Phonetica
Editor: K. Kohler, Kiel

Separatum
Publisher: S. Karger AG, Basel
Printed in Switzerland

D. H. Whalen[a]
Yi Xu[a, b]

[a] Haskins Laboratories, New Haven, Conn., USA
[b] University of Connecticut, Storrs, Conn., USA

Original Paper

Phonetica 1992; 49: 25–47

Information for Mandarin Tones in the Amplitude Contour and in Brief Segments

Abstract

While the tones of Mandarin are conveyed mainly by the F_0 contour, they also differ consistently in duration and in amplitude contour. The contribution of these factors was examined by using signal-correlated noise stimuli, in which natural speech is manipulated so that it has no F_0 or formant structure but retains its original amplitude contour and duration. Tones 2, 3 and 4 were perceptible from just the amplitude contour, even when duration was not also a cue. In two further experiments, the location of the critical information for the tones during the course of the syllable was examined by extracting small segments from each part of the original syllable. Tones 2 and 3 were often confused with each other, and segments which did not have much F_0 change were most often heard as Tone1. There were, though, also cases in which a low, unchanging pitch was heard as Tone 3, indicating a partial effect of register even in Mandarin. F_0 was positively correlated with amplitude, even when both were computed on a pitch period basis. Taken together, the results show that Mandarin tones are realized in more than just the F_0 pattern, that amplitude contours can be used by listeners as cues for tone identification, and that not every portion of the F_0 pattern unambiguously indicates the original tone.

Received:
April 1, 1991
Accepted:
June 17, 1991

D. H. Whalen, PhD
Haskins Laboratories
270 Crown Street
New Haven, CT 06511 (USA)

Introduction

The four tones of Mandarin Chinese are distinguished primarily by the direction of the F_0 contour during the vowel [Howie, 1976]. Many tone languages, for example Cantonese [Gandour, 1984], Gaoba Dong [Shi et al., 1987], and Thai [Abramson, 1975], have some tone contrasts that depend on the height of the pitch contour within the speaker's pitch range. Mandarin, however, does not seem to depend on pitch register. Despite the importance of F_0 movement as a source of tone information, there are other correlates in the phonetic signal. Most prominently, duration is moderately well correlated with tone category [Howie, 1974; Ho, 1976]. Tones 1 and 4 are generally shorter than Tone 2, and Tone 3 is usually the longest. Though less well described, there are also differences in the amplitude contours for syllables with the different tones [Howie, 1976, p. 242; Coster and Kratochvil, 1984]. The present work will examine the distribution of tone information through the syllable in Mandarin, first through the changes in amplitude over time and then through the availability of F_0 information over the course of the syllable.

In the first two experiments, we examine the ability of native speakers of Mandarin to use the information in the amplitude contour to identify the four tones. It has been shown, by Abramson [1972] and Lin [1988], that with the presence of vocal pulses, the role of pitch variation is so prominent that it is very difficult to demonstrate the effect of other phonetic aspects upon the identification of tones. Lin [1988] directly manipulated the amplitude contour but found no effect on tone perception in the presence of the F_0 patterns. Some of the studies on perception of whispered speech [e. g. Wise and Chong, 1957; Kloster Jensen, 1958, Abramson, 1972; Howie, 1976, Test16] found that tonal contrasts were not well preserved in whispered speech, indicating that without vocal phonation during production, it is difficult to convey tonal contrasts. An interesting comparison in Abramson's study was that of synthetic 'whispered' utterances in Thai which were, in fact, normal speech resynthesized through the vocoder with the voice source replaced by a broadband noise source. The original utterance was used as a template for the filters of a synthesizer. With a voiceless source, the original amplitude contour and formant structure remained, but the syllables had no F_0. Subjects were more accurate in tone identification with those stimuli than with naturally whispered utterances. This result suggests that, unlike whispered speech, normal speech in Thai does provide some tonal information through aspects other than pitch variations. Howie's [1976, p. 236] similar study of Mandarin showed relatively poor discrimination of the four tones [39.2%].

While the use of voiceless resynthesis to eliminate F_0 retains the amplitude contour, it also retains the formant structure. The tones might be accompanied by vowel quality changes as well as amplitude changes [Rose, 1988]. The present study makes use of another way of avoiding the presence of F_0, namely, by using signal-correlated noise [Schroeder, 1968]. In this technique, a speech signal is digitized, and then the sign of half of the samples, selected at random, is reversed. The result is a signal with a flat spectrum which nonetheless has exactly the same duration and amplitude contour as the original. Both formant and F_0 information are completely lost. With such stimuli, we can examine just the amplitude contour.

Amplitude seems to vary systematically across the syllable in Mandarin and may thus carry tone information. Conversely, while it is clear that F_0 does vary systematically across the syllable, the tone information in that pattern may be distributed unevenly, with certain parts of the pattern being better indicators of the tone than other parts. We examined this question by extracting brief segments from various points within the syllable. This also allowed a direct comparison of the distribution of F_0 information with the distribution of amplitude, to see if there is any correlation. The perceptual strategies used by Mandarin listeners should be clarified by these results.

Table 1. Duration (in ms) and peak amplitude (dB for a 10-ms window) of the six tokens of each tone on /ba/ used in experiment 1

	Tone 1		Tone 2		Tone 3		Tone 4	
	Dur	Ampl	Dur	Ampl	Dur	Ampl	Dur	Ampl
Token 1	317	49.9	350	49.9	485	47.9	302	52.8
Token 2	335	52.2	357	52.6	414	47.1	286	53.3
Token 3	353	52.3	368	52.8	417	49.3	247	54.2
Token 4	348	53.3	336	51.4	428	48.1	267	54.0
Token 5	293	50.7	324	50.6	443	49.2	287	54.2
Token 6	296	50.9	330	51.8	454	50.1	255	54.6

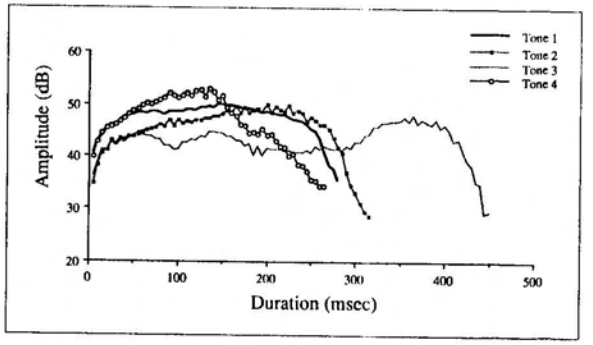

Fig. 1. Typical amplitude contours for the syllable /ba/ with each of the four tones, experiment 1

Experiment 1

In the first study, the ability of subjects to perceive the tone from signal-correlated noise versions of the syllable /ba/ was examined. In this experiment, duration was allowed to vary with tone as it normally does.

Method

Stimuli. A male speaker of Mandarin (the second author) produced 12 repetitions of the syllable /ba/ with each of the four tones. These were read from cards containing the character for the words /bā/ 'eight', /bá/ 'to pull', /bǎ/ 'to hold', and /bà/ 'father'. These productions were input into the Haskins VAX PCM system [Whalen et al., 1990] at a 10-kHz sampling rate without preemphasis of the high frequencies. From these, six tokens were selected for each of the four tones. For Tones 1, 2 and 4, these were simply the first six tokens. For Tone 3, the six were the first six that did not contain creaky voice as part of the realization of the tone. The duration and peak amplitudes (over a 10-ms window) are shown in table 1. Typical amplitude contours are plotted in figure 1. The signal-correlated noise versions were created by randomly inverting about half of the samples of each of these tokens. As described above, the resulting stimuli contained neither F_0 information nor formant information. Such signals sound rather like speech on a telephone line when things have gone disastrously bad.

