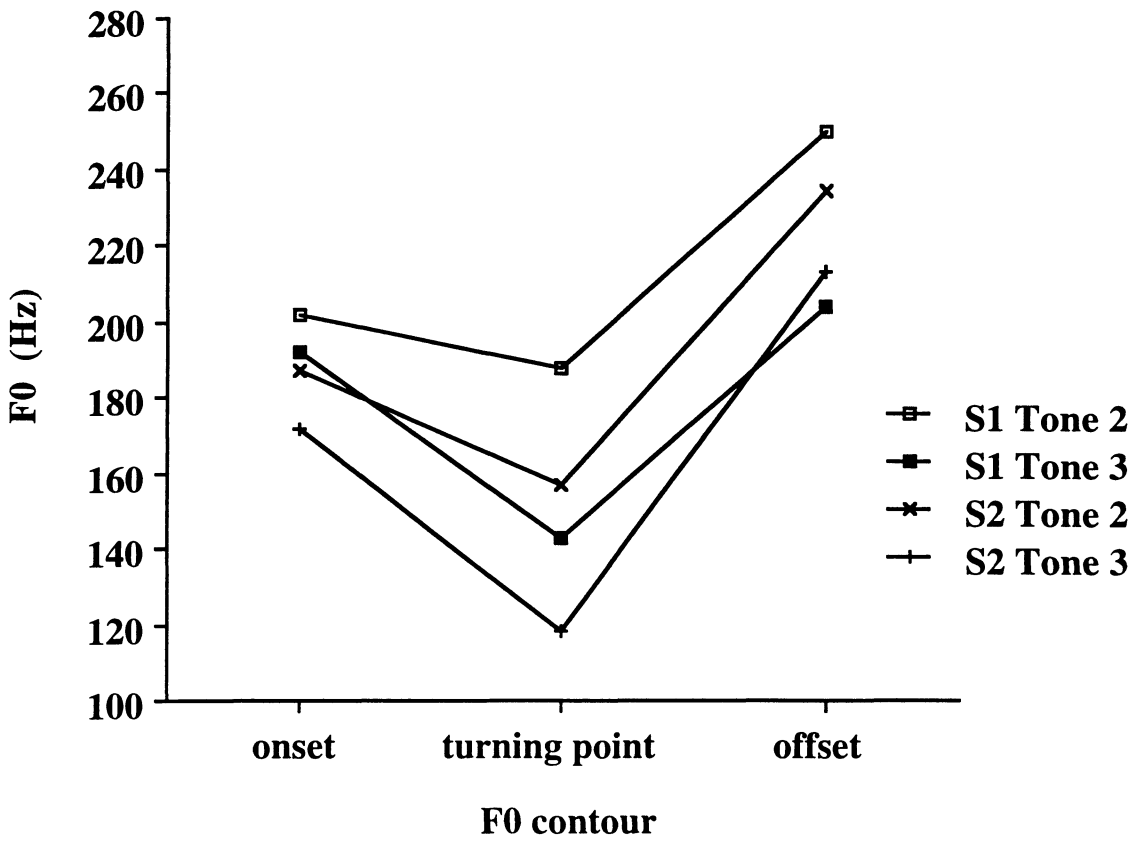



**Figure 5.** Averaged F0 contours of Tones 2 and 3 for the two female speakers (S1 and S2), across all syllable types.

The F0 contours in Figure 5 represent Tone 2 and Tone 3 average F0 onset, Turning Point and offset values for all syllable types in the isolation and carrier conditions. Although the figure does not include a time dimension to show realistic F0 contours, it shows that the two tones are very similar in F0 at onset, and have a similar falling-rising contour. The crucial observation in Figure 5 is that 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 192 Hz, falling to 157 Hz at the Turning Point, and ending at 234 Hz; the Tone 3 of S1 has an onset of 187 Hz, falling to 143 Hz, and ending at 204 Hz. The other two tones for each speaker, Tone 2 for S1 and Tone 3 for S2, are produced outside of the region of tonal overlap.

Recall that the F0 range data in Figure 4 indicate that the two female speakers share 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 occur precisely in this region, a pattern that is predicted by the F0 range data. Figure 6, which contains both F0 and duration information, exemplifies how tones in this region are produced at similar F0 heights.

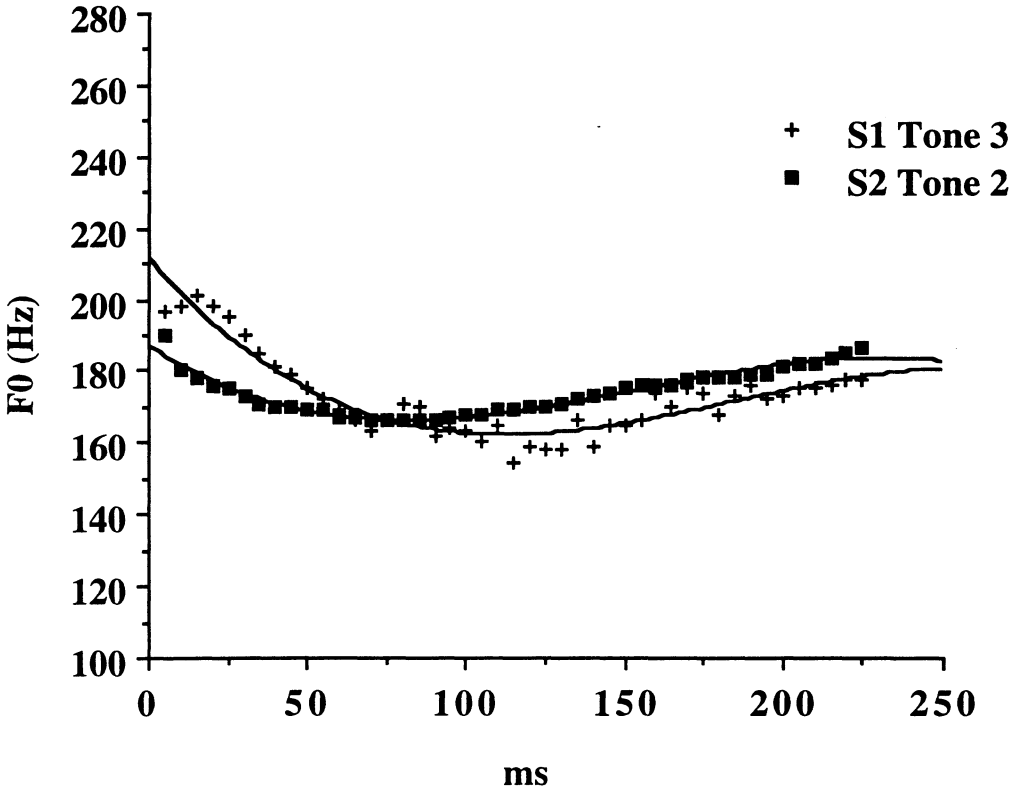



**Figure 6.** F0 contours for one [u] token each of Tone 2 of S2 and Tone 3 of S1, from the isolation production task.

The data in Figure 6 show both duration and F0 aspects of representative tone tokens. The similar F0 level of these two tokens attests to the likelihood that Tone 2 for S2 and Tone 3 for S1 may be acoustically identical. They are, thus, appropriate to use in subsequent tests for normalization. The following section describes them in more phonetic detail.

