Contextual Effect on Perception of Lexical Tones in Cantonese

Joan K-Y Ma, Valter Ciocca & Tara Whitehill

Division of Speech and Hearing Sciences
University of Hong Kong, Hong Kong
joanma@hkusua.hku.hk

10.21437/Interspeech.2005-272

Abstract

The present study investigated the role of tonal context (extrinsic information) in the perception of Cantonese lexical tones. Target tones at three separate positions (initial, medial and final position) were recorded by two speakers (one male and one female). These sentences were edited and presented in three conditions: original carrier (target within the original context), isolation (target without context) and neutral carrier (target word as appended at the final apposition within a new carrier). Nine female listeners were asked to identify the tones by matching targets with Chinese characters. Perceptual data showed that tones presented within the original carrier were more accurately perceived than targets presented in isolation, showing the importance of extrinsic information in the perception of lexical tones. In the neutral carrier condition, tones of the final position showed perceptual accuracy significantly above targets of the initial and medial positions. The perceptual error patterns suggested that listeners placed more emphasis on the immediate context preceding the target in tone identification. When tones were presented without an extrinsic context, the proportion of errors for each tone differed. Most of the errors involved misidentifying targets as tones of same F0 contour but different level. The results showed that the importance of extrinsic information on the perception of lexical tones was mainly on identification of F0 level while the intrinsic acoustic properties of the tone helped in identifying the F0 contour.

1. Introduction

In tone languages, F0 variation at the syllabic level is used to mark lexical meaning. For example, in Cantonese the word /ma/ means "mother" when produced in high-level tone (/ma_{55}/), but it means "horse" when produced with low-rising tone (/ma_{23}/). Fundamental frequency (F0) is the primary acoustic correlate of tone identification in Cantonese [1, 2]. F0 values for different tones are not absolute but are relative to the F0 range of each talker [1, 3]. Therefore, it is possible that a high-level tone produced by a male speaker has similar F0 level to a low-level tone produced by a female speaker. Besides inter-speaker variability in tone production, the F0 level of a tone produced by the same speaker may vary due to reasons such as intonation [4], downdrift [4, 5], as well as temporary F0 fluctuations. Therefore, an overlap in the F0 patterns of different tones is expected owing to both inter-and intra-speaker variability.

Although the absolute F0 level of different tones can overlap across talkers, listeners seem to have little difficulty in resolving this potential ambiguity in tone perception. Several studies have shown evidence that Mandarin and Cantonese listeners make use of the tonal context of an utterance (extrinsic information) to identify lexical tones accurately ("tone normalization") [2, 6, 7]. Leather [6] used two synthetic Mandarin tones in four natural speech carriers in a tone identification task. He found that tone identification was affected by differences in carrier phrases, and hypothesized that the F0 range inferred from the carrier phrase is an important factor for tone perception. Moore and Jongman [7] used three sets of stimuli varying between tone 2 (mid-rising) and tone 3 (falling-rising) in terms of turning point (or the inflection point of the tone), ΔF0 (the difference in F0 between the onset and the turning point) or both. Targets were presented following two precursors that differed only in F0 level. Their findings supported the existence of an effect of extrinsic tonal context on tone identification. Similar findings were obtained in a study on perception of Cantonese tones [2].

Leather [6] proposed that listeners infer the F0 range of a talker when processing an utterance, and that the extracted F0 range is used to normalize for tone identification. Wong and Diehl [2] further investigated this proposal by presenting the listeners targets following a carrier sentence. The carrier was divided into two halves, and the F0 levels of the first and the second halves were raised or lowered independently. For example, the first half was raised by two semitones and the second half was lowered by three semitones. They claimed that listeners used a "recency" strategy in tone identification, i.e., listeners relied on the F0 level of the second half of the carrier for tone normalization. Listeners were less dependent on the immediate context only when the F0 level of the carrier violated downdrift, (i.e., when the first half of the carrier had lower F0 level than the second half). In this case, listeners used the first half of the carrier for tone normalization. These findings suggest that both halves of the carrier may be used for tone normalization. It is possible that shifting the two halves of the carrier sentence, as Wong & Diehl [2] did, resulted in an unnatural sounding sentence. Therefore, listeners may have depended more on the immediate context as they were unable to infer F0 information from both halves of the carrier equally. One purpose of this study was to further investigate the role of contextual cues such as immediate F0 context using natural sentences.

