The effect of precursor duration on tone normalization in Cantonese

ERAMELA Elaine

A dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, May, 10, 2002.

Abstract

In Cantonese, listeners' tonal judgment has been shown to depend on the acoustic information (speakers' pitch range) carried by the context preceding or following the target word. However, no empirical attention has ever been given to the amount of acoustic information needed for Cantonese tone normalization. In the investigation reported here, a Cantonese semantically neutral preceding context was cut into six different lengths and varied in three pitch levels. Twenty normal subjects aged 21 - 25 were recruited to perform a lexical identification task. The results provided robust evidence for the presence of tone normalization in Cantonese. A period of 450 ms was found to be a crucial duration for successful tonal judgment when the fundamental frequency of the context was raised or lowered by one musical semitone.

Introduction

In Cantonese, syllables are constituted phonologically not only of consonants and vowels but also of tones. Tone has its principal acoustic correlate of fundamental frequency (F0) (Leather, 1983; Bauer & Benedict, 1997). In each syllable, tone possesses lexical significance in which it is an indispensable phonetic dimension for differentiating words with the same segmental features (Fok, 1974). To be more precise, we call the tones in Cantonese *lexical tones*.

According to Vance (1976), Gandour (1981), Matthews & Yip (1994) and Bauer & Benedict (1997), the Cantonese tone system comprises of six contrastive tones, namely High Level (HL), High Rising (HR), Mid Level (ML), Low Falling (LF), Low Rising (LR) and Low Level (LL). The tone letters assigned according to their pitch values are 55, 35, 33, 21, 23, 22 respectively according to Chao (1947) and Kao (1971). For example, a word /fu_{35}/ (tiger) is described as having a HR tone, because the speaker has to change his F0 from a relatively mid frequency to a higher frequency, with two tone letters 35 to delineate the starting and ending F0 levels. The word /fu_{22}/ (father) is a LL tone, as the tone contour of this word begins at a relatively low F0 and remains fairly level towards its endpoint.

The primary, acoustic correlate of tone is fundamental frequency, which is the result of the number of the openings and closings of the vocal cords in one second (Bauer & Benedict, 1997). However, since different people have different tension of their vocal cords, size of the vocal tracts and vocal cords, and volume of airflow from the lungs, their F0 in the vocal cords vibration vary tremendously which affect their voice quality and pitch. A speaker with a

naturally high pitch produces a particular tone with a higher F0 than a speaker with a low pitch voice. For example, a woman with average F0 of 200 Hz produces a HL tone at 250 Hz but a man with average F0 of 100 Hz may produce the same HL tone at 150 Hz. In this example, people are referring to the same word, but with two different F0, 250 Hz and 150 Hz. Theoretically, Cantonese-speaking people should have problems in distinguishing one word from another due to the diversity of individual absolute range of F0. However, the inter-speaker variation does not seem to cause listeners any difficulties in tone identification. On the one hand, all speakers of the same tone language still preserve the relative distances that separate tones and the general shapes of the tone contours, which makes it possible for the speakers to maintain the consistency of their utterances for the listeners (Bauer & Benedict, 1997). On the other hand, the movement and the height of the tone contour are relative and not absolute features (Fok, 1974; Bauer & Benedict, 1997). Listeners will automatically adjust their perception of the tone relating to the average pitch of the speaker in order to identify the different tones produced within the utterances (Bauer & Benedict, 1997; Wong 1998; So, 2001; Wong, 2001). The process that listeners adjust their perception according to speaker-specific acoustic information is called speaker normalization (Leather, 1983; Moore & Jongman, 1997; Nusbaum & Magnuson, 1997).

Speaker normalization compensates for differences among speakers regarding the different extrinsic information the preceding context carries, like speaking rate, stress pattern and average F0 of the speakers (Nusbaum & Magnuson, 1997). There have been many studies providing experimental evidence of the presence of speaker normalization in English speech perception on

vowel quality, place of articulation, manner of articulation etc (Eimas, Tartter, Miller & Keuthen, 1978; Johnson, 1990; Ladefoged & Broadbent, 1957; Miller & Eimas, 1977; Summerfield & Haggard, 1975). Would the same context effect occur when listeners are processing suprasegmental aspects of speech such as tone? To look into more specifically whether tone normalization is present, we have to review past literature concerning tone perception on tonal languages, like Mandarin and Cantonese.

In the studies of Mandarin tone perception (Fox & Qi, 1990; Leather, 1983; Lin & Wang, 1985; Moore & Jongman, 1997), the role of extrinsic acoustic information carried by the context on tone normalization was examined. Experiments were carried out which aimed at investigating how listeners perceived a particular tone in isolation or within context. For example, Lin & Wang (1985) presented listeners with pairs of Mandarin Chinese tones in which the tones in the first syllables were held constant whereas the onset F0 of the second syllables varied. Listeners were asked to identify the tones in the first syllable in each pair. Although it was claimed that different F0 onset of the second syllable affected listeners' perception of the tone in the first syllable, there was no statistical analysis provided following the experiment to show how robust the results were. By using a similar anchoring paradigm as Lin & Wang (1985), Fox & Qi (1990) also presented listeners with pairs of Mandarin Chinese tones. The first syllable in the pair was either Tone 1 or Tone 2 whereas the F0 onset of the second syllable varied. Listeners were asked to identify the tones in the second syllable in each pair. Results only showed borderline significance of contextual effect on tone normalization. The pattern of responses in the experiment did not

occur reliably for all subjects nor for a particular stimulus.