Procedure. Five repetitions of the 24 signal-corre-

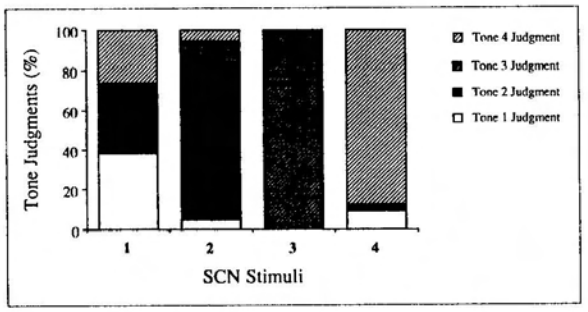

Fig. 2. Identification of the signal-correlated noise (SCN) stimuli of the four tones, experiment 1. Numbers indicate original tones.

lated noise stimuli were randomized and recorded on audiotape for presentation to the subjects. There was an interstimulus interval of 2.5 s, with 6 s after every eight, corresponding to two open lines on the answer sheet. The answer sheets contained rows of four Chinese characters, representing the words/bā/ 'eight', /bá/ 'to pull', /bǎ/ 'to hold', and /bà/ 'father', for each syllable the subjects were to hear. The subjects were instructed to circle the character which they thought (or guessed) was the one intended.

Subjects. The subjects were 10 native speakers of Mandarin (5 male and 5 female), enrolled at the University of Connecticut. All had acquired Beijing Mandarin as their first language.

Results

Figure 2 presents the identification results for these stimuli. As can be seen, Tones 2, 3 and 4 are quite well identified, averaging 87.6% correct. Analysis of variance based on percent correct for each of the four tones shows that these decisions are well above chance [$F(1,9) = 907.16$, $p < 0.001$], as are the decisions that include Tone 1 as well [$F(1,9) = 348.22$, $p < 0.001$]. There were significant differences among all the four tones [$F(3,27) = 54.25$, $p < 0.001$], and also among Tones 2, 3 and 4 [$F(2,18) = 8.44$, $p < 0.01$].

Responses to Tone 1 were fairly evenly divided among Tone 1, 2 and 4 judgments. At 38.5% correct, the Tone 1 judgments are above the chance level of 25% [$F(1,9) = 5.38$, $p < 0.05$]. However, it is apparent from the figure that the Tone 1 contours are decidedly different from those of Tone 3, but barely distinguishable from Tones 2 and 4.

Discussion

When Mandarin listeners are presented with a signal that consists solely of an amplitude contour (which intrinsically has a duration), they are able to identify Tones 2, 3 and 4 fairly well. They can tell that Tone 1 is not Tone 3, but are otherwise uncertain about its identity. These results lead to the opposite conclusion from that of Lin [1988], who found no effect of amplitude contour. As mentioned, he manipulated amplitude on a synthetic syllable with voice source, which of necessity had an F_0 pattern. It seems that when an F_0 pattern is present, it is difficult to ignore it in making a tone judgment. When there is no F_0, as in our stimuli, the amplitude contour can be effective in tone identification.

The stimuli in this experiment had two characteristics which limit the generality of claims that can be made. The first is that the carrier syllable was /ba/. It might be that the overall contour is not so distinct, but rather the character of the contour at the stop release. The next experiment addresses that question by using a carrier syllable that has a glide as its onset (in this case, /yi/).

Table 2. Duration (in ms) and peak amplitude (dB for a 10-ms window) of the five tokens of each tone on /yi/ used in experiment 2.

	Tone 1		Tone 2		Tone 3		Tone 4	
	Dur	Ampl	Dur	Ampl	Dur	Ampl	Dur	Ampl
Token 1	354	58.6	350	53.8	354	52.0	351	55.9
Token 2	401	57.7	403	52.9	417	49.3	416	56.6
Token 3	378	55.7	378	59.5	406	44.6	366	56.4
Token 4	488	58.1	470	57.6	472	50.5	460	60.4
'Typical' token	345	58.2	407	55.1	453	49.1	342	57.4

Token 1 was chosen to have a duration typical of Tone 1 syllables, Token 2 for Tone 2, 3 for 3 and 4 for 4. The 'typical' tokens were the closest to the mean for the tones as produced by our speaker.

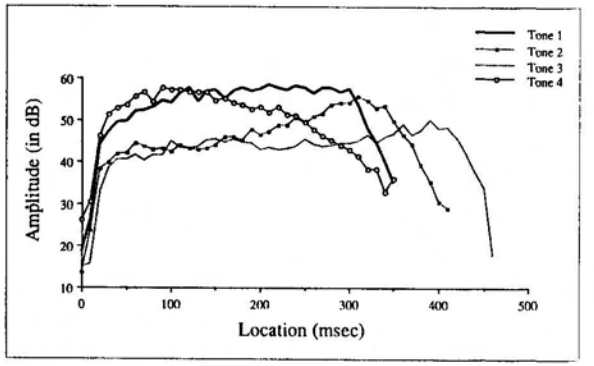

Fig. 3. Typical amplitude contours for the syllable /yi/ with each of the four tones, experiment 2.

The second characteristic of these stimuli is that amplitude contour and duration were correlated. Tone 4 stimuli were short and Tone 3 were long, with the others in between. The next experiment dissociates these two dimensions by choosing tokens at various durations.

Experiment 2

In the second study, signal-correlated noise versions of the syllable /yi/ were used. In this experiment, duration was controlled, so that every tone occurred at every duration used.

Method

Stimuli. A male native speaker (not the same one as in the first experiment) of Beijing Dialect of Mandarin produced 12 repetitions of the syllable /yi/ with each of the four tones. These were read from a list containing the characters for the words /yī/ 'clothing', /yí/ 'suspicion', /yǐ/ 'chair', and /yì/ 'meaning'. Since we knew we would have to exclude any Tone 3 productions which had creaky voice, we obtained twice as many Tone 3 syllables as the others. In order to introduce small variations in the duration of the syllable, we had the speaker produce the syllables in the intervals between the clicks of a metronome. During the recording session, the pace of the metronome was increased by about 10% after each complete reading of the list [see also Maddieson, 1980, p.121]. Although the talker was not explicitly told to alter his rate, he did anyway.

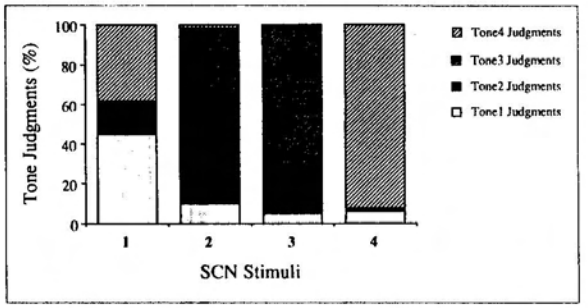

Fig. 4. Identification of the signal-correlated noise (SCN) stimuli arranged by the original tone, experiment 2. Numbers indicate original tones.

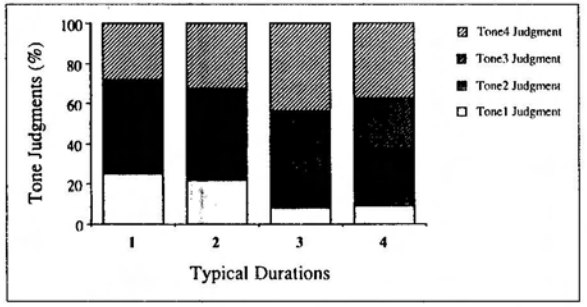

Fig. 5. Identification of the signal-correlated noise (SCN) stimuli arranged by which tone's duration the stimulus represented, experiment 2.

In this fashion, we obtained small changes in rate without changing our instructions. As it turned out, there were enough tokens of each of the tones at various durations in just the first rate to suit our purposes, and those were the only ones used. These productions were digitized as before, and 20 tokens were selected. For each tone, we selected one token which had the typical duration. Then, four additional tokens were selected so that the duration was as similar as possible to the typical duration of each of the tones in order. That is, in addition to the first four tokens, we selected one token of each tone to have Tone 1's typical duration, then one token of each to have Tone 2's typical duration, and so on. The duration and peak amplitudes (over a 10-ms window) are shown in table 2. Typical amplitude contours are plotted in figure 3. The signal-correlated noise version were created as before.

Procedure. Five repetitions of the 20 stimuli were randomized and recorded on audiotape for presentation to the subjects. There was an interstimulus interval of 2.5 s, with 6 s after every eight, corresponding to two open lines on the answer sheet. The answer sheets contained four characters, representing the words /yī/ 'clothing', /yí/ 'suspicion', /yǐ/ 'chair', and /yì/ 'meaning' for each syllable the subjects were to hear. The subjects were instructed to circle the character which they thought (or guessed) was the one intended.