In Cantonese, there are six contrastive tones: high-level (55), high-rising (25), mid-level (33), low-falling (21), low-rising (23) and low-level (22) [8]. Moore & Jongman [7] stated that the intrinsic acoustic properties (F0 level and contour) specific to each lexical tone might be sufficient for the correct identification of Mandarin tones that contrast in both F0 height and contour. However, the presence of three level tones and two rising tones in Cantonese makes perception based only on intrinsic acoustic properties unlikely, as F0 height is a significant factor in differentiating tones of the same contour. For example, Fok-Chan [1] reported confusion between tones 33 and 22 in her perceptual experiment in Cantonese with the targets presented in isolation; she proposed that the confusion is due to the absence of extrinsic context cues with these stimuli. Fok-Chan [1] also reported that tones 55 and 21 were more resistant to confusion in the absence of extrinsic context when compared with other tones. This suggested that perception of the six tones might depend on the extrinsic context to different degrees as the

intrinsic F0 changes of some tones (e.g. the falling contour of tone 21, as it is the only falling tone in Cantonese) may be of saliency to the listeners for correct tone identification. Therefore, another purpose of this study was to investigate the relative importance of extrinsic context cue in perception of all six tones in Cantonese.

2. Method

2.1. Listeners

Nine females were recruited as listeners. They were all first year undergraduates in the Division of Speech and Hearing Sciences, the University of Hong Kong. Their age ranged from 18 to 19 years old. They were considered naïve listeners as the experiment was carried out within the first two months of their enrollment, during which they received limited training in speech perception. Cantonese was the native language of all the listeners. Hearing screening was performed with all the listeners by the first author. All listeners passed the hearing screening (≤ 20 dBHL at 250, 500, 1000, 2000 and 4000 Hz).

2.2. Stimuli

Speech materials were collected from two native Cantonese speakers (one male and one female). Three carrier phrases were used, with the target word in initial position (/X tsi_{22} hou_{25} lan_{21} $sɛ_{35}$/ 'X is difficult to write'), in medial position (/$sɛ_{25}$ $kɔ_{33}$ X tsi_{22} sin_{55}/ 'Write the word X first') and in final position (/lei_{55} $kɔ_{33}$ tsi_{22} $hɐi_{22}$ X/ 'This word is X'). Three sets of target words, derived from the syllables /si/, /ji/, /jɐu/, were embedded in the above sentence contexts. Each set consisted of six words that differed only in tone. With three contexts and eighteen target words, there were a total of 54 different stimuli produced by each speaker.

Three presentation conditions were used in the current experiment – the original carrier, the neutral carrier and the isolation form. For the original carrier, the target tones were presented within the original sentence frame that the speaker produced. Target tones occurred at the initial, medial or final positions as described above. Wong and Diehl [2] hypothesized that listeners would be able to normalize for F0 changes in downdrift within the original carrier, as listeners were found to place more emphasis on the immediate context preceding the target. Therefore, it is predicted that all six tones would be relatively well perceived in this condition. The second presentation condition, target tones in isolation, was designed to evaluate how (intrinsic) F0 changes are perceived when extrinsic contextual cues are removed. Target tones placed at the initial, medial and final positions were manually extracted from the original carriers using the Praat software [9]. The beginning and the end of the target tones were selected by looking at the waveform display and the wide-band spectrogram. Fok-Chan [1] noted that tones can be difficult to perceive in isolation. In the present experiment, the loss of cues for speaker normalization was compensated for by blocking stimuli by speaker. Because of the lack of extrinsic context, we expected that tones in this condition would be less accurately identified than tones within the original carrier. In the third presentation condition, all targets of the speaker were presented within the same neutral carrier. In this presentation condition, all the target tones extracted from the initial, medial and final positions of the original carriers were placed at the final position of the carrier / lei_{55} $kɔ_{33}$ tsi_{22} $hɐi_{22}$/ (This word is). A neutral carrier was synthesized from the production of the original carrier for targets in final position by the two speakers. For each speaker, the average F0 of each syllable of all the 18 productions of /lei_{55} $kɔ_{33}$ tsi_{22} $hɐi_{22}$/ was calculated. Among the 18 productions, the utterance with F0 closest to the average in all four syllables was chosen. The F0 of each syllable within the selected utterance was then resynthesized using the PSOLA algorithm of Praat software [9] to be within 2 Hz of the average F0 for each syllable. The two synthesized carriers were judged to sound natural by the first author and two other native Cantonese speakers (all qualified speech therapists). All the target tones were then added as the final syllable of the neutral carrier. This presentation condition was designed to investigate the strategies listeners used in normalizing for tone perception. Wong and Diehl [2] proposed that listeners placed emphasis on the immediate context in tone perception. Therefore, target tones in initial and medial positions should be less accurately perceived than tones presented with the original carrier and tones in final positions. This is because targets of the initial and medial positions had higher F0 than naturally produced targets at the final position.