When it comes to Cantonese, there were three recent studies about Cantonese tone perception (Wong, 1998; So, 2001; Wong, 2001). Wong (1998) compared the tonal judgment of the listeners when the target words were presented in isolation under the mixed-speaker condition and the one-speaker condition. The target words were 師 (/si$_{55}$/, means 'teacher', in HL tone), 試 (/si$_{33}$/, means 'try', in ML tone) and 是 (/si$_{22}$/, means 'yes', in LL tone). He found that tones in isolation were more difficult to be identified under the mixed-speaker condition due to the unknown speaker identity. Wong (1998) also examined the effect of the presence of a preceding context 下一個字係 (/ha$_{22}$ jɐt$_{5}$ kɔ$_{33}$ tsi$_{22}$ hai$_{22}$/, means 'The next word is') on tonal judgment of the target word 試 (/si$_{33}$/, means 'try', in ML tone). He showed that by raising the F0 of the preceding context by two semitones and lowering it by three semitones (we call such shifting as 'perfect shift'), listeners perceived the target word as in LL tone and HL tone respectively. His assumption was made based on the fact that the ML tone is approximately three semitones lower than the HL tone and two semitones higher than the LL tone (Chao, 1947). Listeners gained significantly higher accuracy in identifying the tone of the target word with the presence of a preceding context, when compared with the two isolation conditions. This suggested that listeners used F0 in the context as a cue for speaker identity and the acoustic information carried in the preceding context assisted listeners' tonal judgment by allowing tone normalization to take place. By also investigating the effect of English context, and with different size of F0 shift of the Cantonese and English context, Wong (1998) suggested a *pitch*

range assessment model for tone perception, which claims that listeners will assess and refer to the pitch range of the speaker before making tonal judgment. They will perceive a tone as high when the tone is at the higher end of the speaker's speaking pitch range while they will perceive a tone as low when the tone is at the lower end of the speaking pitch range of the speaker.

To extend Wong's (1998) study, Wong (2001) continued to examine the role of context on listeners' tonal judgment. She embedded the target word 意 (/ji$_{33}$/, means 'meaning'), which was in ML, in the middle of the context 我會讀 __ 俾你聽 (/ŋɔ$_{23}$ wui$_{23}$ tuk$_{2}$ __ pei$_{35}$ nei$_{23}$ t^{h}ɛŋ$_{55}$/, means 'I'll read __ for you to hear'). She intended to find if tone identification depended on the F0 of the context by shifting the F0 of the context perfectly (upward by two semitones and downward by three semitones). Similar to the result obtained by Wong (1998), a high accuracy of expected responses was obtained (more than 90%) for the perfect shift of F0 of the context, which showed strong evidence for the context effect on tone perception. Wong (2001) also examined the effect of size of pitch shift on tone normalization by raising and lowering the F0 of the context with small-shifting (0.5 semitone), medium-shifting (1 semitone) and large-shifting (2 semitones). Ceiling effect was obtained for the large-shifting condition, and the accuracy decreased in line with the decreasing in size of the shift. Wong (2001) also investigated whether listeners relied on pitch range or the average pitch of the context for tone normalization by using a dynamic context (with a full range of pitch variation) and a monotone context (with no pitch variation). She argued that to identify tone, an estimate of the speaker's average pitch was rather sufficient, but there may also be a role for the speaker's pitch range.

In investigating the contextual effect on tone perception, So (2001) also shifted the F0 of the context with the same size and direction as in Wong's (2001) study. The results were consistent with that of Wong's (2001). So (2001) also looked into the effect of position of the context on tone perception. She embedded the target word 意 (/yi$_{33}$/, means 'meaning', in ML tone) in three contexts: a 'preceding context' 下一個字會係 (/ha$_{22}$ jɐt$_{5}$ kɔ$_{33}$ tsi$_{22}$ wui$_{23}$ hai$_{22}$/, means 'The next word is'), a 'following context' 係一個中文字 (/hai$_{22}$ jat$_{5}$ kɔ$_{33}$ tsɔŋ$_{55}$ mɐn$_{21}$ tsi$_{22}$/, means 'is a Chinese word') and a 'both context' 我會讀 __ 俾你聽 (/ŋɔ$_{23}$ wui$_{23}$ tuk$_{2}$ __ pei$_{35}$ nei$_{23}$ tʰɛŋ$_{55}$/, means 'I'll read __ for you to hear'). She found that there was no significant difference of the listeners' responses among the three types of context. She agreed with the *concept-driven, top-down active process model* suggested by Nusbaum and Schwab (1986) and the *pitch range assessment model* suggested by Wong (1998) that listeners will wait for enough acoustic information about the pitch range of the context before making any decision on the tone of the target word. Therefore, any acoustic information in the context will be useful for tonal judgment regardless of the position they appear.