## 2.5 Acoustic characteristics of Tones 2 and 3

The preceding section established that Tone 2 for S2 and Tone 3 for S1 may be produced at equivalent F0 levels, and thus satisfy the scenario conducive to testing for normalization. Further acoustic analysis of these tones defined parameters used in subsequent tests using synthesized stimuli. The measurements taken included tone duration (taken to be equivalent to vowel duration), Turning Point (TP) and ΔF0.

Mean duration measurements taken for each tone according to the vowel in each syllable type are shown in Table 1.

| Tone2 (S2) | ma 'hemp' | yi 'move' | bi 'nose' | wu 'nothing' |
|---|---|---|---|---|
| mean | 278 | 356 | 331 | 363 |
| | | | | |
| Tone3 (S1) | ma | yi | bi | wu |
| mean | 288 | 388 | 320 | 377 |

**Table 1.** Average duration (ms) of Tone 2 for S2 and Tone 3 for S1, according to syllable type.

Table 1 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. One token of S1's *ma* was excluded because the presence of creak made location of vowel offset impossible.[2] S2's Tone 2 durations range from 244 ms to 415 ms; S1's Tone 3 range is 324 to 485 ms. Average duration for the Tone 2 (S2) tokens is 332 ms, as compared with 346 ms for Tone 3 (S1) tokens. An unpaired, two-tailed t-test showed this difference not to be significant [$t(45) = -.89$, $p > .38$]. A large area of overlap is thus represented in these Tone 2 and 3 tokens: from 324 to 415 ms.

Within each syllable type, statistical analyses between Tone 2 (S2) and Tone 3 (S1) show that the differences are not significant in each case (*yi*: [$t(10) = -1.73$, $p > .12$]; *bi*: [$t(10) = .49$, $p > .64$]; *wu*: [$t(10) = -.48$, $p > .65$]; and *ma*: [$t(9) = -0.49$, $p > .64$]).

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 milliseconds, and also as a percentage of tone duration. Tone 2 Turning Point values averaged 66 ms, occurring at an average of 20% into the tone. Tone 3 showed an

[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.

average Turning Point of 139 ms, occurring at 35% into the tone on average. Turning Point values for Tone 2 ranged from 25 to 96 ms, as compared to 105 to 200 ms for Tone 3, demonstrating a significant difference between the two tones [$t(27)= -6.22$, $p<.001$]. Calculated as a percentage of total tone duration, Tone 2 ranges from 7-24% of the tone, while the Tone 3 range is 28-41%. This difference between the two tones is also significant [$t(27)= -4.76$, $p<.001$]. The Turning Point regions for Tones 2 and 3 are therefore quite distinct.[3]

In addition to Turning Point, the other acoustic parameter being observed is the decline in F0 from the onset of the tone to the Turning Point, or ΔF0. These data showed an average ΔF0 of 35 Hz for Tone 2 (S2), and 51 Hz for Tone 3 (S1). For Tone 2, ΔF0 ranged from 4 to 67 Hz, and from 24 to 106 Hz for Tone 3. There is, thus, an area of overlap occurring between 24 to 67 Hz. Unpaired two-tailed t-test results show that ΔF0 differences between Tone 2 (S2) and Tone 3 (S1) are not significant [$t(23)= -1.7$, $p>.09$], but at a level suggesting a strong trend.

## 2.6 Discussion

Results of the tone analysis show Tones 2 and 3 to be similar in F0 height and contour. Both tones show a decline in F0 from the onset to the Turning Point, as well as a final rise. The data also show that Tone 2 for S2 and Tone 3 for S1 may be produced at a virtually equivalent F0 height in terms of onset and overall contour. Thus, Tone 2 for S2 occurs at roughly the same F0 height as Tone 3 for S1. Finally, duration differences between these two tones are not significant.

Differences were found between the tones, however. The intrinsic property of F0 showed that Tone 2 tends to have smaller decreases in F0 from the onset to the Turning Point than Tone 3, though these differences between the tones did not reach significance. The most conspicuous difference is in timing of the Turning Point: Tone 2 tokens showed significantly earlier Turning Points than Tone 3 tokens, both in absolute duration and as a percentage of tone duration.

The data in Experiment 1 show that the two female speakers share a region of overlap in 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 is that if listeners are to correctly identify these tones they must

[3]Since Turning Point is one of the acoustic parameters that will be manipulated in the creation of a Tone 2 - Tone 3 continuum, it would be desirable to have an area of overlap, rather than the completely distinct regions observed. Nevertheless, the 4% difference between the regions is probably under the JND for duration in this case (Henry, 1948; Ruhm, Mencke, Milburn, Cooper, and Rose, 1966; Lehiste, 1970).

normalize for speaker identity (F0 range). To test this hypothesis, perception of tone stimuli in high and low F0 precursors is examined, where tone stimuli form a continuum from Tone 2 to Tone 3, varying in ΔF0 and Turning Point characteristics. The next section describes how test items for these experiments were created, beginning with the tone stimuli.

### 3. Experiment 2: Perception of Turning Point and ΔF0 in isolation

Results of Experiment 1 suggest that timing of the Turning Point and ΔF0 distinguish Tones 2 and 3. Earlier studies using Tones 2 and 3 have also considered timing of the Turning Point and ΔF0 to be perceptual cues for these two tones. Shen and Yan (1991) constructed two Turning Point continua, one with a fixed ΔF0 of 15 Hz, the other 30 Hz. They found that subjects' perception shifted toward Tone 3 earlier for the stimuli whose ΔF0 was 30 Hz, than for stimuli with a ΔF0 of 15 Hz. Blicher, Diehl and Cohen (1990) created a Tone 2 to Tone 3 continuum which manipulated three dimensions: timing of the Turning Point, ΔF0, and F0 offset. However, neither of these earlier studies have documented a systematic investigation of these parameters which addresses whether ΔF0 and Turning Point covary, whether perception based on each of these parameters was equally categorical, or what combinations of Turning Point and ΔF0 trigger shifts in identification from one tone to the other. The values of Turning Point and ΔF0 used in the previous studies were based on production data, but data for the present experiment show enough variability in the production of Tones 2 and 3 to allow either tone to possess the particular F0 and duration characteristics of the previous studies' exemplars. Therefore, a perception experiment was devised to determine the relative importance of timing of the Turning Point and ΔF0. The experiment tests perception of isolated synthetic stimuli in which timing of the Turning Point and ΔF0 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 manipulations. In addition, results of this experiment will be used to more accurately model the synthetic stimuli used in subsequent tests.