2.3. Procedures

The experiment was carried in a sound-attenuated room (IAC booth), with speech materials presented to listeners through a Sennheiser HD 545 headset, connected to an Apple Macintosh G4 computer. A HyperCard programme was used to run the experiment. For each trial, one set of six Chinese characters representing the six words contrasted by tone was presented on the screen. For original carrier and neutral carrier conditions, the carrier was also displayed at the top of the screen.

The stimuli were divided into six blocks according to presentation condition (original carrier, neutral carrier and isolation) and talker. Within each block, there were a total of 108 trials, as each of the 54 stimuli was repeated once. In each session, two blocks (male and female) of one presentation condition were presented. The three sessions were scheduled at least one week apart to minimize learning effects. The order of presentation of blocks was counterbalanced across listeners.

Before each session, the 18 Chinese characters were introduced and read aloud to the listeners by the first author, as some Chinese characters have more than one pronunciation in different contexts. Within each trial of the experimental blocks, listeners were asked to match the word they heard with one of the Chinese characters by clicking on the button representing that character. Each stimulus was presented once; the listener could listen to each stimulus a second time by clicking on a "repeat sound" button. Each block took about fifteen minutes to finish, and each session took about half an hour.

2.4. Data analysis

Confusion matrices were compiled separately according to presentation conditions (original carrier, neutral carrier and isolation form) and position (initial, medial and final). The identification patterns of each individual listener were constructed, and group confusion matrices were composed by summing confusion matrices across the nine listeners. The confusion matrices for the group for neutral carrier and isolation conditions are displayed in Tables 1 to 6. The numbers in the cells represents the percentage of responses being realized as that particular tone. For example, in Table 1, tone 25 of the initial position presented in isolation was accurately identified 101 times out of 108 trials, resulting in an identification percentage of 93.5%.

3. Results

3.1. Presentation condition

The mean percentages of correct response were 98.2% for targets presented with original carrier (SD = 1.2), 81.1% for

targets presented with neutral carrier (SD = 3.8), and 78.8% for targets presented in isolation (SD = 3.1). A series of Wilcoxon matched pair tests was performed to compare the results between the three conditions. The results demonstrated statistically significant differences between the original carrier and the other two presentation conditions (between original carrier and neutral carrier: $T = 0$, $p < 0.05$; between the original carrier and the isolation condition: $T = 0$, $p < 0.05$). However, there was no significant difference between the neutral carrier and the isolation condition ($T = 11$, $p > 0.05$),

3.2. Original carrier

All six tones presented with the original carrier were accurately perceived. Tone 55 was perceived with 100% accuracy (SD = 0), while both tone 21 and tone 22 were perceived with 99.4% mean accuracy (SD = 1.9 for tone 21 and SD = 1.2 for tone 22). The perceptual accuracy was slightly lower for tone 25 (mean = 97.8%, SD = 1.9) and lowest for both tone 33 (mean = 96.3%, SD = 5) and tone 23 (mean = 96.3, SD = 2.8). Perceptual accuracy of tones was also compared across position of the target word within the phrase. Targets at the final position (mean = 97.5%, SD = 2.2) were most accurately perceived, followed by that of the initial position (mean = 99.3%, SD = 1.4) and the medial position (mean = 97.8%, SD = 1.9).

3.3. Isolation

For targets presented in isolation, tone 23 (mean = 96.6%, SD = 6.2) was most accurately perceived, followed by tone 55 (mean = 88.6%, SD = 10.1). Wilcoxon matched pair test showed no significant difference between tones 55 and 23 ($T = 3.5$, $p > 0.05$). Performance for these two tones was significantly above that for tones 25, 33, 21 and 22 ($p < 0.05$ for all), except for the difference between tones 55 and 21 ($T = 20$, $p > 0.05$). Tone 21 was perceived with mean accuracy of 87.4% (SD = 11.9) while the mean perceptual accuracy for tone 25 was 75% (SD = 7.7). There was no significant difference between tones 21 and 25 ($T = 5$, $p > 0.05$); accuracy for these tones was significantly higher than that of tones 33 (mean = 59.9%, SD = 14.77) and 22 (mean = 65.1, SD = 5.8) ($p < 0.05$). Tones 33 and 22 were the least accurately identified among the six tones when presented in isolation. The difference between tones 33 and 22 was not significant ($T = 11.5$, $p > 0.05$). In the isolation condition, targets were perceived with similar accuracies across the three positions. Targets were perceived with mean percentage accuracy of 78.2% (SD = 7) at the initial position, 78.6% (SD = 5.3) at the medial position, and 79.5% (SD = 8.6) at the final position. A series of Wilcoxon matched pair tests was used to compare the results, and no statistically significant difference was found between the three positions ($p > 0.05$ for all tones).

