Speech Prosody 2006
Dresden, Germany
May 2-5, 2006

Perception of Cantonese level tones influenced by context position

Hongying Zheng[a], *Gang Peng*[b], *Peter W-M. Tsang*[a] & *William S-Y. Wang*[b]

[a]Department of Electronic Engineering
City University of Hong Kong, China
[b]Department of Electronic Engineering
The Chinese University of Hong Kong, China
H.Y.Zheng@student.cityu.edu.hk

Abstract

When humans perceive speech sounds, they categorize the sounds into one or another phoneme category. Perception of speech sound depends on context. Previous studies on categorical perception of lexical tones were mainly done in an absolute manner without context. In these experiments we explore the influence of context on the categorical perception of lexical tones. In particular, we ask whether the position of the context with respect to the target syllable influences the categoricalness of the perception. Two experiments on natural and synthesized speech both show that categorical boundaries of identification curves are sharper when the context is to the right of the target syllable than when the context is to the left of the target syllable. Moreover, steeper peaks are obtained in the discrimination curve from right context continuum. They agree with and enhance the identification results. Explanations of the phenomenon are suggested in the paper.

1. Introduction

In tone languages, words are made up of consonants, vowels, as well as tones. There are 3 level tones in Cantonese. They are high-level (tone1, /55/), mid-level (tone3, /33/) and low-level (tone6, /22/)* [1].

When comprehending speech, a listener uses strategies to categorize a continuum of variants along some acoustic dimensions into 'one or another of the phoneme categories that his language allows' [5]. This behavior is called categorical perception [CP].

When the target phoneme is embedded in context, perception of the target phoneme not only depends on changes on the phoneme itself but also depends on changes in the context. We call this behavior relative perception [RP].

Studies of CP for consonants and vowels with context influence are documented (e.g. [7]). Although there have been some studies of CP on lexical tones on an isolated syllable (e.g. [2] [8]), there has been none done with contextual influence. In these experiments we explore the influence of context on the CP of tones. In particular, we investigate how the position of the context with respect to the target syllable [TS] influences the categoricalness of the perception.

Two experiments, experiment I and experiment II, on Cantonese three level tones were carried out, based on natural and synthesized speech respectively. Experiment Ia was an identification task and experiment Ib was a discrimination task. Experiment II was an identification task only. All of the tasks contained two subtasks. One was to investigate perception of TS with left context [LC], i.e., TS was at the end of the context sentence. The other was to investigate perception of TS with right context [RC], i.e., TS was at the beginning of the context sentence.

2. Experiments

2.1. Experiments Ia and Ib

Experiment Ia was an identification task; Experiment Ib was a two-step discrimination task (AX or same/different task).

2.1.1. Subjects

16 native Cantonese speakers, university students in Hong Kong, with no reported history of speaking or hearing disability, participated in this pair of experiments (aged 18-33). 13 subjects (6 male and 7 female) participated in experiment Ia. 4 subjects (3 male and 1 female) participated in experiment Ib. One subject participated in both tasks.

2.1.2. Stimuli

Table 1: F_0 distribution of stimuli based on natural sentence in LC and RC continua

No. of Stim.	LC sentence $F_{0/TS/}$=127Hz		RC sentence $F_{0/TS/}$=136Hz	
	F_0(Hz) of /hai^6/	Dist. (Hz)	F_0(Hz) of /hai^6/	Dist. (Hz)
11	133	-6	140	-4
10	128	-1	135	1
9	123	4	130	6
8	118	9	125	11
7	***113***	***14***	***120***	***16***
6	108	19	115	21
5	103	24	110	26
4	98	29	105	31
3	93	34	100	36
2	88	39	95	41
1	83	44	90	46

Stimuli in experiments Ia and Ib were resynthesized sentences based on natural speech, which were recorded from a native male Cantonese speaker. The LC was represented by the sentence: /ni^1 go^3 zi^6 hai^6 TS/$^\nabla$ (This word is TS). The RC was represented by the sentence: /TS hai^6 mat^1 ji^3 si^3/ (What's the meaning of TS). There were three TSs for each of LC and RC sentences. The TSs were: /fan^1/ (divide), /fan^3/ (sleep) and /fan^6/ (share), which differed from each other only in pitch value. The 6 sentences uttered each three times were recorded.

* In Chao's notation, voice pitch (F_0) ranges from /1/(low) to /5/ (high).

$^\nabla$ Transcriptions enclosed between slashes are written in Jyut ping, according to the Linguistic Society of Hong Kong [LSHK]. The superscript numerals denote the tone category of the syllable.

The best one among the 3 utterances for each sentence, according to judgment by another two native Cantonese speakers, was chosen for further manipulation.

The difference in mean F_0 between the /hai^6/ and /fan^1/ was measured, which we might call anchor 1, and the values were 44Hz and 46Hz in LC and RC sentences respectively. Similarly, anchor 2 was obtained from the difference in mean F_0 between the /hai^6/ and /fan^6/. They were -6 Hz and -4Hz in LC and RC sentences respectively. LC and RC sentences with the TS of /fan^3/ were baselines. The difference in F_0 between anchor 1 and the baseline was equally divided at 5 Hz step. Hereby, we obtained 7 points as reference (No.1- 7 in Tab. 1). Similarly, 4 reference points (No.8-11) were obtained between anchor 2 and the baseline. Totally there were 11 points including 2 endpoints. No.7 was the baseline point.