Wong (1998), Wong (2001) and So (2001) have investigated the role of context on Cantonese tone normalization. However, no experimental work has ever been done to examine the amount of acoustic information needed for such normalization. The present study extends work from the examination of the role of context to the duration of context on tone normalization. By presenting different lengths of a preceding context to listeners and examining if listeners perform differently with respect to different lengths of context, the answer for the research

question 'How much of the speaker's voice do we have to hear to make relatively accurate tonal judgment?' can be addressed. It is predicted that the longer the length of context is, the higher percentages of expected responses will be obtained because more acoustic information is available for tone normalization.

Method

Subjects

Twenty native Cantonese speakers aged from 21 to 25 with normal hearing were recruited. Nine were students in the Department of Speech and Hearing Sciences in the University of Hong Kong, with some training in linguistics. Ten of them came from other departments in the same university, and one was a secondary school teacher. Of these 11 subjects from outside the Department of Speech and Hearing Sciences, one had studied linguistics for three years, and reported that he had some knowledge about tone perception. Therefore, ten of the subjects were classified as having linguistic knowledge (nine Speech and Hearing Sciences students and one linguist) but not the others.

Stimuli

The stimulus used in the experiment consisted of a precursor phrase (context) and a target word. The precursor phrase was semantically neutral and composed of four syllables 阿志要個 (/a_{33} tsi_{33} jiu_{33} $kɔ_{33}$/, means 'Ah Chi needs a'), each with the ML tone. The target word was 意 (/ji_{33}/, means 'meaning'), also in ML tone. According to Wong's (2001) study, there was a significant difference between the listeners' performance towards a monotone context and a

dynamic context with pitch shifted for 0.5 and 1 semitone. The precursor phrase in this study was designed to eliminate any other factors related to the pitch range of the context to influence the tonal judgment by the listeners. By using a sentence with only one tone, the way the listeners' judgment interacts with the contour shapes of the context is not an issue.

The stimulus was produced by a native Cantonese male whose age was 22 and with no known speech or hearing impairment. He was a student at the University of Hong Kong. He volunteered to participate in the study. The precursor phrase 阿志要個 (/a_{33} tsi_{33} jiu_{33} $kɔ_{33}$/, means 'Ah Chi needs a') and the target word 意 (/ji_{33}/, means 'meaning') were recorded as a whole phrase. The speaker was asked to read out the whole phrase with the target word once after several times of practice. The recording was done in a quiet room. The phrase was recorded via a Macintosh external microphone directly to disk using Praat 3.9.27 (Boersma & Weenink, 2001) via the built-in sound card of an Apple Macintosh G4 at a sampling rate of 44.1k Hz.

The phrase was modified using the Praat 3.9.27 software (Boersma & Weenink, 2001). The length of precursor phrase was determined to last for 1043 ms. The F0 of the target word remained constant in all conditions while the F0 of the precursor phrase was shifted to create three levels of F0: unshifted, raised (one semitone upward) and lowered (one semitone downward). The size was chosen to be one semitone because, according to the studies of So (2001) and Wong (2001), shifting the context upward or downward with two semitones would cause a ceiling effect. It would be too easy for the listeners to identify the tone of the target word according to the pitch level of the context. In this case, we would not be able to conclude

whether the judgment of the listeners was based on the pitch shifted or the length of context. Therefore, a smaller pitch shift with one semitone was used in order to allow ample room for performance differences to be measured across different length of context After the F0 shifting, each of the three sentences was then cut so that six different lengths of the precursor phrase were created. The six different lengths were 125 ms, 250 ms, 450 ms, 650 ms, 850 ms and 1043 ms. This also roughly corresponded to zero, one, two, three and four syllables preceding. Thus, there were totally 18 conditions (six length of precursor x three levels of F0 shifting).

Procedures

The experiment was conducted in an IAC single-walled sound booth. An instruction sheet was provided and the experiment was started with the consent from listeners. The complete phrase was not included in the instruction, to avoid any disturbance to the effect of the length of the context to the tonal judgment of the target word by the listeners.

There were six parts in the experiment, which were different in the length of context of the stimuli. Part 1 consisted of trials with the shortest length of context (125ms) while part 6 consisted of the longest (1043ms). The other lengths were presented in order, e.g. 250ms, then 450ms etc. In each part of the experiment, there were 11 blocks of trials. Block 1 was considered a practice and the result was not analyzed although the subjects were not informed of that. Block 2-11 were for experimental purpose. Within each block, there existed three trials with three different pitch levels (raised, lowered and unshifted) arranged in a randomized order. To sum up, there were 33 trials in each part, giving a total of 198 trials in the whole experiment (33 trials x 6

parts).

The six parts of the experiment were presented according to the ascending order of the length of context (i.e. Part 1 was presented first while Part 6 was the last). The reason to present the shortest part first was that listeners might estimate and memorize the pitch level from the context when they were doing the longer parts, and the memory of the pitch level of the context might be recalled when they performed parts with shorter context. In such circumstances, listeners would not actually be using only the context provided for tonal judgment for parts with shorter context, but also the memorized pitch level of the context presented earlier.