## 3.1 Method

### 3.1.1 Subjects

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

### 3.1.2 Stimuli

In order to reduce or eliminate speaker-specific or tone-specific cues in the stimuli, the syllable [u] was chosen for synthesis because of the relatively similar durations between speakers and tones. 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, Hertz and Weber 1992; Zsiga 1994) and a Klatt synthesizer (1980) in the Cornell Phonetics Laboratory.[4] 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 formant. 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 2.

[4]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.

| | onset | offset |
|---|---|---|
| F1 | 345 | 304 |
| F2 | 703 | 628 |
| F3 | 2940 | 2940 |
| F4 | 4320 | 4320 |
| F5 | 4840 | 4840 |

**Table 2.** Formant frequency values (Hz) for synthesized stimuli

Duration of the stimuli was constant at 400 ms, well within the duration range for either Tone 2 or Tone 3. Amplitude of voicing began at 55 dB and declined to 53 dB over the duration of the token.

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 twelve stimuli. This continuum contained Turning Point information which 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. ΔF0 was varied from 10 to 70 Hz in steps of 5 Hz, generating a continuum of thirteen stimuli. Because Tone 2 typically exhibits a shallower ΔF0, it was expected that tones with a ΔF0 equal to 10 Hz would produce more Tone 2 responses than tones with a ΔF0 of 70 Hz.

The two continua together allowed for testing of both timing of the Turning Point and ΔF0, in an effort to understand how these acoustic parameters are used in the perception of Tones 2 and 3. Figure 7 represents all combinations of these parameters which were included in Experiment 2.

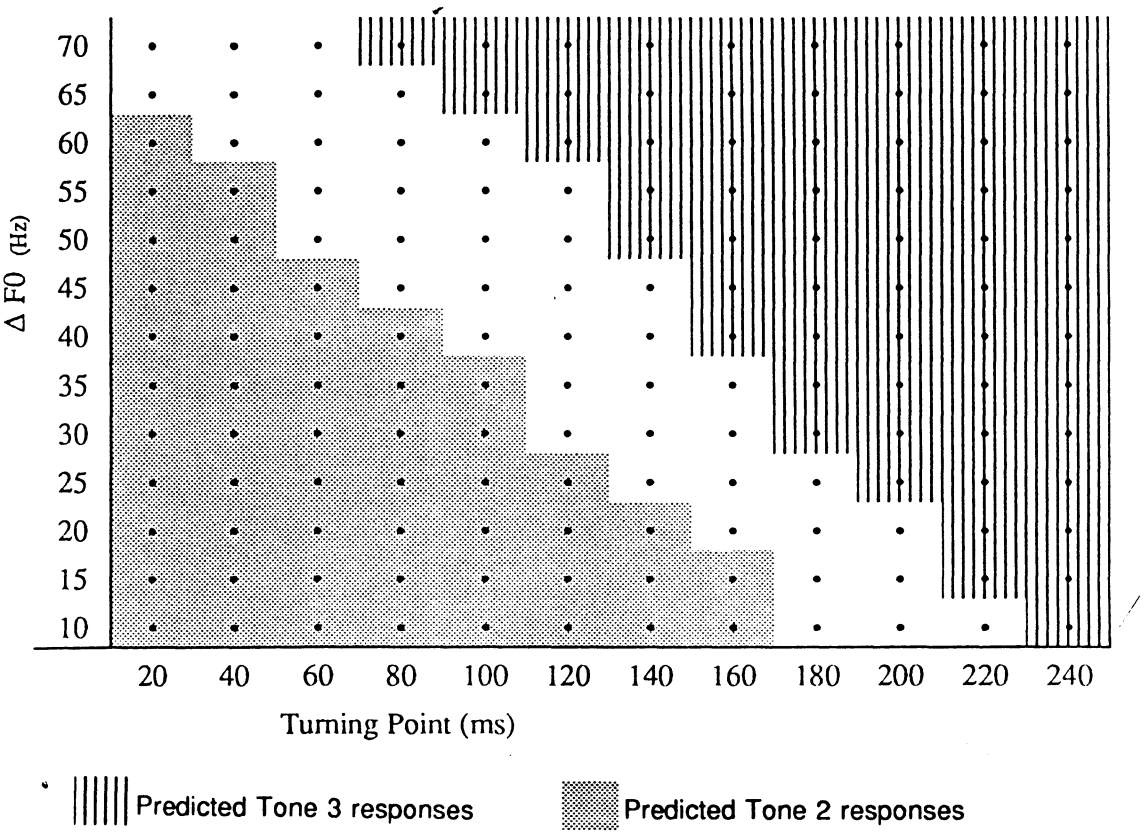



**Figure 7.** Combinations of Turning Point and ΔF0 manipulations for synthesized stimuli. Turning Point manipulations are represented along the horizontal axis, ΔF0 manipulations on the vertical axis. The shaded region corresponds to predicted Tone 2 responses, and the lined region corresponds to predicted Tone 3 responses.

Based on the duration and F0 manipulations represented in Figure 7, 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 Figure 7 which corresponds to the Tone 2 characterization contains stimuli with Turning Points from 20ms to approximately 100ms, along with lower ΔF0 values. It might also be expected that a high ΔF0 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 Figure 7.

### 3.1.3 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 tokens in total (12 Turning Point x 13 ΔF0 x 3 repetitions). Stimuli were low-pass filtered at 5.2 kHz and played out 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). A practice session consisting of 23 test items preceded the test. These practice items provided listeners with endpoints 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 therefore 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.