Subjects. The subjects were 12 native speakers of Mandarin (5 male and 7 female), enrolled at the University of Connecticut. All had acquired Beijing Mandarin as their first language. Three of them had participated in experiment 1.

Results

Figure 4 presents the identification results for these stimuli, arranged by original tone. As can be seen, Tones 2, 3 and 4 are identified well above change, at 55.3, 69.5 and 92.3% correct. Responses to Tone 1 were mostly divided between Tone 1 and Tone 4 judgments, with the former at 45.0% and the latter at 38.2%. Thus, unlike the stimuli in experiment 1, the Tone 1 stimuli here were basically ambiguous between Tone 1 and Tone 4. Across the four tones, identification was well above chance [$F(1, 11) = 215.72$, $p < 0.0001$)].

Figure 5 reorganizes the data in accordance with the duration of the production rather than the tone produced. Thus the tokens which had the duration typical of Tone 1 are in the first column, all those that had the typical duration of Tone 2 are in the second, and so on for Tone 3 and Tone 4 (excluding the fifth token, which varied in duration). We examined responses that would be correct for the typical duration. As can be seen, there is no tendency for the 'correct' tone to be perceived based on the duration. The percentages for the correct answer are 24.8, 21.5, 31.3, and 36.9. The only significantly preponderant response is that of Tone 4 to the tokens with Tone 3's duration. This indicates that duration is not being used, since Tone 4 has the shortest typical duration and Tone 3 the longest.

In an analysis of variance of all the judgments organized by original duration rather than original tone, identification was found to be above chance [$F(1, 11) = 15.91$, $p < 0.001$], but only because there are some of the original tones included. We therefore also analyzed the percent correct responses to those syllables whose tone category was not the same as its original tone. For example, responses for Tone 1 would include the productions of Tones 2, 3 and 4 at the typical duration for Tone1, but would exclude those Tone 1's at the Tone 1 duration. Here, there is a significant *below*-chance accuracy of 15.1% [$F(1, 11) = 69.19$, $p < 0.001$]. In short, these judgments were not influenced by duration.

Discussion

Even when the onset of the syllable is a glide and duration is not a cue, Mandarin listeners can identify a syllable's tone with reasonable accuracy solely from the amplitude contour. Unlike the first experiment, Tone 1 is most often heard either correctly or as Tone 4. Even when duration is removed as a cue, the amplitude contour carries a good deal of information about tone.

These results should not be taken as indicating a lack of importance of duration as a cue. If we had neutralized the amplitude contours and varied the duration, we might have seen an effect of duration. Indeed, Yang [1989] reports that both the Tone 1–Tone 2 distinction and the Tone 2–Tone 3 can be made to depend on duration. Blicher et al. [1990] found similar results with the Tone 2 –Tone 3 distinction. The point here is that there is information in the amplitude contour aside from simple duration. However, the study of Coster and Kratochvil [1984] also did not find significantly accurate discrimination of the tones based on duration.

Experiment 3

The distribution of energy across the syllable has been shown to be an effective cue to tone identity in the first two experiments. In the next experiments, we examine the distribution of information in the F_0 pattern at various points in the syllable. This was accomplished by excising brief segments from different locations so that the F_0 pattern at those locations can be examined by itself. To the extent that listeners are able to recover the correct tone from only part of its contour, we will have an indication of the specificity not just of the contour as a whole but of the pieces of it as well. If only one part of the contour can successfully signal a particular tone, then we can conclude that the tone information is potentially restricted to that part of the contour.

Method

Stimuli. The syllables from the first experiment served as the basis for those in the third. One token of

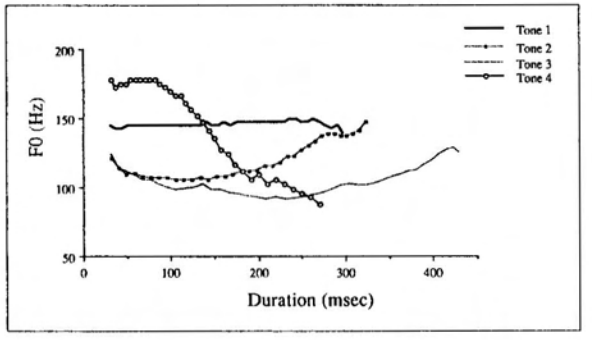

Fig. 6. F_0 patterns of the four syllables that served as the basis for the brief segments used in experiment 3.

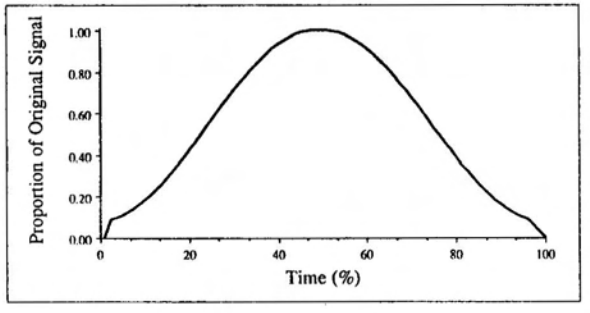

Fig. 7. Shape of the Hamming window used on the extracted segments of experiments 3 and 4.

each of the /ba/ syllables for each of the tones was selected. The F_0 patterns for the four syllables are shown in figure 6. The token we selected had a typical duration for that tone. From each of these syllables, we extracted brief segments of 100, 80, 60 or 40 ms depending on the condition. These were selected from various points in the syllable. The first segment began at the syllable onset, the next began at a point 50 ms into the syllable, the next at 100 ms, and at as many further 50-ms intervals as the syllable would allow. If the remainder of the syllable at any point was less than the duration of the window, that segment was not used. This resulted in 5 or 6 segments for Tone 1, 6 or 7 for Tone 2, 8 or 9 for Tone 3 and 5 or 6 for Tone 4. Since simple extraction from syllables is like using a square window in acoustic analysis (which results in high frequency artifacts), we applied a Hamming window to the resulting sections. By applying this function (displayed in fig. 7) to our segments, we obtained portions of syllables that sounded like extremely short syllables.

Procedure. Stimuli were grouped according to window size. Within each size, five repetitions of the stimuli were randomized and recorded on audiotape for presentation to the subjects. There was an interstimulus interval of 2.5 s, with 6 s after every ten, corresponding to two open lines on the answer sheet. The answer sheets contained four characters, representing the words /bā/ 'eight', /bá/ 'to pull', /bǎ/ 'to hold', and /bà/ 'father', for each syllable the subjects were to hear. The subjects were instructed to circle the character which they thought (or guessed) was the one intended. They were told that the stimuli had been excised from a larger syllable, which they were to identify to the best of their ability. The 100-ms stimuli were tested first, then the 80-, 60- and 40-ms ones.

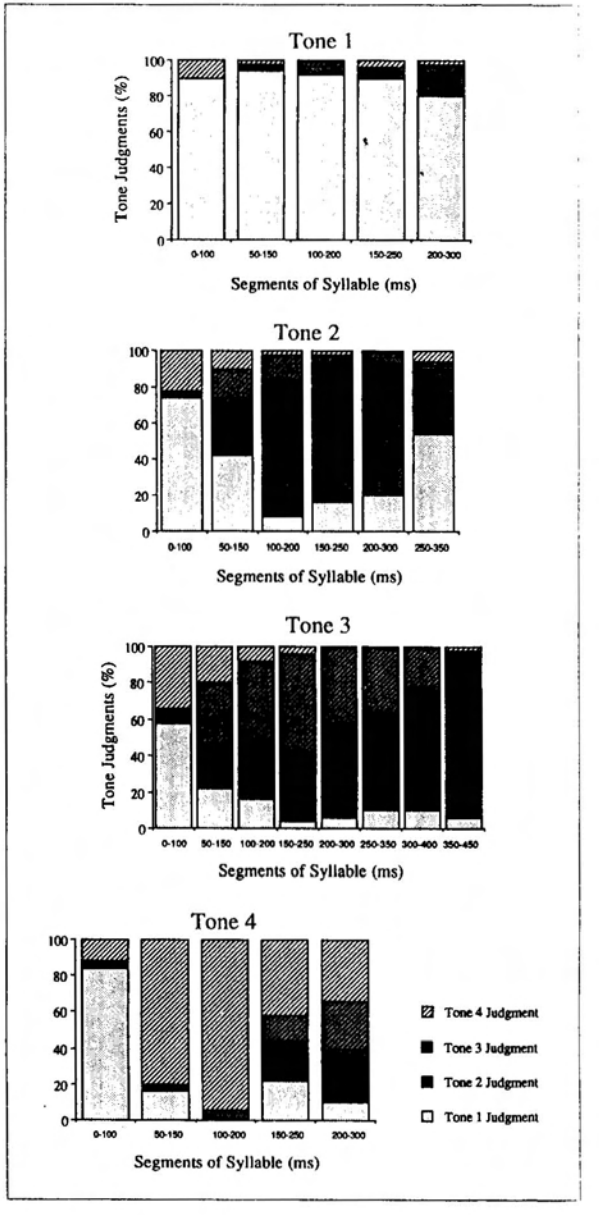

Fig. 8. Identification of the 100-ms segments, experiment 3.