In each trial, listeners listened to the stimuli (the context and the target word) through the headphone and were asked to identify the target word using three choices 醫 (/ji$_{55}$/, means 'doctor', in HL tone), 意 (/ji$_{33}$/, means 'meaning', in ML tone) and 二 (/ji$_{22}$/, means 'two', in LL tone) shown on the screen. The three choices were only different in the tone level (high, mid and low). Each participant needed to give tonal judgment for 198 trials in the experiment individually and the session lasted for about 15 minutes.

Results

The responses from the listeners were recorded and scored according to whether they matched with the expected responses in the particular situations. It was expected that if the F0 of the context was lowered by one semitone, listeners would respond with a HL tone because the unshifted F0 of the target word would be higher than the F0 of the context sentence. In contrast, if the context was raised by one semitone, listeners would respond with a LL tone because the

unshifted F0 of the target word would be lower than the F0 of the context sentence. For the unshifted context, listeners would respond with a ML tone. The percentages of the expected responses of each condition were calculated.

The twenty listeners were having different education background in which their linguistic knowledge varied. Ten of them had linguistic knowledge while ten of them had not. To ensure that their linguistic knowledge did not affect their performance, one of the independent variable in this study was the between-groups factor with the two levels: with and without linguistic knowledge. There were two other independent variables, including the direction of pitch shift (raised, lowered and unshifted) and the length of the context (125 ms, 250 ms, 450 ms, 650 ms, 850 ms and 1043 ms). A three-way analysis of variance (ANOVA) with a between groups factor of linguistic experience and repeated measures on the factors of direction of pitch shift and the length of context was employed. Since the dependent measure of percent correct was bounded by zero and one, all the data was first processed through the arcsine transformation (Kirk, 1995). The arcsine transformation was used to allow the data to more closely fit the assumptions of ANOVAs. The results will be reported separately according to the effect of i) listeners' linguistic knowledge, ii) direction of pitch shift, and iii) length of context.

Results showed that there was no main effect of the listeners' linguistic knowledge on their tonal judgment in this study ($F(1,18)=.29$, $p>.05$). There was no interaction effect between listeners' linguistic knowledge and the direction of pitch shift ($F(2,36)=1.03$, $p>.05$). There was neither any interaction effect between listeners' linguistic knowledge and the length of context

($F(5,90)=1.38$, $p>.05$). Therefore, this factor can be ignored in subsequent analyses.

As shown in Figure 1, listeners gave expected responses with 73% in the raised context, 87% for the unshifted context and 44% for the lowered context. The main effect of direction of pitch shift was significant ($F(2,36)=15.48$, $p<.01$), suggesting that listeners made their tonal judgment based on the pitch level of the context. Post-hoc analysis by using Tukey HSD tests showed that the percentage of expected responses in the lowered context was significantly smaller than that in the raised context ($p<.05$). The same percentage was also significantly smaller that that in the unshifted context ($p<.05$). In contrast, the difference of the percentages of expected responses between the raised context and the unshifted context was not significant ($p>.05$). It implies that although the F0 of the context was raised and lowered with the same size (one semitone), the two directions of the shift did not affect listeners' tonal judgment to the same degree.