### 3.2 Results

Figure 8 gives number of Tone 2 responses for all stimuli, arranged according to values for timing of the Turning Point and ΔF0.

| Δ F0 (Hz) | 20 | 40 | 60 | 80 | 100 | 120 | 140 | 160 | 180 | 200 | 220 | 240 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 75 | | | | | | x | | | | | x | |
| 70 | 16 | 12 | 13 | 4 | 6 | 2 | 0 | 0 | 1 | 0 | 1 | 0 |
| 65 | 14 | 15 | 14 | 9 | 13 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| 60 | 18 | 17 | 13 | 8 | 9 | 6 | 3 | 3 | 0 | 4 | 0 | 0 |
| 55 | 16 | 18 | 18 | 13 | 7 | 4 | 3 | 1 | 0 | 1 | 1 | 0 |
| 50 | 18 | 14 | 16 | 14 | 13 | 9 | 9 | 6 | 1 | 1 | 1 | 0 |
| 45 | 18 | 16 | 18 | 14 | 12 | 15 | 5 | 3 | 7 | 3 | 2 | 1 |
| 40 | 18 | 17 | 17 | 17 | 16 | 16 | 15 | 15 | 11 | 5 | 0 | 4 |
| 35 | 17 | 17 | 18 | 17 | 17 | 17 | 10 | 5 | 3 | 2 | 3 | 0 |
| 30 | 17 | 18 | 18 | 18 | 18 | 18 | 15 | 14 | 17 | 14 | 14 | 15 |
| 25 | 18 | 18 | 18 | 18 | 18 | 18 | 17 | 16 | 9 | 16 | 14 | 10 |
| 20 | 18 | 18 | 18 | 16 | 18 | 18 | 18 | 18 | 16 | 10 | 12 | 8 |
| 15 | 18 | 17 | 17 | 17 | 18 | 17 | 15 | 18 | 18 | 16 | 16 | 18 |
| 10 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 17 | 18 | 17 | 15 |

Turning Point (ms)

**Figure 8.** Tone 2 responses for isolated stimuli varying in timing of the Turning Point and ΔF0. Eighteen responses were possible for each stimulus (6 subjects x 3 repetitions). Responses to stimuli from Experiment 2 determined which tone continua to use in Experiments 3a - 3c. These continua are enclosed in boxes: the diagonal boxes indicate stimuli varying along both Turning Point and ΔF0, the horizontal boxes represent stimuli varying only in Turning Point, and the vertical row of boxes shows stimuli varying only in ΔF0. An "x" denotes stimuli added after Experiment 2.

Figure 8 shows that, as expected, stimuli are clearly identified as Tone 2 in the region where Turning Point and ΔF0 values are low. Along the Turning Point dimension, unambiguous Tone 2 responses span virtually the entire continuum, up to a ΔF0 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.

Along the ΔF0 dimension, decreases of up to 70 Hz are still identified as Tone 2. According to Kratochvil (1971), who tested perception of synthetic tones with durations from 90 - 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 .026 by Ruhm et al. (1966), and for 400 ms signals as .12 by Stott (1935), corresponding to approximately 10 - 60 ms for the 400 ms stimuli in Experiment 2. If the initial 20 to 40 ms portion of the tone is imperceptible (below threshold), subjects may hear only a rise, not the initial fall, since ΔF0 is equivalent to 0 at the earliest Turning Points. ΔF0 begins to trigger Tone 3 responses when the Turning Point reaches approximately 80 ms into the tone. At this point, Tone 3 responses increase as a function of both ΔF0 and timing of the Turning Point. When the Turning Point occurs as late as 200 ms into the tone, however, relatively low ΔF0 values (approximately 30 Hz or greater) elicit Tone 3 responses. Thus, the later the Turning Point, the easier it is for ΔF0 to effect a change, though even late Turning Points are resilient to the effects of a ΔF0< 30 Hz. On the other hand, ΔF0 appears to trigger more Tone 3 responses in stimuli with later Turning Points rather than early ones.

### 3.3 Discussion

The purpose of Experiment 2 was to determine how the acoustic dimensions of timing of the Turning Point and ΔF0 contribute to perception of Mandarin Tones 2 and 3. As for which of the two acoustic dimensions might be more important, the data suggest that there is an interdependency between ΔF0 and timing of the Turning Point. It appears that ΔF0 is more relevant as Turning Point increases, while Turning Point is more relevant for a ΔF0 of more than 30 Hz. For example, a tone with an initial fall of 30 Hz will be perceived the same as a tone with an initial fall of 10 Hz for any Turning Point, but a tone with a Turning Point at 160 ms will be perceived differently if its ΔF0 is 30 Hz or 45 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 ΔF0 appear to be interdependent, there are places where either dimension alone is sufficient to produce categorical functions. For example, stimuli with a constant Turning Point of 140 ms show all Tone 2 responses for a ΔF0 of 20 Hz, moving to 50% responses for a ΔF0 of 50 Hz, and all Tone 3 responses for the

largest ΔF0. Along the Turning Point dimension, Tone 2 responses move categorically from 100 to 0% for the continuum of stimuli with a constant 50 Hz ΔF0. While either of the two acoustic parameters are robust enough to trigger categorical identification functions, it is clear from both the production and perception data that timing of the Turning Point and ΔF0 operate in tandem as perceptual cues to Tones 2 and 3.

In summary, the results of this perception test show how tone stimuli which vary in ΔF0 and timing of the Turning Point are perceived. Subjects made categorical responses based on these two acoustic dimensions, such that identification functions may be 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 ΔF0 and Turning Point examined above.

## 4. Experiments 3a-3c: Perception of stimuli in precursor phrases

The following three experiments test the hypothesis that listeners perceive tones in part by normalizing for speaker F0 range.

### 4.1 Method

Experiments 3a-3c employ a design which compares how identical stimuli are identified in two contexts differing only in speaker identity. This type of test ensures that the effect is caused by normalization of different talker characteristics; the test uses naturally spoken carrier phrases from different speakers to serve as precursors. Normalization effects in this experiment would cause a shift in identification of stimuli as a function of which precursor phrase is heard, high or low F0.

As previously noted, earlier experiments on tones have provided evidence that external acoustic information may influence perception, though only Leather (1983) examined normalization effects due to perceived speaker identity. Experiments 3a-3c of the present study expand on Leather's work by testing two different Mandarin tones, Tones 2 and 3, and seek to provide more robust evidence of normalization. This will be done in several ways: by examining the direction of any shift in identification relative to the precursor, by measuring shifts in identification over the entire function, rather than arbitrarily selected points in the middle, and by presenting stimuli in a mixed condition, rather than blocked by speaker.

### 4.1.1 Tone stimuli for Experiments 3a-3c

The three experiments were conducted based on the perception data from Experiment 2. Each experiment is distinguished according to the stimuli used: Experiment 3a employed a continuum of stimuli containing cues about both timing of the Turning Point and ΔF0; stimuli for Experiment 3b included a continuum of stimuli varying only ΔF0, and Experiment 3c used a continuum varying only timing of the Turning Point. The continua marked by boxes in Figure 8 were used in these experiments. All three continua share a common midpoint which has a Turning Point of 120 ms, and a ΔF0 of 50 Hz. This midpoint stimulus received 50% Tone 2 responses for the corresponding identification functions resulting from Experiment 2.

## 5. Experiment 3a: Perception of stimuli varying in ΔF0 and Turning Point

### 5.1 Method

### 5.1.1 Subjects

Eleven subjects, seven male and four female, aged between 19 and 40, participated in this experiment. All are 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.

### 5.1.2 Stimuli: Tone continuum

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 ΔF0. This continuum is the diagonal set of stimuli shown in Figure 8. A schematized version of these stimuli is shown in Figure 9.