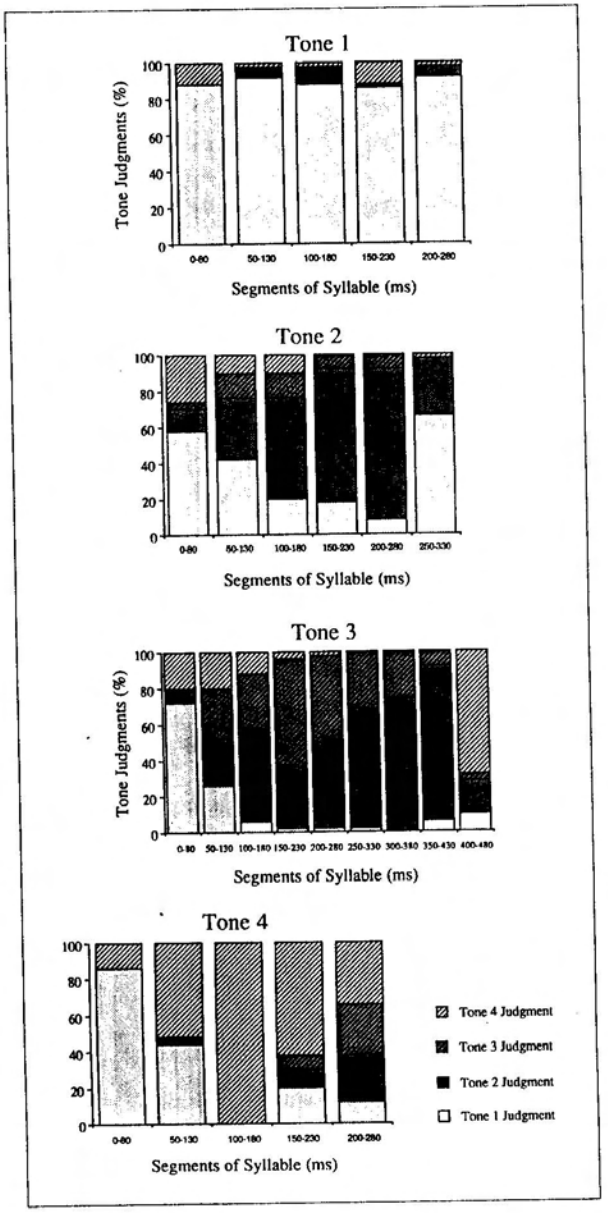

Fig. 9. Identification of the 80-ms segments, experiment 3.

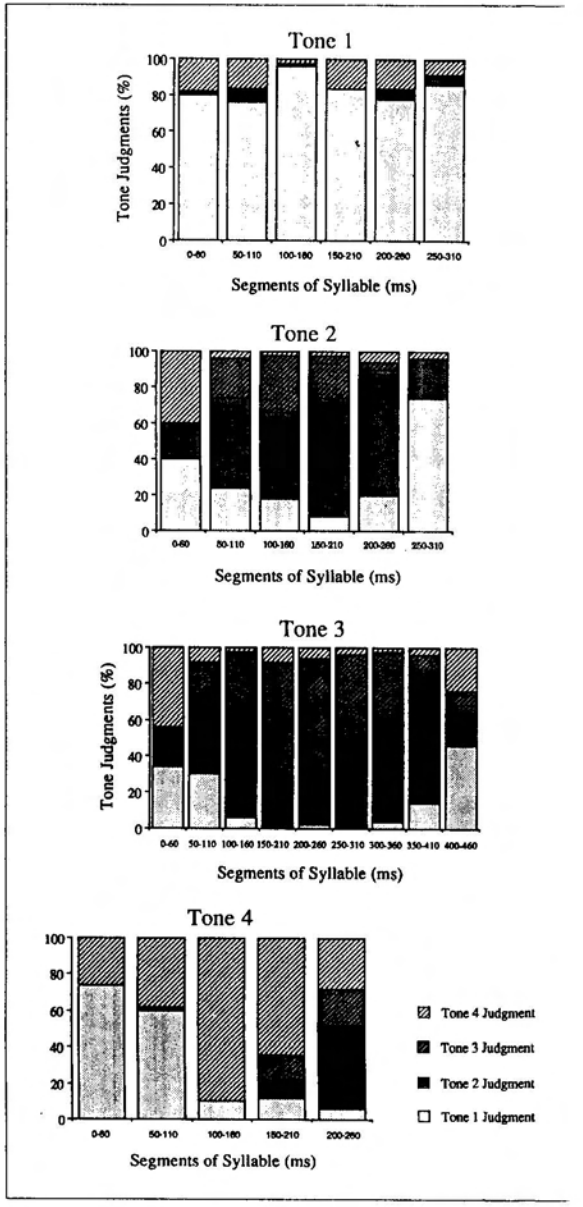

Fig. 10. Identification of the 60-ms segments, experiment 3.

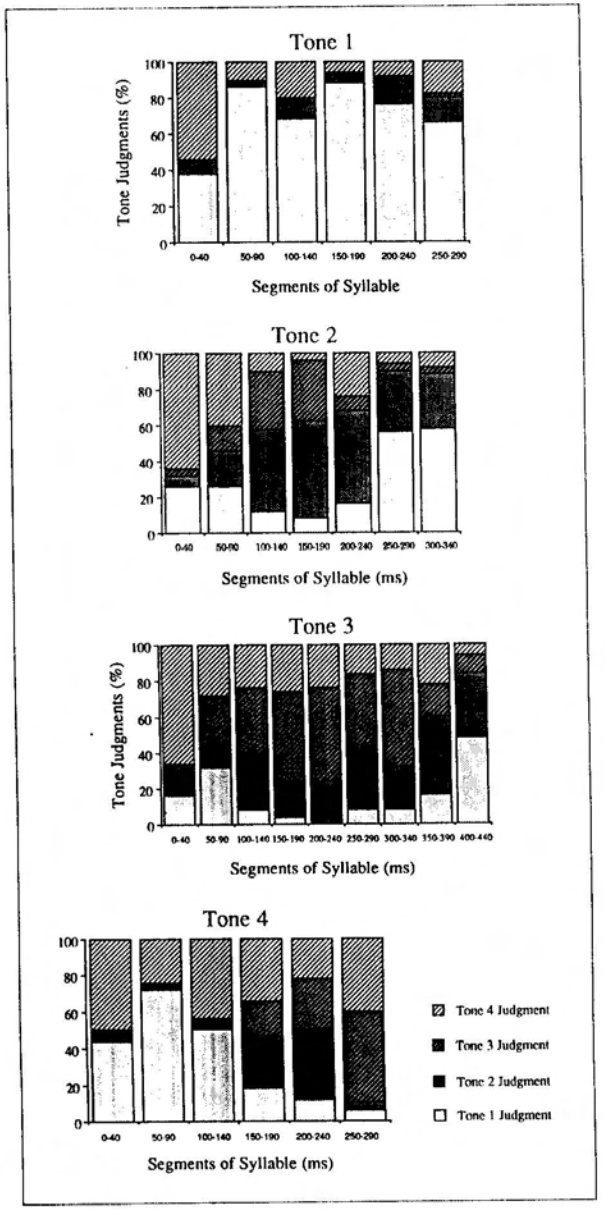

Fig. 11. Identification of the 40-ms segments, experiment 3.