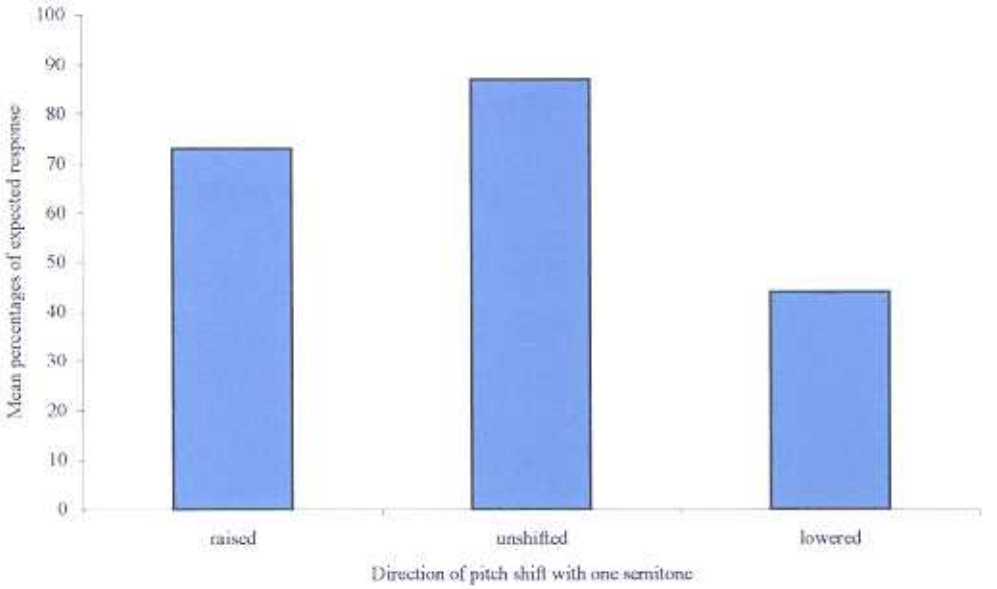

Figure 1. Mean Percentages of Expected Responses by Direction of Pitch Shift

As shown in Figure 2, listeners gave expected responses with the mean percentages of 85% when the length of context was 1043 ms and 83% when it was 850 ms. A mean percentage of 74% was obtained both when the length of context was 650 ms and 450 ms. When the length of context was 250 ms, a mean percentage of 56% was obtained while when the length of context was 125 ms, a mean percentage of 36% was obtained. There were statistically significant differences among the effects of lengths of the context in eliciting expected responses across different directions of pitch shift ($F(5, 90)= 50.75$, $p<.01$). Tukey HSD test was done and the results revealed that the percentage of expected responses in the 125 ms context (36%) was significantly lower than that in all longer context durations ($p<.05$ for all cases). The percentage of expected responses at 250 ms context duration (56%) was also significantly lower than that in all higher durations ($p<.05$ for all cases). However, the percentages of expected responses for the 450 ms context (74%) and 650 ms (74%) were not significantly different ($p>.05$), and they also did not differ significantly from those at 850 ms (83%) and 1043 ms (85%) ($p>.05$). The percentages of expected response at 850 ms (83%) and 1043 ms (85%) also did not show significant difference between each other ($p>.05$).

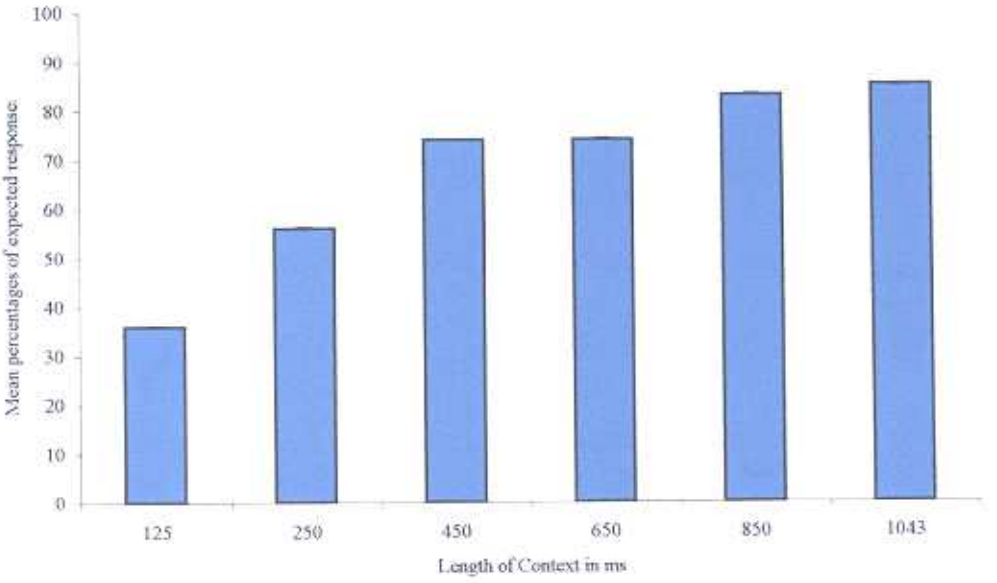

Figure 2. Mean Percentages of Expected Responses by Different Length of Context

An interaction effect was found between the two independent variables of the direction of pitch shift and the length of context ($F(10,180)=7.89$, $p<.01$). This implies that the effect of the length of context depended on the direction of pitch shift. In order to locate the source which contributed to such pattern, post-hoc comparisons were performed by using the Tukey HSD test.

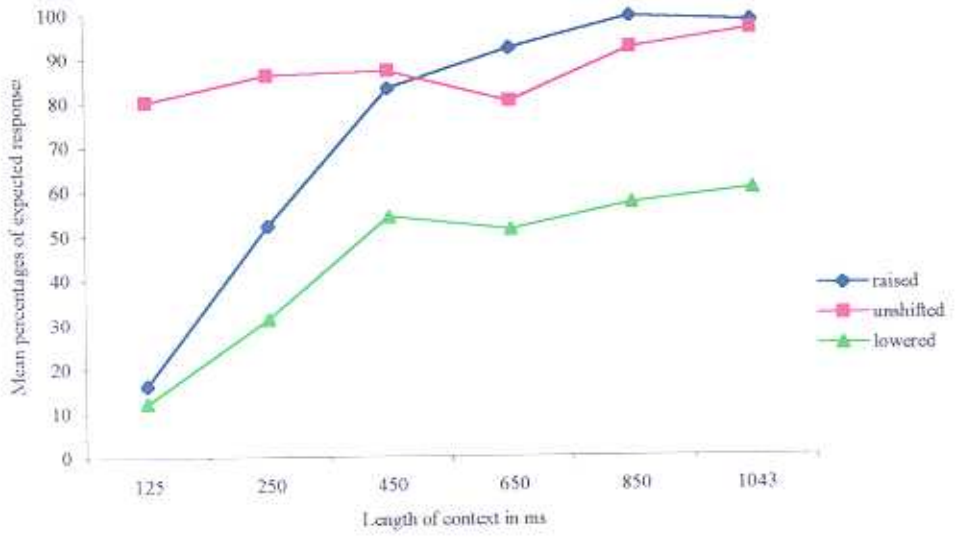

Figure 3. Mean Percentages of Expected Responses in Different Length of Context With Different Direction of Pitch Shift