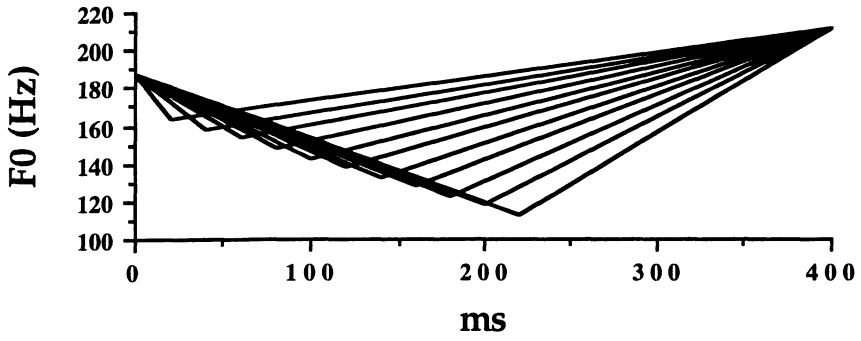



**Figure 9.** F0 contours of the Experiment 3a stimuli which varied in both the Turning Point and ΔF0 acoustic dimensions.

The schematized tone contours in Figure 9 illustrate manipulations in both the Turning Point and ΔF0 acoustic dimensions. 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. ΔF0 ranged from 25 Hz to 75 Hz, in steps of 5 Hz.

### 5.1.3 Stimuli: Precursor phrases

In addition to these stimuli, 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). Analysis revealed that phrases for the two speakers were not sufficiently distinct in average F0, in contrast with the data from the reading task. Instead, a phrase from another female participant in the study was selected to replace S2. The mean F0 for the carrier phrase of the substituted speaker was 187 Hz, which was very similar to the overall mean of 186 Hz for S2. Table 3 summarizes the acoustic information for the two precursors.

| 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 |

**Table 3.** F0 and duration information for precursors used in Experiments 3a-3c.

Table 3 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 differ by 39 Hz in average F0, but are further distinguished by the range; the high precursor spans 192-272 Hz, as compared to 170-229 Hz for the low precursor. In order to visualize how the stimuli are situated with respect to these F0 ranges, recall that the synthesized stimuli have a fixed onset and offset of 188 Hz and 212 Hz, respectively, levels which were based upon production data. The ΔF0 value decreases from 163 to 113 Hz in the continuum containing both ΔF0 and Turning Point cues, and also in the ΔF0 continuum.

Because they are naturally produced, the phrases differ in voice quality as well as F0 range. (Formant frequencies of the target stimuli, however, were synthesized to be ambiguous relative to the precursors to control for an effect of voice quality (see Table 3.2).) Other than these differences, the phrases were identical. Each 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 hear one instance of the test tones, rather than possibly comparing precursor examples of the test tones with the stimuli.

[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).

### 5.1.4 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 x 10 repetitions), with an inter-trial 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 continuum of manipulations.

## 5.2 Results

Results of Experiment 3a are summarized in figure and table form below. Figure 10 shows the percentage of Tone 2 responses for stimuli in the two presentation types (high and low precursor conditions), averaged across subjects. Table 4 lists probit values for each subject as a function of the precursor conditions.

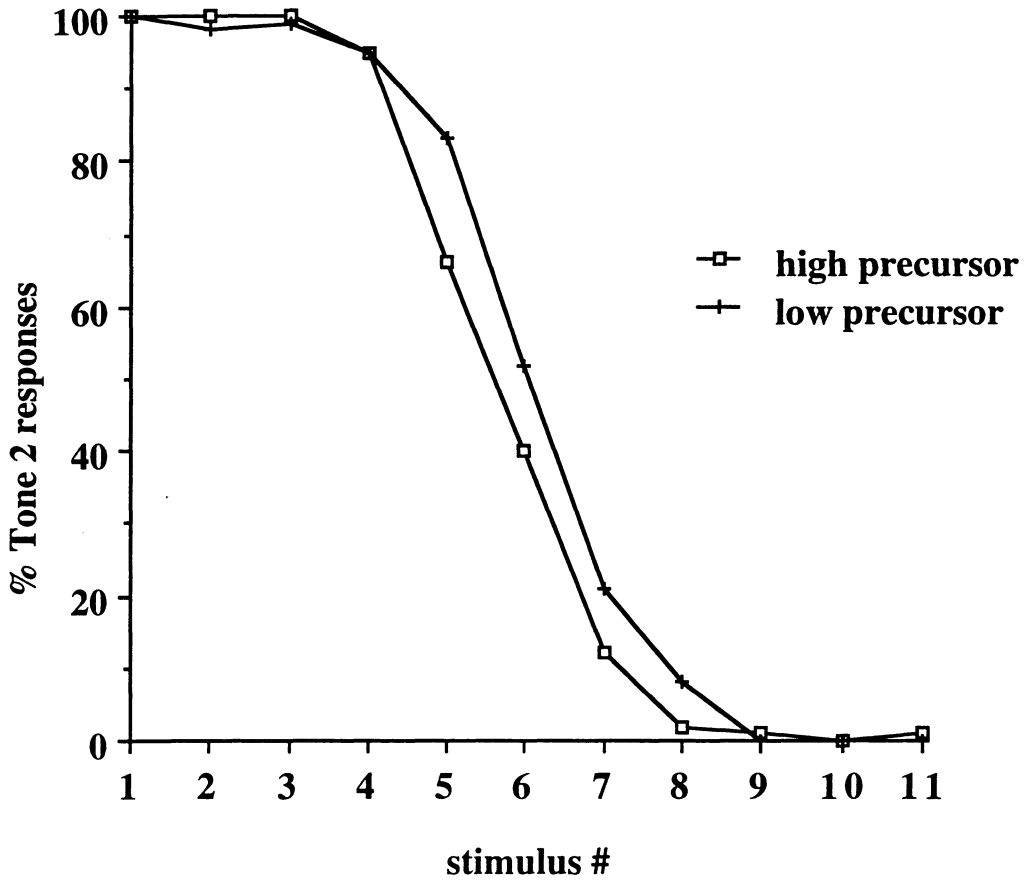



**Figure 10.** Experiment 3a Turning Point/ΔF0 continuum identification functions for high and low precursor conditions, averaged across subjects. Stimulus 1 corresponds to predicted Tone 2 responses.

| 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 4.** Experiment 3a probit values for Turning Point/ΔF0 stimuli in high and low precursor conditions.