3.4. Neutral carrier

For targets presented with a neutral carrier, tones 55 (mean = 97.2%, SD = 5.6) and 23 (mean = 93.2%, SD = 4.8) were perceived most accurately. Wilcoxon matched pair test showed no significant difference between these tones ($T = 5$, $p > 0.05$). Performance for tones 55 and 23 was significantly better than for the other four tones (tones 25, 33, 21 and 22) ($p < 0.05$ for all). Tones 25 (mean = 80.9%, SD = 6), 33 (mean = 79.7%, SD = 9.8) and 21 (84%, SD = 9.8) had similar identification accuracy ($p > 0.05$). Tone 22 was perceived with the lowest accuracy (mean = 51.5%, SD = 13.8); differences with all other five tones were significant ($p < 0.05$ for all comparisons with tone 22). In comparing accuracy across positions, targets at the final position were most accurately identified (mean = 93.8%, SD = 5.3), followed by those in

Table 1. Initial position presented in isolation.

Target	Perceived Tones					
	55	**25**	**33**	**21**	**23**	**22**
55	100					
25		93.5			1.9	4.6
33	9.3		83.3			7.4
21				68.5		31.5
23		4.6	0.9		94.4	
22			68.5		1.9	29.6

Table 2. Medial position presented in isolation.

Target	Perceived Tones					
	55	**25**	**33**	**21**	**23**	**22**
55	86.1		12		1.9	
25		38.9		0.9	60.2	
33			63.9		0.9	35.2
21				97.2		2.8
23			0.9		99.1	
22		0.9	8.3	4.6		86.1

Table 3. Final position presented in isolation.

Target	Perceived Tones					
	55	**25**	**33**	**21**	**23**	**22**
55	79.6		20.4			
25		92.6	0.9	2.8	3.7	
33			32.4	1.9	1.9	63.9
21				96.3	1.9	1.9
23		2.8	0.9		96.3	
22				20.4		79.6

Table 4. Initial position presented with neutral carrier.

Target	Perceived Tones					
	55	**25**	**33**	**21**	**23**	**22**
55	100					
25		94.4	4.6		0.9	
33	30.6		67.6			1.9
21				58.3		41.7
23		7.4	5.6		86.1	0.9
22			86.1		2.8	11.1

Table 5. Medial position presented with neutral carrier.

Target	Perceived Tones					
	55	**25**	**33**	**21**	**23**	**22**
55	97.2		2.8			
25		50			50	
33			89.8		1.9	8.3
21				98.1		1.9
23			0.9		97.2	1.9
22			52.8	0.9		46.3

Table 6. Final position presented with neutral carrier.

Target	Perceived Tones					
	55	**25**	**33**	**21**	**23**	**22**
55	94.4		4.6	0.9		
25		98.1		0.9	0.9	
33			81.5			18.5
21				95.4	0.9	3.7
23		3.7			96.3	
22				2.8		97.2

Tables 1-6. Confusion matrices for targets at different positions (initial, medial and final) and presentation conditions (isolation and neutral carrier). Cell numbers represent the percentage of responses for each target tone.

medial position (mean = 79.8%, SD = 3.5) and in initial position (mean = 69.6%, SD = 10.8). Statistical analysis showed that the difference between all three positions were significant (initial and medial positions: $T = 2$, $p < 0.05$; initial and final positions: $T = 1$, $p < 0.05$; and, medial and final positions: $T = 0$, $p < 0.05$).

4. Discussion

The difference in performance across presentation conditions demonstrated the role of extrinsic context in tone perception. Among the three presentation conditions, tones presented with the original carrier were the most accurately perceived. While all six tones at the three positions (initial, medial and final) were accurately identified within the original carrier, performance was poorer for tones presented in isolation across the three positions. This finding was expected since Leather [6] pointed out the importance of talker normalization in tone perception. In the original carrier condition, the carrier provided listeners cues to infer the F0 range of the speakers and therefore, helped the listeners to normalize for tone perception. When presented in isolation, the lack of context for the listeners to normalize for the speaker's voice reduced the listeners' ability to perceive tones accurately.