The curve in Figure 3 representing the unshifted context (red line with filled squares) shows a flat line. Results revealed that there was no significant difference of the listeners' performance when different lengths of context were presented under the unshifted context ($p>.05$ for all cases). A mean percentage of 87% (ranging from 80% to 96% of percentages of expected responses) was obtained in this context for the six lengths. Under the lowered condition, the curve (green line with filled triangles) shows an upgoing trend from 125 ms to 450 ms and became level for contexts longer than 450 ms. However, only the percentage of expected responses of the context 450 ms was significantly higher than that of the context 125 ms ($p<.01$). There was no significant difference found among the remaining adjacent pairs. For the raised context, a steeper upgoing trend can be observed from 125 ms to 450 ms in the curve (blue line with filled rhombus). Only the percentages of expected responses of context length 250 ms was significantly higher than that

of the length 125 ms, and that of 450 ms was significantly higher than that of the length 250 ms (both with $p<.01$). For all the rest of the adjacent pairs within this context, no significant difference was observed ($p>.05$ for all cases).

The results of the Tukey HSD test revealed that the percentages of expected responses of the raised context were significantly higher than that of the lowered context from 450 ms onwards ($p<.05$). When the lengths of context were 125 ms and 250 ms, the percentages of expected responses between the raised and lowered conditions were not significantly different ($p>.05$).

Discussion

The aim of the present study was to investigate the amount of contextual cues needed for Cantonese tone normalization. Previous studies about Cantonese tone perception (Wong, 1998; So, 2001; Wong, 2001) suggested that tone normalization is operating in assisting listeners in tone perception and identification. The experiment conducted in this study supported the notion and proved that the difference in length of context did affect listeners' tonal judgment. There existed a critical length of context in which tone normalization can function optimally. The interaction of the effect of length of context, the effect of the direction of pitch shift and the listeners' tonal judgment will be discussed further.

Evidence of tone normalization

The results of this study coincide with previous researches about tone perception that listeners' tonal judgment is made based on the F0 of the context preceding the target word. In the

experiment in this study, the context was shifted to create three conditions, namely raised context (one semitone upward), unshifted context, and lowered context (one semitone downward). Listeners identified the target word as LL tone with 73% of the time in the raised context while they identified the target word as HL tone with 44% of the time in the lowered context; and they identified the target word as ML tone with 87% of the time in the unshifted context. The result is consistent with Wong's (1998), Wong's (2001) and So's (2001) studies that listeners perceptually adjust their expectation according to the acoustic information provided by the context. Hence, it is proved that tone perception is a talker-contingent process.

How does the F0 of the context affect tone normalization?

Listeners may have an expectation on the location of the tone of the target word after the preceding context was presented (So, 2001). Raising the F0 of the preceding context may result in raising the listeners' expected pitch range (the highest and lowest ends) of the six Cantonese tones. Similarly, lowering the F0 of the context may result in lowering their expected pitch range of the tones. There should be an acceptable range of expected location of the F0 of the target word according to the size of F0 shifting of the context (So, 2001). When the F0 shifting in the context was large, a larger shift of range of the expected location of the F0 of the target word would result. In the studies of Wong (1998) and Wong (2001), they raised the F0 of the context by two semitones. Listeners perceived the 'actual' ML tone as a LL tone with more than 90% of the time, suggesting that listeners might have shifted their expectation of the F0 of the target word upward for about two semitones. Similarly, the F0 of the context was lowered by three

semitones and listeners perceived the 'actual' ML tone as a HL tone with similarly high percentage, suggesting that listeners might have shifted their expectation of the pitch of the target word downward for about three semitones.

The F0 of the context was only shifted upward and downward by one semitone in this study. Listeners may have shifted their expected location of the pitch of the target word with a smaller degree. When the context was raised by one semitone, the one-semitone upward shift of expected location of the pitch of the target word may not be large enough for listeners to perceive the 'actual' ML tone as a LL tone, according to the fact that the distance between a ML tone and a LL tone is two semitones (Chao, 1947). Likewise, when the context was lowered by one semitone, the one-semitone downward shift of expected location of the pitch of the target word may not be large enough for listeners to perceive the 'actual' ML tone as a HL tone, according to the fact that the distance between a ML tone and a HL tone is three semitones (Chao, 1947). Nevertheless, the same target word 意 (/ji$_{33}$/, means 'meaning', in ML tone) could still be perceived as different tones depending on the relative pitch level of the context at least for some incidences (73% of LL responses under the raised context and 44% of HL responses under the lowered context). This suggests that tone normalization is not an all-or-none matter. Otherwise, listeners would either have perceived the target word as in ML tone (the original tone) for most of the trials because the shifting was not enough, or have judged the target word according to the shift with an expected response for most of the trials because the small amount of shift is already enough for 'perfect' normalization.