The probit values in Table 4 represent category boundaries for the listeners who participated in Experiment 3a. As shown in Figure 10, 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, at 5.60, as compared to 5.99 for the low precursor. A paired two-tailed t-test shows this difference between boundaries to be significant [$t(10) = -2.57$; $p<.03$]. Subjects thus appear to refer to the F0 range of the precursor in perception of the tones. Moreover, the normalization effect is robust enough to be obtained in a mixed block condition, in comparison to 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 causes 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.

The next section reports results of subject responses for stimuli in which only ΔF0 was manipulated.

## 6. Experiment 3b: Perception of stimuli varying in ΔF0

### 6.1 Method

#### 6.1.1 Subjects

22 native speakers of Mandarin Chinese participated in this experiment. Twelve of these were excluded from the results on the basis of criteria outlined in section 6.1.3. The ten remaining subjects included 5 males and 5 females. One of the subjects is from Taiwan, and nine are from Mainland China. None reported any hearing disorders.

#### 6.1.2 Stimuli

Test items were sentences composed of the two precursors used in Experiment 3a, followed by a test word taken from the ΔF0 continuum in the pretest. This continuum varied ΔF0 in 11 steps of 5 Hz from 163 Hz. The timing of the Turning Point was fixed at 120 ms.

#### 6.1.3 Procedure and Analysis

The test procedure and analysis of data were identical to those used in Experiment 3a. However, this experiment seemed to be more difficult for subjects than Experiment 3a, judging both by number of missed trials and failure to achieve categorical identifications at continuum endpoints. 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 endpoints. Failure to meet these criteria led to the disqualification of 12 subjects.

### 6.2 Results

The averaged identification functions for the ΔF0 continuum in the high and low precursor conditions are presented in Figure 11, and the probit values are listed in Table 5.

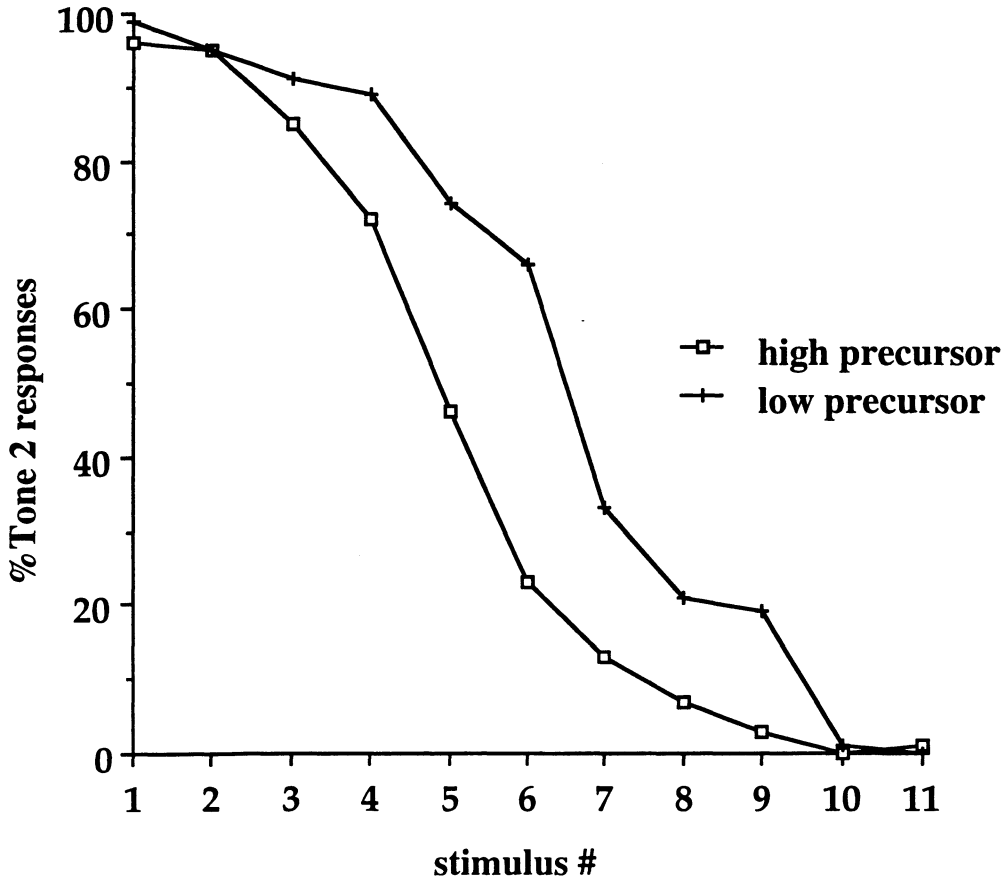



**Figure 11.** Experiment 3b ΔF0 identification functions for high and low precursor conditions, averaged across subjects.

| 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 |

**Table 5.** Experiment 3b probit values by subject for ΔF0 stimuli in high and low precursor conditions.

The data in Table 5 show that for the Tone 2 to Tone 3 continuum varying only in ΔF0, there is an earlier shift to Tone 3 responses in the high precursor condition for nine of the ten subjects: 4.46 as compared to 5.91 in the low precursor condition. This difference is significant [$t(9) = -4.69$, $p<.001$]. The shift is one of contrast--the high precursor prompts more low tone responses and vice versa. If an assimilatory effect had been observed, there would have been more high tone responses when stimuli were preceded by the high precursor. 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 is much greater than the shift in Experiment 3a. These differences, computed as the difference between the low and high precursor probits, were shown to be significant in a two-tailed, unpaired t-test of the shifts for each subject [$t(19)=-3.33$, $p<.003$]. These results indicate that listeners relied more on speaker F0 range to disambiguate the tones when the stimuli provided less intrinsic acoustic information. The implications of this finding will be discussed in section 8. The next section reports results of subject responses for stimuli in which only the Turning Point dimension was manipulated.

## 7. Experiment 3c: Perception of stimuli varying in Turning Point

### 7.1 Method

#### 7.1.1 Subjects

20 native speakers of Mandarin Chinese participated in Experiment 3c. Eight subjects were disqualified according to the criteria outlined in section 6.1.3. The twelve remaining subjects included six males and six females. Four subjects are from Taiwan, and eight are from Mainland China. None reported any hearing disorders.

#### 7.1.2 Stimuli

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 3a and 3b. The continuum from Tone 2 to Tone 3 varied only timing of the Turning Point, from 20 ms to 220 ms into the tone. The decrease in $\Delta F0$ was constant at 50 Hz.

#### 7.1.3 Procedure and Analysis

Test procedure and analysis of results are identical to those of Experiments 3a and 3b.

### 7.2 Results

Figure 12 displays average percent Tone 2 responses for the Turning Point continuum in the high and low precursor conditions. Table 6 lists probit values for each subject in both the high and low precursor conditions.