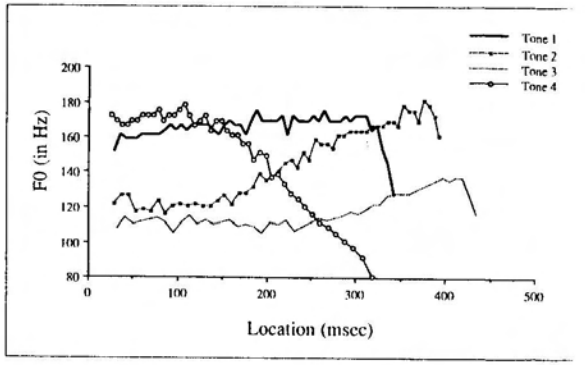

Fig. 12. F_0 patterns of the four syllables that served as the basis for the brief segments used in experiment 4.

Subjects. The subjects were those of experiment 1, and this tape was run after that of experiment 1.

Results

Figure 8 presents the identification results for the stimuli of 100 ms duration, arranged by original tone. It can be seen that Tones 1, 2 and 4 are quite well identified from the segments beginning at 100 ms, over 75% correct in each. Tone 1 was well identified at all locations, while identification of the others was better at the middle to late portions of the syllable. The first segment was most often identified as Tone 1 regardless of the original tone. Tone 3 was often perceived as Tone 2, especially toward the end of the syllable.

Figures 9–11 show the results for the 80-, 60- and 40-ms segments, respectively. The results for the 80- and 60-ms windows are almost identical to those of the 100-ms window. With the shortest window, Tones 1 and 4 become somewhat more confusable with each other, and Tone 4 judgments begin to dominate in the first segment of the syllable. It is possible that with such a short stimulus, the amplitude reduction imposed by the Hamming window was perceived as a drop in F_0, which would sound more like Tone 4.

Discussion

When there is not much movement in the F_0, Tone 1 percepts tend to predominate in these brief, extracted segments. When there is definite movement downward, Tone 4 percepts dominate, while upward movements tend to elicit Tone 2 judgments. Tone 3 judgments are relatively infrequent and tend to be associated with the lowest F_0 values. Thus there is some effect of the level of the F_0 on the tone percept. For the 100-ms window beginning at the 150-ms point in Tone 3, for example, there is almost no F_0 movement, yet Tone 1 percepts are virtually absent. It seems that the relatively low F_0 is responsible, since only Tone 3 tokens without creaky voice were selected for this experiment.

While register is important in some tone systems, it is not often reported to be so in Mandarin. However, Cheng and Sherwood [1982] found that their automatic tone recognition system worked best if it coded F_0 level as well as F_0 movement. Massaro et al. [1985]

showed an effect of F_0 level on Tone 1–Tone 2 distinction. Of course, we can assume that F_0 movement would override F_0 level, just as it does in overriding creaky voice [Gårding et al., 1986]. However, the appearance of a register effect was unexpected.

Experiment 4

The final experiment used the syllable /yi/ to explore the brief segments from different locations so that the F_0 pattern would be relatively free from influences of the release of constriction. In this way, any peculiarity in the patterns from the previous experiment that might be due to the stop release would be eliminated.

Method

Stimuli. The syllables from the second experiment served as the basis for those in the fourth. One token of each of the /yi/ syllables for each of the tones was selected. The F_0 patterns for the four syllables are shown in figure 12. The token we selected had a typical duration for that tone. From each of these syllables, we extracted brief segments of 100, 80, 60 or 40 ms, depending on the condition, using the Hamming window described earlier. These were selected from the same points in the syllable as in experiment 3: The first section was at the beginning, the next began at 50 ms, the next at 100 ms, and at as many further 50-ms intervals as the syllable would allow. If the remainder of the syllable at any point was less than the duration of the window, that segment was not used. For Tones 2 and 3, the 40-ms window at the beginning of the /yi/ was not loud enough to be heard, so these two stimuli were not used. This resulted in 6 or 7 segments for Tone 1, 7 or 8 for Tone 2, 8 or 9 for Tone 3 and 6 or 7 for Tone 4.

Procedure. The stimuli were once again grouped by window size. Five repetitions of the stimuli were randomized and recorded on audiotape for presentation to the subjects. There was an interstimulus interval of 2.5 s, with 6 s after every eight, corresponding to two open lines on the answer sheet. The answer sheets contained four characters, representing the words /yī/ 'clothing', /yí/ 'suspicion', /yǐ/ 'chair', and /yì/ 'meaning' for each syllable the subjects were to hear. The subjects were instructed to circle the character which they thought (or guessed) was the one intended. They were told that the stimuli had been excised from a larger syllable, which they were to identify to the best of their ability. The 100-ms stimuli were tested first, then the 80-, 60- and 40-ms ones.

Subjects. The subjects were those of experiment 2, and these tapes were run after that of experiment 2.

Results

Figure 13 presents the identification results for the stimuli of 100 ms duration, arranged by original tone. It can be seen that the results are quite similar to those of experiment 3, except that the most accurate segments tend to be 50 ms later (those beginning at 150 ms). Another difference from the previous experiment is the accurate identification of Tone 3 from the first segment. For the /yi/ syllable, the Tone 3 token began with a low F_0 and stayed at that level, rather than beginning around the same level as Tone 2 and dropping as in the previous experiment.

Figures 14–16 show the results for the 80-, 60- and 40-ms segments, respectively. The results for the 80- and 60-ms windows are almost identical to those of the 100-ms window. With the shortest window, Tone 3 becomes somewhat more identifiable than it is when longer.

Overall identification of the tones was well above chance [48.0%, $F(1, 11) = 158.71$, $p < 0.001$]. The tones were not equally well identified [$F(1, 11) = 38.20$, $p < 0.001$]. For Tones 1–4, respectively, they were 78.7, 39.3, 30.6, and 43.3% correctly identified. The accuracy differed across durations [$F(3, 33) = 30.38$, $p < 0.001$]. For the 100-ms segments, the rate was 53.5%, the 80, 54.7%, the 60, 57.3 % and the 40, 36.4%. Although this difference seemed to be due primarily to the 40-ms segments, an analysis without that duration still revealed a significant duration effect [$F(2, 22) = 11.42$, $p < 0.001$]. The different tones were differently affected by the changes in duration [$F(9, 99) = 4.61$, $p < 0.001$].

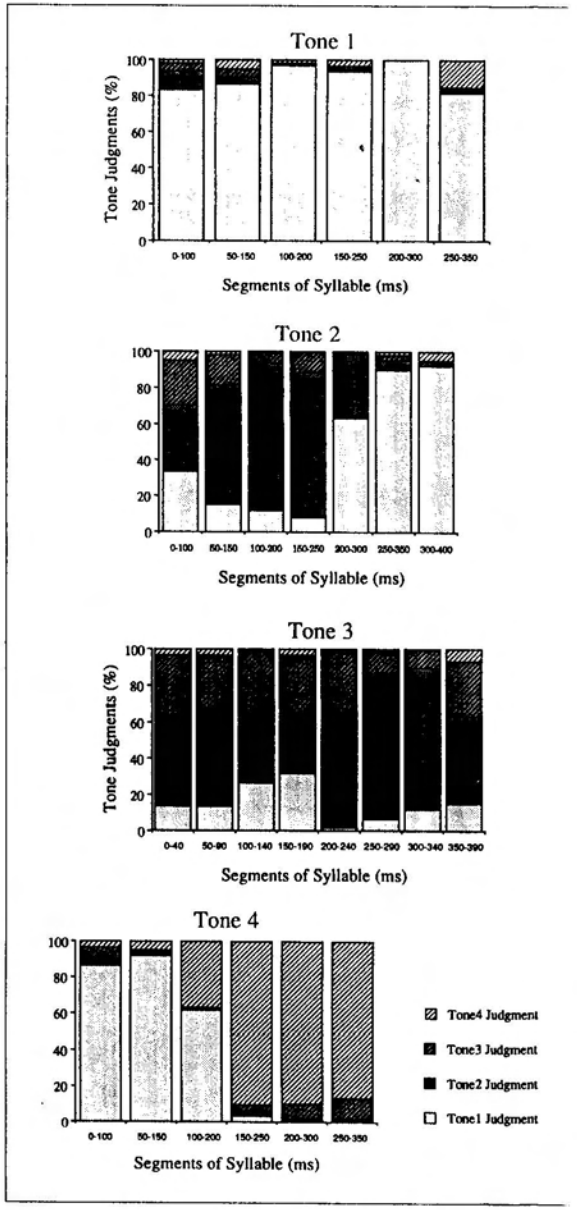

Fig. 13. Identification of the 100-ms segments, experiment 4.