Performance for targets presented with neutral carrier was significantly worse than for those presented with the original carrier. While the accuracy for targets of the final position of the neutral carrier condition was above 90%, the accuracy for targets of the initial and medial positions was significantly lower than that of the final position. By looking at the confusion matrices, most of the errors at the initial and medial positions involved misperceiving targets as a tone of the same contour but of higher F0 level (e.g. tone 22 → tone 33); errors were most common for tones 33, 21 and 22 in initial position and tone 22 in medial position. Because of downdrift, words at the initial and medial positions have typically higher F0 level than the same tone at the final position [4, 5]. Therefore, if listeners use the recency strategy for tone normalization, they should perceive these targets in final position as tones of the same contour but higher F0 level. Listeners in the current experiment used the context immediately preceding the final syllable for tone identification, thereby supporting Wong and Diehl's claim [2] of the existence of a recency strategy in tone perception. Except for tone 22 of the initial position, perceptual errors for tones 33 and 21 in initial position and tone 22 in medial position accounted for 50% or less of the trials of the specified tones. Therefore, some listeners were able to correctly identify these tones although F0 levels were higher than the listeners' expectation given the immediate preceding context and the effect of downdrift. It is hypothesized that there are other acoustic cues (e.g. the initial F0 level of the carrier) in addition to cues provided by the immediate extrinsic context that are employed by these listeners for correct tone identification.

Performance for the six tones in isolation differed, showing that some tones are more affected than others by the absence of extrinsic context. Moore and Jongman [7] suggested that, for some tones, the intrinsic acoustic properties (F0 level and contour) might provide sufficient information for correct identification. However, in Cantonese, the three level tones share the same F0 contour and differ only in F0 level. Among the three level tones, tone 55 was least affected when contextual cue was removed, and its perceptual accuracy was significantly higher than the other two level tones. Fok-Chan [1] suggested that tone 55 is distinctive owing to its relatively high F0 level and the greater relative distance when compared with the other two level tones. By contrast, the absence of contextual cues affected the perception of tones 33 and 22. As tones 33 and 22 are relatively close in F0 level [1], the lack of context caused frequent confusions between these tones. For the two rising tones, tone 23 was more accurately perceived than tone 25 when presented in isolation. Referring to the Tables 4 to 6, most of the perceptual errors for tone 25 involved targets at the medial position and was most often confused with tone 23. Similar to the level tones, the error response was of the same tone contour but differ only in tone level. For tone 21 at the initial position, listeners showed confusion with the low-level tone (tone 22), which has a slightly falling contour [3]. Tone 21 has the lowest F0 level among the six Cantonese tones [3]. When produced at the beginning of an utterance, these targets have relatively high F0 level compared to tone 21 produced in other positions. Therefore, confusion in tone identity with a tone of similar contour but higher F0 level was noted when tone 21 of the initial position was presented without extrinsic context. The confusion pattern between the six tones in the absence of context demonstrated that the intrinsic acoustic properties of tone do have impact on tone perception, as most of the confusion found in this presentation condition was within the same tone contour but confused in tone level.

5. Conclusions

The present study showed that extrinsic tonal context plays a significant role in tone perception. When presented with a natural context, listeners relied heavily on the immediate preceding context (recency effect) in identifying the lexical tones. Among the six tones, the absence of extrinsic context exerts the smallest effect on the perception of tones 55 and 23. For the other tones, the intrinsic acoustic properties helped in identifying the contour of the tone, but contextual cues played a significant part in correct identification of F0 level.

6. References

[1] Fok-Chan, Y. Y. (1974). *A perceptual study of tones in Cantonese*. Hong Kong: University of Hong Kong Press.

[2] Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter-and intratalker variation in Cantonese level tones. *Journal of Speech and Hearing Research, 46*, 413-421.

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

[4] Ma, J. K.-Y., Ciocca, V. & Whitehill, T. (2004). *The effects of intonation patterns on lexical tone production in Cantonese by acoustic analysis.* Paper presented at the International Symposium on Tonal Aspects of Languages, Beijing, China.

[5] Ohala, J. J. (1978). Tone Rules. In V. A. Fromkin (Ed.), *Tone: A linguistic approach* (pp. 5-39). New York: Academic Press.

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

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

[8] Chao, Y. R. (1947). *Cantonese primer*. New York: Greenwood Press.

[9] Boersma, P. & Weenink, D. (2003). *Praat Program.* Retrieved February, 27, 2003, from University of Amsterdam, Institute of Phonetics Sciences website: http://www.fon.hum.uva.nl/praat/