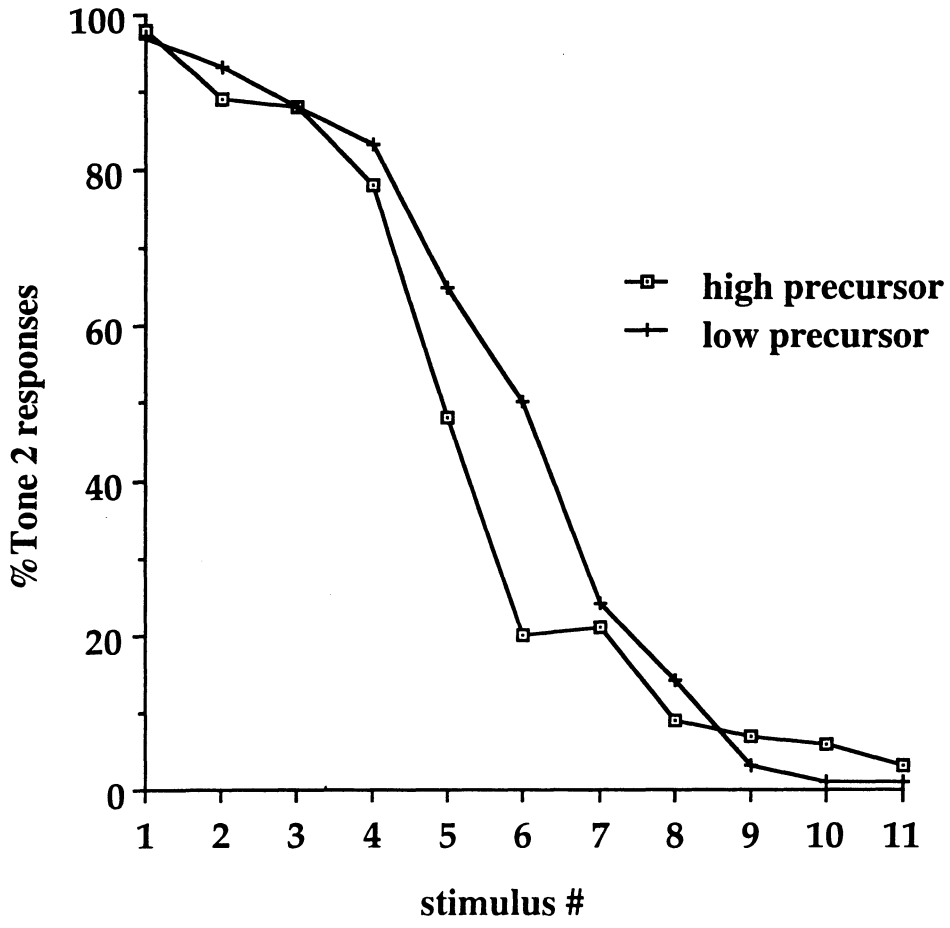



**Figure 12.** Experiment 3c Turning Point continuum identification functions for high and low precursor conditions, averaged across subjects.

| 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 |

**Table 6.** Experiment 3c probit values for Turning Point stimuli in high and low precursor conditions.

Table 6 lists probit values for subjects in the high and low precursor conditions. The average boundary in the high precursor condition as compared to the low is 4.78 vs. 5.20. Only eight of the twelve subjects show a shift in the direction predicted by the normalization hypothesis, and the difference between the probits in the two conditions is not significant [$t(11)= -1.55$, $p>.15$]. These results suggest that stimuli varying only in a duration dimension will not cause a normalization effect for contexts that vary in an F0 dimension.

## 8. General Summary and Discussion

The hypothesis of this study is that listeners use acoustic information about the speaker in the perception of lexical tones. In particular, the study investigates whether listeners use speaker F0 range in perception of intrinsic acoustic properties of Mandarin Tones 2 and 3. The hypothesis predicts that tone identification is affected by changes in speakers. If speaker information is not relevant in tone perception, on the other hand, changes in speakers should cause no significant shift in identification.

To examine the hypothesis, a series of production and perception experiments were conducted. First, production analyses from Experiment 1 located two speakers who share a region of F0 range overlap. The analysis revealed that within the area of F0 range overlap, a low tone for a high-pitched speaker and a high tone of a low-pitched speaker may occur at equivalent F0 heights. Further investigation of intrinsic acoustic properties of Tones 2 and 3 showed that the two tones may be distinguished in both F0 ($\Delta$F0) and temporal (Turning Point) dimensions. Experiment 2 demonstrated that while both dimensions are used in production, either $\Delta$F0 or Turning Point cues alone are sufficient to distinguish the two tones in perception. Isolating intrinsic cues was essential to the subsequent experiments in determining whether these cues were mediated through speaker F0 range, exhibiting a shift, or whether they were used independently of speaker F0 range, showing no effect of speaker condition.

The study then investigated whether changes in speaker identity affect tone perception by presenting tone continua in precursor phrases from two speakers, and observing whether identification shifted as a function of speaker F0 range. Results of Experiments 3a and 3b, which examined perception of both F0 and temporal properties of Tones 2 and 3 in high F0 and low F0 precursor phrases, showed a 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 speaker identity, 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 3c, however, which varied the temporal dimension of Turning Point. The stimuli in Experiment 3c differed in only one aspect from the stimuli in Experiments 3a and 3b: they did not vary in $\Delta$F0. These results suggest that normalization is triggered only when both stimuli and precursors vary in the same acoustic dimension. A context differing along a temporal dimension, such as speaking rate, may trigger normalization processes in perception of the Turning Point stimuli. This hypothesis is investigated in Moore (1995).

If the temporal dimension was not relevant for normalization in the Turning Point stimuli, it is tempting to assume that temporal information may not have contributed to the normalization effect in Experiment 3a, where stimuli varied along both dimensions. However, the larger magnitude of the effect in Experiment 3b as compared to Experiment 3a 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 is degraded. Such differences between effects have been observed in rate normalization work on vowel perception by Gottfried, Miller and Payton (1990), as well as rate effects in the perception of [b] - [w] continua in Shinn, Blumstein and Jongman (1985). Both of these studies show reductions in normalization effects as stimuli more closely resemble natural speech. Thus, it is possible that when listeners are given accompanying temporal information in Experiment 3a, they do not refer to speaker identity as much as when intrinsic tonal cues are restricted, as in Experiment 3b. 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 study, findings of this study support the conclusions of Leather (1983). First of all, the present study observes 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 compare crossover boundaries based on the entire identification function, so that reliable shifts may be observed 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 are assimilatory or contrastive, or whether changes were consistent for all speakers, the present study provides conclusive evidence that shifts in identification are contrastive--in a direction opposite to the precursor F0--and that this shift is consistent enough across subjects in Experiments 3a and 3b to be statistically significant.