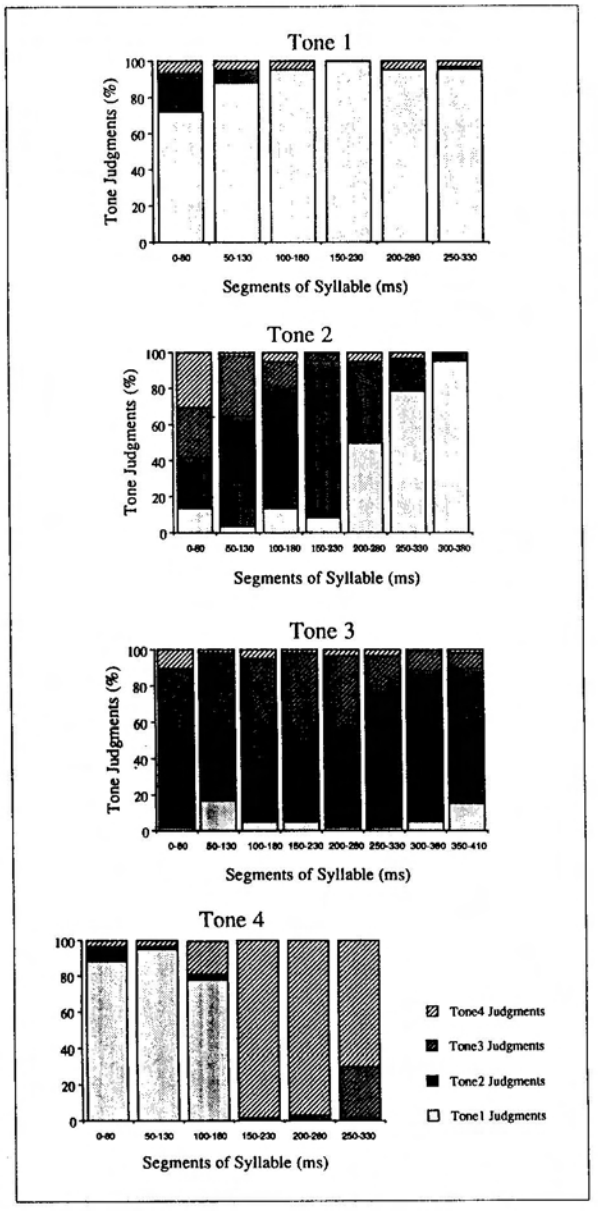

Fig. 14. Identification of the 80-ms segments, experiment 4.

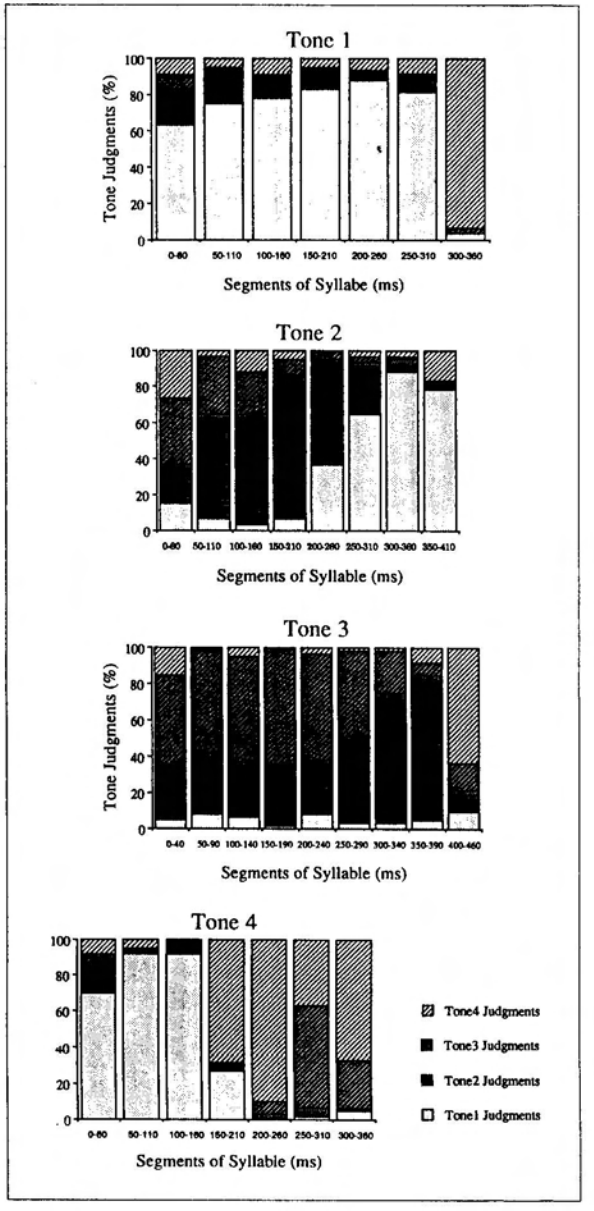

Fig. 15. Identification of the 60-ms segments, experiment 4.

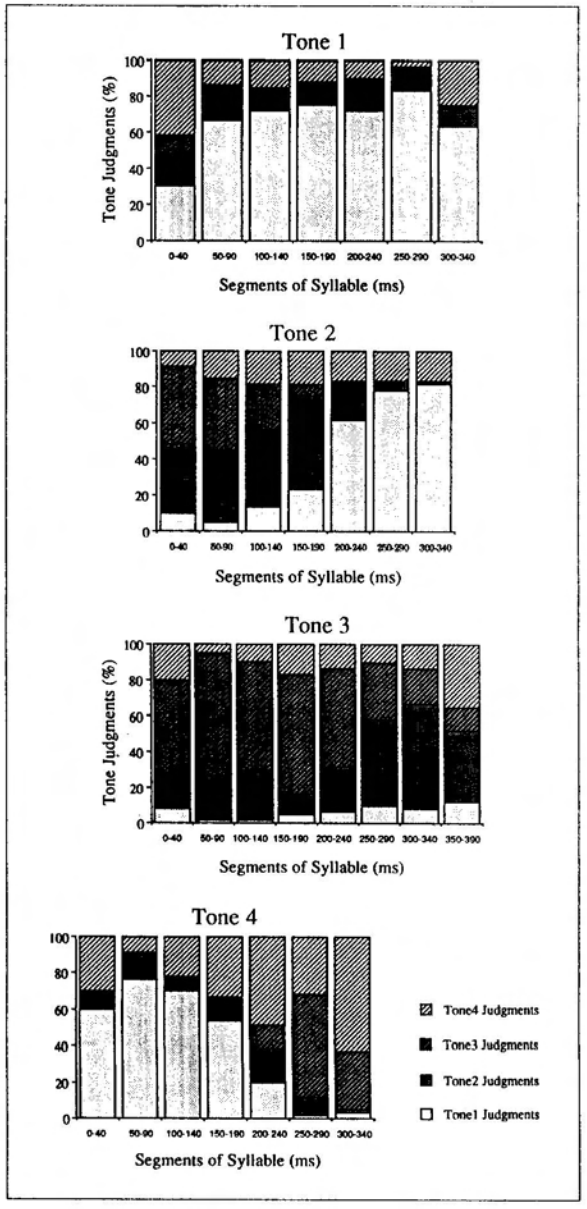

Fig. 16. Identification of the 40-ms segments, experiment 4.

Table 3. Factors obtained in the 'best' subset of independent variables in the regression analysis, experiment 4

Tone 1	Tone 2	Tone 3	Tone 4
Average F_0	Slope	– High	F_0 difference
– Slope	– Average F_0	Slope	– Duration
– F_0 difference	Amplitude	– Duration	– Slope
Duration			– Average F_0
			Amplitude
With the F_0 prediction factor			
Average F_0	Prediction factor	– Average F_0	– Slope
– Slope	Slope	Slope	– Prediction
Duration	– High	– Prediction factor	factor
– F_0 difference	Duration	– Duration	– Average F_0
Prediction factor	– Amplitude		High
			– Duration

A minus sign preceding the variable name indicates that the variable received a negative weighting in the solution.