Listeners in this study performed significantly better under the raised context condition (a LL tone was expected) than under the lowered context condition (a HL tone was expected). Note that under the raised context condition, the distance between the actual F0 of the target word and the expected F0 location of the LL tone is about one semitone. However, under the lowered context condition, the distance between the actual F0 of the target word and the expected F0 location of the HL tone is about two semitones. Therefore, it was more difficult for listeners to perceive the ML tone as a HL tone in the lowered context (with a gap for about two-semitone size) than to perceive the ML tone as a LL tone in the raised context (with a gap for about one-semitone size) when the size of F0 shifting of the contexts were the same.

The critical length of context for tone normalization

Recall that when we looked at the main effect of the length of context on listeners' tonal judgment, only the mean percentages of expected response between 125 ms/250 ms (36%/56%) and 250 ms/450 ms (56%/74%) were significantly different. There was a plateau of listeners' performance in tonal judgment when the duration of context was longer than 450 ms.

To identify tones, listeners have to refer to their knowledge of the speakers' pitch range and adjust their perception according to the speaker-specific information (Moore & Jongman, 1997). The speakers' pitch range can be determined from the information of the context. The longer the context is, the more acoustic information will be available. 250 ms length could elicit significantly more expected responses than the 125 ms length due to the fact that the increase in the length of context provides more acoustic information for listeners to refer to for further

calibration and normalization to take place. The same principle can explain why the 450 ms length has elicited more expected responses than the 250 ms length.

When the context presented was 650 ms, 850 ms and 1043 ms, listeners' performance did not differ significantly from that of 450 ms. In this case, although more acoustic information was available for listeners' reference, the extra information was not as helpful as we expected. The result is apparently showing that 450 ms of the context was relatively enough for listeners' tone normalization to operate for tonal judgment. Listeners responded with an expected response for 74% of the time. This response pattern suggested that any additional context beyond the 'minimum duration' of 450 ms did not assist in listeners' tone normalization anymore and was left unused.

Figure 3 shows that there is a change in slope for the raised and lowered context at 450 ms. At 125 ms and 250 ms, the difference in the percentage of expected response between the raised context and the lowered context was not significant. It implies that when the context is shorter than 450 ms, the effect of the length of context dominates listeners' performance. Although we expect a performance difference between the raised context condition and the lowered context condition, due to the different size of gap between the expected F0 location of the target word from listeners and the actual F0 of the target word, it does not apply to the situation in which the context is not entirely sufficient for tone normalization. The insufficient acoustic information did not allow listeners to be affected by the difference between the pitch levels. When the contexts were 450 ms, 650 ms, 850 ms and 1043 ms, the percentages of expected responses under the

raised condition were significantly larger than that under the lowered condition in each time interval. Tukey HSD analysis showed that for both contexts, all points lying on the same curve did not differ statistically from 450 ms onwards. This suggests that the effect of the direction of pitch shift dominates listeners' tonal judgment for contexts of 450 ms and above rather than the effect of the length of context. The sufficient acoustic information provided allows the difference between the pitch levels of the context presented to affect listeners' performance.

What happened when the length of context was at 125 ms and 250 ms?

When the length of context was only 125 ms, the mean percentage that listeners responded with a ML tone when the context was raised (for which a LL tone was expected) was 74% whereas when the context was lowered (for which a HL tone was expected) was 79%. When the context was unshifted, 80% of expected response (a ML tone) was obtained. The ML tone appeared to be a 'default' response in the 125ms condition. Although it may be controversial that under the unshifted 125ms context, the context may be enough for listeners' tone normalization as a high percentage (80%) of ML tone response was obtained. However, when the same pattern of responses of the ML tone with similarly high percentages was obtained under the raised and lowered context, such behavior may be better explained by the fact that the context with 125 ms is rather short and listeners may have just perceived the target words as if they were in their isolation form. Therefore, with no doubt, a ML tone was perceived most of the time in whatever context.

When the length of context was 250 ms, the mean percentage that listeners responded with

an expected responses for the raised, unshifted and lowered condition were 52%, 86% and 31% respectively. They responded with a ML tone for the raised context with 42% of the time and for the lowered context with 63%. Since the percentages of the expected responses under the raised and the lowered conditions increase significantly when comparing with the condition with 125 ms length, the increase in the length of context is believed to be helpful in tone normalization of the listeners. However, since 250 ms is not entirely sufficient in assisting listeners to have stable and accurate tonal judgment, although it is already much better than 125 ms, listeners were still less certain about the pitch of the context. This contributes to the phenomenon that listeners responded with a ML tone for most of the trials in which they failed to identify the expected tone. For example, in the raised condition, listeners heard a LL (expected) tone 52% of the time, but a ML tone 42% of the time (and HL only 6% of the time). This means that virtually all of their 'mis-' identifications corresponded to the actual tone of the stimulus as originally produced.