Although the context effects shown here are contrastive in direction, these results differ from the direction of shifts 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 argue 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 a category-level perceptual identification, such as for continua whose endpoints do not represent different phonemes, or for non-speech stimuli. For continua whose endpoints 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 as compared to the present study, the opposite shifts in identification reported in the results 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.

Results of this study also support those of Johnson (1990), who found that both intrinsic and extrinsic F0 contributes to vowel perception. Johnson found that when precursor F0 corresponded with two different speakers but intrinsic F0 corresponded with one speaker, a shift in identification occurred, indicating that listeners were mediating intrinsic F0 differences through perceived speaker identity. As the results of Experiments 3a and 3b 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.

## 9. Conclusion

The conclusion of this series of experiments is that perception of tones is a talker-contingent process. Evidence was provided to show that listeners use extrinsic F0 information corresponding to speaker identity in perception of lexical tones, supporting the hypothesis 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 identical, as illustrated in this study, or when the same speech sound exhibits different acoustic characteristics. Further 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, Pisoni, and Martin 1989; Sommers, Nygaard, and Pisoni 1992), blocked conditions (Strange, Verbrugge, Shankweiler, and Edman 1976; Assmann, Nearey and Hogan 1982) or when listeners have increased familiarity with the talkers' voices (Verbrugge, Strange, Shankweiler, and Edman 1976; Nygaard, Sommers, and Pisoni, 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, Strange, Shankweiler, and Edman 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 obtain 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 (1968), 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 avoids the "bootstrap" problem (Nearey, 1989; 2092), which confronts the issue of how listeners are able to establish a representation of speaker identity without precursor acoustic information.

Finally, this investigation raises another issue concerning the degree to which the speaker normalization process is conditioned by language experience. Language differences in tone perception have been reported in terms of the importance placed on intrinsic cues (Gandour and Harshman 1978). In the present study, speaker normalization was demonstrated to occur for native speakers of Mandarin Chinese, a language which uses F0 to distinguish lexical items. It may then be hypothesized that for non-native listeners, or those whose native language does not use F0 to make lexical contrasts, there may be no utilization of speaker F0 range. This hypothesis assumes that normalization is dependent on the presence of phonetic categories for tones. In an examination of this hypothesis, Moore (1995) finds evidence for normalization of speaker F0 range for tone stimuli by English listeners. However, results of that study indicate that the pattern of normalization is different between native and non-native listeners. The difference in patterns derives from the native listeners' greater degree of dependence on contextual F0 range information for less natural stimuli, a pattern not exhibited by the non-native listeners. That is, Mandarin listeners show the greatest effect of speaker F0 range for stimuli varying one acoustic dimension, rather than two (the magnitude of effect differences observed between Experiments 3a and 3b in the present study); English listeners show effects only when stimuli varies two acoustic dimensions. The findings of Moore (1995), therefore, demonstrate that language experience influences normalization.

One of the interesting observations exhibited in the present experiments has been a pattern of speaker normalization which occurs only when both precursors and stimuli vary in the F0 dimension. Normalization for stimuli varying in Turning Point may occur if precursors then vary in a temporal dimension. This hypothesis is also investigated in Moore (1995) for tone stimuli identical to those used in the present study. Findings from Moore (1995) indicate that when precursors vary in a temporal dimension (in this case, speaking rate), listeners normalize for rate by shifting category boundaries for the temporal cue (Turning Point). These results point to the importance of temporal

information in tone identification, information that is overlooked when discussions of tone focus on F0 properties alone.

## 10. References

Assman, 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. *Journal of Phonetics* 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 Press, Berkeley).

Charif, R. A., Hertz, S. R., and Weber, T. J. (1992) *Delta System User's Guide*, (Eloquent Technology, Inc., 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), 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), 119-132.

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

Finney, D. J. (1971). *Probit analysis* (Cambridge University Press, Cambridge).

Fox, R., and Qi, Y. Y. (1990) Context effects in the perception of lexical tone. *Journal of Chinese Linguistics* 18, 261-283.

Fujisaki, H., and Kawashima, T. (1971) A model of the mechanisms for speech perception: Quantitative analysis of category effects in discrimination. In *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* , V.A. Fromkin, ed. (Academic Press, NY), 41-76.

Gandour, J., and Harshman, R. (1978) Cross-language differences in tone perception: a multidimensional scaling investigation. *Language and Speech* 21, 1-33.

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

Henry, F. (1948) Discrimination of the duration of a sound. *Journal of Experimental Psychology* 38, 734-743.

Howie, J. M. (1976) *Acoustical studies of Mandarin vowels and tones* (Cambridge University Press, Cambridge).

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.

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. *Journal of Phonetics* 11, 373-382.

Lehiste, I. (1970) *Suprasegmentals* (MIT Press, Cambridge).

Li, C., and Thompson, S. (1981) *Mandarin Chinese: A functional reference grammar* (University of California Press, Berkeley and Los Angeles, CA).

Li, C.N., and Thompson, S. (1977). The acquisition of tone in Mandarin-speaking children. *UCLA Working Papers in Phonetics* 33, pp. 109-130.

Lin, T. and Wang, W. Y.-S. (1985). Shengdiao ganzhi wenti, [tone perception]. *Zhongguo Yuyan Xuebao* 2, 59-69.

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.

Lyovin, A. (1978). Review of Tone and Intonation in Modern Chinese by M. K. Rumjancev. *Journal of Chinese Linguistics* 6, 120-168.

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

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 normalization in tone perception.* Ph.D. dissertation, Cornell University .

Mullennix, J. W., Pisoni, D. B., 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), pp. 105-111.

Norman, J. 1988. *Chinese* (Cambridge University Press, Cambridge).

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science* 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. Cognit.* 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. *Journal of Speech and Hearing Research* 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). *Journal of Chinese Linguistics* 6, 120-168.

Shen, X. (1990) Tonal coarticulation in Mandarin. *Journal of Phonetics* 18, 281-285.

Shen, X. and Lin, M. (1991) A perceptual study of Mandarin Tones 2 and 3. *Language and Speech* 34, 145-156.

Shen, X., Lin, M., and Yan, J. (1993) F0 turning point as an F0 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. *Perception and Psychophysics* 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. *Japanese Psychological Research* 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]. *Perception and Psychophysics* 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 J. J. Ohala, T M. Nearey, B. L. Derwing, M. M. Hodge, and G. E. Wiebe, eds., *ICSLP 92 Proceedings: 1992 International Conference on Spoken Language Processing* (Priority Printing, Edmonton, Canada) vol. 1, 217-220.

Stott, L. H. (1935) Time-order errors in the discrimination of short tonal durations. *Journal of Experimental Psychology* 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.* University of Wisconsin Ph.D. dissertation.

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.

Zsiga, E. (1994) *Syllt: The Delta Syllable Tool.*, (Eloquent Technology, Ithaca, NY).