The change of accuracy over the time course of the syllable was tested in two complementary analyses. Since analysis of variance with repeated measures requires complete cells, it was not possible to analyze all the results at once. Different tones at different durations had different numbers of segments. We therefore compared the first six locations, and also the last six locations. For the 100- and 80-ms segments of Tones 1 and 2, these two sets were identical, while the other combinations had more segments available. In addition, the segment that began at 50 ms for Tones 2 and 3 at the 40-ms duration were considered the first segment, since the segments at location 0 were excluded.

The tests for the tone and duration factors replicated those of the previous test and will not be further elaborated. The percent correct differed across the syllable both in the beginning six segments [$F(5, 55) = 49.49$, $p < 0.001$] and in the final six segments [$F(5, 55) = 13.88$, $p < 0.001$]. All the interactions with tone, duration and the two together were significant in both analyses (we will spare the reader the numbers).

To explore more systematically the contributions of F_0 level and F_0 change on tone perception, we analyzed the results from experiment 4 with an all-possible-subsets regression [Frane, 1988]. This technique takes any number of independent variables, and tries all combinations of them to see which set does the best job of accounting for the variance. The independent measures we used were the highest F_0 in the segment, the average F_0 across the segment, the difference between the highest and lowest F_0 (regardless of position within the segment), the slope of the F_0 (which would be positive for rising F_0 and negative for falling), the duration of the segment, and the peak amplitude of the segment. Another analysis used all those factors plus a prediction factor, based on F_0 movement, that varied for each tone (these will be described in the next paragraph). The dependent measure was the percentage of times a particular tone was chosen by our 12 subjects for that segment. Thus there were 113 segments, and each had a dependent measure for each of the four tones, each of which needed its own analysis. That is, all the Tone 1 responses were analyzed, then all the Tone 2, etc. In this way, the 'best' subset of predictors can help establish the important factors in these stimuli for each of the tone percepts.

The prediction factors that we made for the tones were based on our interpretation of the importance of F_0 movement, and the fact that simple correlations with the slope would be less successful for Tone 1, where a lack of slope is positively predicted. The lack of movement seemed to be important for Tone 1, so we established a prediction that would make for a small number when there was a lot of tone movement (up or down) and large when there was little F_0 movement. Specifically, we used the formula:

$$\text{Prediction} = 50 - (\text{slope}^2),$$

where 50 was chosen because it was a reasonably small number which was nonetheless larger than most of the squares of the slopes. (If the number happened to be less than zero, it was set to zero.) For Tones 2 and 3, the prediction was that rising slopes would be good indications for that tone, while flat and falling would be equally bad. So if the slope was 0 or less, the prediction was set to 0. Otherwise, it was the square of the slope, so that bigger slopes would give an even bigger effect. For Tone 4, the prediction is the opposite of that for Tones 2 and 3. Thus if the slope was positive or 0, the prediction was 0. Otherwise, it was the square of the slope, so that bigger (negative) slopes would give an even bigger (positive) effect.

Before reporting the subset analyses, we should point out that the prediction factors were significantly correlated with the actual judments. For Tones 1, 2 and 4, the correlations were 0.30, 0.58, and 0.38, respectively, all significant at the 0.001 level. For Tone 3, there was a small, nonsignificant negative correlation (–0.14), due to the small number of Tone 3 judgments overall and the lack of a difference (based on F_0 movement) between Tone 2 and Tone 3. However, of our seven independent measures, the ones to obtain the highest separate correlations with the tone judgments were average F_0 for Tone 1 (0.90), the prediction for Tone 2 (0.48), the high F_0 for Tone 3 (–0.85), and slope for Tone 4 (–0.67).

The 'best' subsets are presented in table 3. The first set of best subsets are those calculated with only the six physical measurements. The second set includes our prediction factor as well. As can be seen, Tones 2 and 4 have an F_0 change metric as their primary contributor, while an F_0 level measure contributes the most to Tones 1 and 3. Each subset has both F_0 level and F_0 change measures, with small contributions of amplitude and duration. So, although changes in F_0 guided many of the tone judgments, there was a surprising contribution of absolute F_0 values to tone perception with these stimuli.

Discussion

Experiment 4 confirms the results of experiment 3: When there is little movement in F_0, Tone1 predominates, while movements down or up yield Tone 4 or 2 judgments, respectively. There is some effect of the level of the F_0 on the tone percept. For the /yi/ stimuli, there were even more cases where the F_0 was fairly flat for 60 ms or more, and yet not all of those cases were identified as Tone 1. It seems that the relatively low F_0 can serve to distinguish Tones 2 and 3 from Tones 1 and 4.

Blicher et al. [1990] hypothesize that Tone 3 perception depends on the initial nonrising portion of the F_0 contour. If we include a register component, this is supported by the present results. That is, a flat onset, if it is low enough in F_0, seems to be perceived more as Tone 3 than as Tone 1 (the flat tone). Since these syllable segments were short, it could well be that the short allophonic version of Tone 3 was perceived, not the full citation form [see Shih, 1988, for an acoustic analysis of the short versions]. In that version, Tone 3 is low and falls slightly, so the contour may not have been as important.

Comparison of Amplitude Contours and F_0 Information

We have seen that the distribution of energy in the amplitude contour is itself a cue to tone identity. Similarly, changes in F_0 specify tone identity more unambiguously than do level F_0s. Segments with little F_0 movement sound like Tone 1. Are the two contours we have examined, amplitude and F_0, correlated with each other and thus with tone identifiability? To examine this question, we correlated the amplitude at the center of the time segments in experiment 4 with the total percent correct, with the slope of the F_0 for that segment, and with the absolute F_0 at the midpoint of the window.

Visual inspection of the amplitude and F_0 plots suggested that large changes in F_0 might be correlated with a large amplitude, but this turns out not to be the case. Using the absolute value of the slope as a measure of the degree of change in F_0, we found a nonsignificant *negative* correlation of amplitude with slope ($r = -0.11$, n.s.). Primarily, the Tone 4 measures were responsible for what trend there was, since there is a significant negative correlation for Tone 4 alone ($r = -0.93$, $p < 0.01$).

There is, however, a striking correlation between amplitude and absolute F_0 ($r = 0.91$, $p < 0.001$). Each of the tones (other than Tone 1, which has little F_0 variation) shows this correlation [Tone 2 (0.94, $p < 0.01$) and Tone 4 (0.94, $p < 0.01$) strongly, and Tone 3 (0.65, $p < 0.10$) marginally]. More strikingly, even though the absolute levels of amplitude differ among the syllables with different tones, the dB per Hertz of F_0 ratio is quite stable, as seen in the high correlation above. Even though the amplitude measurements were taken for the window as a whole, there is still some possibility that the character of the pitch periods of human speech can artifactually introduce this difference. Lower F_0s will have fewer pitch periods in a fixed duration than will higher F_0s. As a final check on this, we correlated F_0 and amplitude on a pitch period basis. Although the correlation is lower (0.71), it is still quite significant ($p < 0.001$, $n = 203$). The pitch periods for each tone also yielded significant correlations. Similar reports have previously appeared for English [Ladefoged, 1967, p. 32; Hirano et al., 1969] as well as Mandarin [Chuang and Hiki, 1975]. Coster and Kratochvil [1984] do not perform a correlation directly, but using their six measures for the four tones, we obtain a correlation of 0.29 (not significant for the 24 observations). Since their measurements were made on all the tones in an extended conversation, many segmental and prosodic factors would be expected to weaken the correlation.

This correlation between amplitude and F_0 raises an interesting possibility for interpreting the results of the first two experiments. Although the most straightforward interpretation of the results is that subjects recognize a consistent correlate of the tones (i.e., the amplitude contour), it is also possible that the subjects interpreted amplitude changes *as* F_0 changes. Thus the large decline in amplitude at the end of Tone 4 might actually give rise to a faint percept of a falling F_0 [Rossi, 1978]. While intriguing, this suggestion is difficult to test directly. Any manipulation of the amplitude contour would, of necessity, remove that contour from the realm of the typical. In addition, these F_0 percepts, if they exist, must be tenuous, since Tone 1 is not well recognized from the amplitude contour itself. This is the case in spite of the fact that the contour is quite flat, which corresponds quite well to the flat F_0. Despite that correspondence, Tone 1 was hard to identify, lending more support to the notion that the amplitude contour itself had to be dynamic in order to be interpreted as F_0 changes.