Conclusion

This study has shown that tones do not rely on their absolute acoustic F0 values to gain their identity. Rather, they contrast with the F0 of the context to attain a relative identity (Moore & Jongman, 1997). Furthermore, there exists a critical duration of context, 450 ms, which allows relatively robust normalization to take place. Any additional information beyond this critical amount may have been left unused.

However, the 450 ms duration is only obtained under the circumstance that the F0 of the context was shifted upward or downward by one semitone. Future research may shed light on

examining if different size of the F0 shifting yields different results.

In addition, the six lengths employed in the study were 125 ms, 250 ms, 450 ms, 650 ms, 850 ms and 1043 ms. Further experiment can specifically locate the more precise length of context needed for tone normalization by using smaller intervals around the 450ms length.

Lastly, the 450 ms duration is about 2-syllable length in this study. It would also be of interest to investigate whether it is the particular length of time (450 ms) providing the sufficient acoustic information for tone normalization, or it is the 2-syllable length. By using fast or slow rate of speech in the context may address the answer for this query.

Acknowledgements

I would like to express my wholehearted thankfulness to my supervisors, Dr. Alexander L. Francis and Dr. Valter Ciocca, for their kind guidance and inspiring comments. Thanks also extend to my friends who voluntarily participated in this study.

I would like to show my sincere appreciation to my classmates, Mr. Wilson Leung, Miss Cecilia Au, Miss Monique Law and Miss Kathy Woo for their continuous support and constructive suggestions throughout the year.

References

Bauer, R.S., & Benedict P.K. (1997). *Modern Cantonese Phonology.* Berlin: Mouton de Gruyter.

Boersma, P. & Weenink, D. (2001). *Praat Program.* Retrieved May 8, 2002, from University of Amsterdam, Institute of Phonetics Sciences Web site: http://www.fon.hum.uva.nl/praat/ manual/Praat_program.html

Chao, Y.R. (1947). *Cantonese Primer.* Cambridge, Mass: Harvard University Press.

Eimas, P.D., Tartter, V.C, Miller, J.L., & Keuthen, N.J. (1978). Asymmetric dependencies in processing phonetic features. *Perception and Psychophysics, 23,* 12-20.

Fok, Y.Y. (1974). *A Perceptual Study of Tones in Cantonese.* Center of Asian studies, the University of HK.

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

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

Johnson, K. (1990). The role of perceived identity in F0 normalization of vowels. *The Journal of Acoustical Society of America, 88,* 642-654.

Kao, D.L. (1971). *Structure of the Syllable in Cantonese.* The Hague: Mouton.

Kirk, R.E. (1995). *Experimental Design.* Pacific Grove, CA: Brooks/Cole.

Ladefoged, P. & Broadbent, D.E. (1957). Information conveyed by vowels. *The Journal of the*

Acoustical Society of America, 29, 98-104.

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

Lin, T. and Wang, W. Y. (1985). Sheungdiao ganzhi wenti [Tone perception]. *Zhongguo Yuyan Xuebao [Chinese Linguistics Journal], 2,* 59-69.

Nusbaum, H. & Magnuson, J. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J.W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp.109-130). San Diego: Academic Press.

Matthew, S. & Yip, V. (1994). *Cantonese: A Comprehensive Grammar.* London: Routledge.

Miller, J.L. & Eimas, P.D. (1977). Studies on the perception of place and manner of articulation. *The Journal of the Acoustical Society of America, 61,* 835-845.

Moore, C.B. & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America, 102,* 1864-1877.

Nusbaum, H.C. & Schwab, E.C. (1986). The role of attention and active processing in speech perception. In E.C. Schwab & H.C. Nusbaum (Eds.), *Pattern Recognition by Humans and Machines: Vol. 1. Speech Perception* (pp.113-157). San Diego, CA.: Academic Press.

So, K.W. (2001). *The Effect of Preceding and Following Contexts on Cantonese Tone Perception.* Unpublished Bachelor's degree dissertation, The University of Hong Kong.

Sommers, M.S. (1998). Spoken word recognition in individuals with dementia of the Alzheimer's type: Changes in talker normalization and lexical discrimination.

Psychology and Aging, 13, 631-646.

Summerfield, A.Q. & Haggard, M.P. (1975). Vocal tract normalisation as demonstrated by reaction time. In G. Fant & M. A. A. Tatham (Eds.). *Auditory Analysis and Perception of Speech* (pp.115-141). London: Academic Press.

Vance, T.J. (1976). An experimental investigation of tone and intonation in Cantonese. *Phonetics, 11,* 368-392.

Wong, P.C.M. (1998). *Speaker Normalization in the Perception of Cantonese Level Tones: Effect of Context-Target Pitch Distance.* Unpublished Master of Art thesis, The University of Texas, Austin, Texas.

Wong, N.K.Y. (2001). *The Effect of Contextual Cues on the Perceptual Normalization of Cantonese Level Tones.* Unpublished Bachelor's degree dissertation, The University of Hong Kong.