General Discussion

Although the main source of information for Mandarin tones is clearly the F_0 contour, we found significant information present in the amplitude contour. This was the case even when we controlled duration by constructing a situation where each of the tones occurred at each of the other tone's typical duration. As for the distribution of tone information in the F_0 contour, those portions with a large change in F_0 are likely to be recognized correctly as Tone 4 (for falling F_0) or Tone 2 (or 3, somewhat interchangeably, for rising F_0). However, despite a large tendency for portions of the contour that lacked F_0 movement to be perceived as Tone1, there was also a contribution of the actual F_0 (which could be localized within the speaker's range over the course of the experiment), so that low F_0 was heard as Tone 2 or 3. Thus there is some evidence for a register component in the primarily dynamic tones of Mandarin.

Previous studies of the contribution of amplitude to tone perception have been somewhat more negative in their results. Tones tend not to be well recovered in whispered speech [e. g. Abramson, 1972], though there are other differences besides the presence or absence of F_0 in whispered speech. Thus whisperers may not reproduce the amplitude contour of voiced utterances at all. In another vein, Lin [1988] found that direct manipulation of the amplitude contour in synthetic syllables did not affect tone judgment, but his stimuli included an F_0 contour. In the presence of F_0, the amplitude is indeed easy to ignore. In the absence of F_0, as in our stimuli, the amplitude information for Tones 2, 3 and 4 is fairly distinct. There is the possibility that these amplitude changes give rise to weak F_0 percepts directly, since there is a strong, direct correlation in our stimuli between amplitude and F_0. However, the flat amplitude contour of Tone 1 should then have given rise to an appropriately flat F_0 percept. In fact, Tone 1 was hard to recover from the amplitude contour. It is certainly possible, though, that the subjects were simply using the typicality of the contour as information. Results for English speakers, for example, have shown that subjects can make use of signal-correlated noise stimuli to aid visual word recognition processing [Frost et al., 1986]. Further experiments would be needed to dissociate these possibilities.

The evidence for differences in the location of tone information across the syllable is less consistent. Howie [1974] found that the F_0 variation before the rhyme portion (consisting of the nuclear vowel and the final glide or nasal) of the syllable does not seem to contribute to the contour. His results do not indicate which portion of the contour itself is most important. More detailed comparisons are rare and somewhat contradictory. Gårding et al. [1986, p. 292] explain the distinction between Tone 3 and Tone 4 as a difference between an early downshift in F_0 for Tone 4, but a late flattening of a downshift for Tone 3. (Note that their Tone 3s were in sentences and therefore of the short variety, that is, not citation form as used here. We do not know how many of our subjects were reporting 'short' Tone 3s when they did report Tone 3.) This would lead to the prediction that the information for Tone 3 should be later in the syllable than that for Tone 4. The results presented here give some confirmation of that prediction. Blicher et al. [1990, p. 46], on the other hand, hypothesize that Tone 3 depends primarily on 'a detectable initial period of nonrising F_0'. (Their Tone 3s, like the present ones, were full citation forms.) In light of the present results, that statement must at least be narrowed to a 'low' F_0.

Not only do listeners normalize for a speaker's typical range of F_0 [Leather, 1983], they can be expected to interpret the range even more narrowly in an experimental situation like ours where the range is more limited still. Thus the fact that we obtain register effects for our experiments may have less than general application to Mandarin. It does raise the possibility, however, that very short syllables, such as we might expect in running speech, would show more of a register tendency.

Acknowledgements

This research was supported by NICHD Grant HD 01994 to Haskins Laboratories. Experiments 1 and 3 were presented at the 63rd Annual Meeting of the Linguistic Society of America, New Orleans, La., December, 1988. Experiments 2 and 4 were presented at the 119th Meeting of the Acoustical Society of America, State College, Pa., May, 1990. Thanks go to Keith Johnson for spurring us on to the amplitude analysis. Arthur S. Abramson provided helpful comments.

References

Abramson, A. S.: Tonal experiments with whispered Thai; in Valdman, Papers in linguistics and phonetics to the memory of Pierre Delattre, pp. 31–44 (Mouton, The Hague 1972).

Abramson, A. S.: The tones of Central Thai: some perceptual experiments; in Harris, Chamberlain, Studies in Thai linguistics in honor of William J. Gedney (Central Institute of English Language, Bangkok 1975).

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

Cheng, C. C.; Sherwood, B.: Technical aspects of computer-assisted instruction in Chinese. Tsing Hua J. Chinese Stud., New Ser. *14:* 35–49 (1982).

Chuang, C. K.; Hiki, S.: Acoustical features of the four tones in monosyllabic utterances of Standard Chinese. J. acoust. Soc. Japan *31:* 369–380 (1975).

Coster, D. C.; Kratochvil, P.: Tone and stress discrimination in normal Beijing dialect speech; in Hong, New papers on Chinese language use, pp. 119–132 (Australian National University, Canberra 1984).

Frane, J.: All possible subsets regression; in Dixon, BMDP statistical software manual, pp. 919–939 (University of California Press, Berkeley 1988).

Frost, R.; Repp, B. H.; Katz, L.: Can speech perception be influenced by simultaneous presentation of print? J. Memory Lang. *27:* 741–755 (1986).

Gandour, J.: Tone dissimilarity judgments by Chinese listeners. J. Chinese Ling. *12:* 235–260. (1984).

Gårding, E.; Kratochvil, P.; Svantesson, J.-O.; Zhang, J.: Tone 4 and Tone 3 discrimination in modern Standard Chinese. Lang. Speech *29:* 281–293 (1986).

Hirano, M.; Ohala, J.; Vennard, W.: The function of laryngeal muscles in regulating fundamental frequency and intensity in phonation. J. Speech Hearing Res. *12:* 616–628.

Ho, A. T.: Mandarin tones in relation to sentence intonation and grammatical structure. J. Chinese Ling. *4:* 1–13 (1976).

Howie, J. M.: On the domain of tone in Mandarin. Phonetica *30:* 129–148 (1974).

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

Kloster Jensen, M.: Recognition of word tones in whispered speech. Word *14:* 187–196 (1958).

Ladefoged, P.: Three areas of experimental phonetics (Oxford University Press, London 1967).

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

Lin, H.-B.; Repp, B. H.: Cues to the perception of Taiwanese tones. Lang. Speech *32:* 25–44 (1989).

Lin, M.-C.: Putonghua shengdiao de shengxue texing he zhijue zhengzhao. (The acoustic characteristics and perceptual cues of tones in Standard Chinese.) Zhongguo Yuwen *204:* 182–193 (1988).

Maddieson, I.: Palato-alveolar affricates in several languages. UCLA Working Papers in Phonetics *51:* 120–126.

Massaro, D. W.; Cohen, M. M.; Tseng, C.: The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. J. Chinese Ling. *13:* 267–289 (1985).

Rose, P. J.: On the non-equivalence of fundamental frequency and pitch in tonal description. Pacific Ling. *C-104:* 55–82 (1988).

Rossi, M.: Interactions of intensity glides and frequency glissandos. Lang. Speech *21:* 384–396.

Schroeder, M. R.: Reference signal for signal quality studies. J. acoust. Soc. Am. *44:* 1735–1736 (1968).

Shi, F.; Shi, L.; Liao, R.: An experimental analysis of the five level tones of the Gaoba Dong language. J. Chinese Ling. *15:* 335–361 (1987).

Shih, C.-L.: Tone and intonation in Mandarin. Working Papers Cornell Phonet. Lab. *3:* 83–109 (1988).

Whalen, D. H.; Wiley, E. R.; Rubin, P. E.; Cooper, F. S.: The Haskins Laboratories' pulse code modulation (PCM) system. Behavior Res. Methods Instruments Computers *22:* 550–559 (1990).

Wise, C. M.; Chong, L. P.-H.: Intelligibility of whispering in a tone language. J. Speech Hear. Disorders *22:* 335–338 (1957).

Yang, Y.-F.: The vowels and the perception of Chinese tones. Acta psychol. sinica *34:* 29–34 (1989).