The F_0 contours of LC syllables (/ni^1 go^3 zi^6 hai^6 /) and RC syllables (/hai^6 mat^1 ji^3 si^3/) in the baseline sentences were adjusted using PRAAT< http://www.fon.hum.uva.nl/praat>, according to the reference points, to make 11 stimuli, while TSs in the sentences were kept unchanged with $F_{0/TS/}$=127Hz in LC and $F_{0/TS/}$ =136Hz in RC. Distribution of F_0 distance between TSs and their immediate neighbors (/hai^6/) is shown in Tab. 1 and stimuli structure is shown in Fig. 1. Distance in Tab. 1 was calculated according to Eq. (1):

$$Dist. = F_{0/TS/} - F_{0/hai^6/} \quad (1)$$

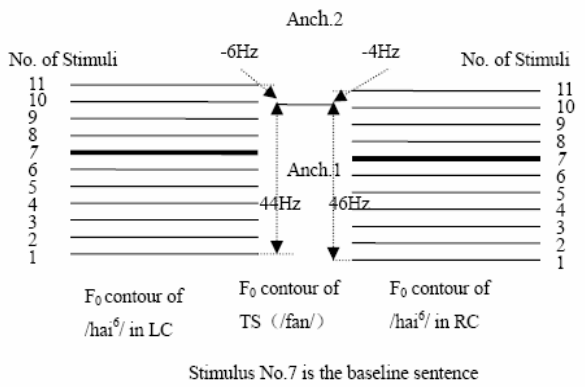

Figure 1: *Stimuli continua structure of experiment I. The step in the continua is 5Hz. We anticipate TS in stimulus No.1 to be perceived as /fan^1/ and TS in stimulus No.11 to be perceived as /fan^6/.*

2.1.3. Procedure

2.1.3.1 *Experiment Ia: identification task*

25 trials (2 endpoints and the baseline sentences repeated 3 times and other sentences repeated twice) for each subtask (LC and RC) were presented to listeners via a loudspeaker. These 25 trials were randomized into 5 blocks. Each block contained 5 trials, separated by 4s silence gap (Inter-trial interval, ITI). On each trial, listeners heard a pair of two identical tokens (sentences) consecutively with a silence gap of 0.5s (Inter-stimulus interval, ISI). After a trial, listeners were asked to identify which TS they heard by circling one from the three given choices, even if they are not sure about the answer. The three choices were the three TSs mentioned above appearing in Chinese characters.

Totally, 325 responses (25trials * 13 subjects) were collected in this task for each of RC and LC continua.

2.1.3.2 *Experiment Ib: Discrimination task*

Stimuli were presented to listeners over a SONY headphone (MDR CD-777). The stimuli for this task consisted of all pairwise combinations of individual sentences separated by zero or two tokens along the continuum, with a 500ms ISI. There was a total of 29 such pairs for each of LC and RC continua. 29 pairs repeated each 3 times were distributed into 15 blocks randomly. Each block contained 6 pairs (with 6s ITI), except for the last block which only contained 3 pairs. The subjects were instructed to select 'yes' or 'no' on paper to indicate whether the TSs in a pair were the same or different.

Totally, 348 responses (29 pairs*3 times *4 subjects) were collected in this task for each of RC and LC continua.

2.1.4. Results

Identification and discrimination responses averaged across listeners are presented in Figs. 2 and 3 for LC and RC continua respectively. The score of discriminations for each pair is calculated following descriptions in [2]. The figures show that listeners exhibit crossovers in their identification function. Three categories are obtained along the continuum. The crossovers correspond roughly to the expected location of boundaries between the three tone categories (at the 3-5 and 9-11 pairs in LC continuum and at 4-6 and 7-9 pairs in RC continuum[*]). Category boundaries in Fig. 3 (RC continuum) are sharper than those in Fig. 2 (LC continuum). Moreover, in RC continuum, peaks in obtained discriminations (dotted line with a circle) are steeper than those in LC. Peaks of the discrimination curve roughly correspond to the crossovers of identification curve in RC but not in LC.

To compare the identification with discrimination performance, discrimination function predicted from the identification results is introduced. The predicted discrimination function for each two-step pair is also superimposed on the identification figure (dotted line with a cross) in Figs. 2 and 3. The predicted discrimination curve is calculated by Eq. (2), taken from [2]:

$$p(disc_{ij}) = 0.5 + 0.25\{[p_{LL}(i) - p_{LL}(j)]^2 + [p_{ML}(i) - p_{ML}(j)]^2 + [p_{HL}(i) - p_{HL}(j)]^2\} \quad (2)$$

Where $p(disc_{ij})$ is the predicted probability of discriminating stimulus No.i and No.j (e.g. 3 and 5 in 3-5 pair). $p_{LL}(i)$ is the measured proportion of low-level responses to stimulus No.i, and so forth. The higher value in the discrimination function indicates listener can more easily perceive the difference in the pair.

A one-way ANOVA provided by SPSS shows there is a significant effect of pair ($F(8,108) = 7.956$, $p<0.001$) for $p(disc_{ij})$ in RC continuum but not ($F(8,108)=0.413$, $p=0.911$) in LC continuum. According to post hoc Tukey HSD tests, 4-6 pair and 7-9 pair are significantly different from other pairs in RC continuum. These two peak values of predicted discrimination function in RC continuum are 0.75 and 0.74. Pearson's R test provided by SPSS shows a significant correlation of discrimination between predicted and obtained in RC continuum ($R=0.785$, $n=9$, $p=0.006$, one-tailed) but not in LC ($R=-0.29$, $n=9$, $p=0.225$, one-tailed).

In this pair of experiments, mean F_0 of TSs in LC and RC continua were different. The difference of F_0 between the

[*] Only two major crossovers (tone1and tone3, tone3 and tone6) were considered

/hai^6/ and TS was also different for each corresponding stimulus in LC and RC sentences. So, it is hard to tell whether the difference of categoricalness is due to context position or simply higher mean F_0 value in RC continuum. Experiment II, where TS and F_0 distance were kept strictly equal, was carried out to rule out the influence of this latter factor.

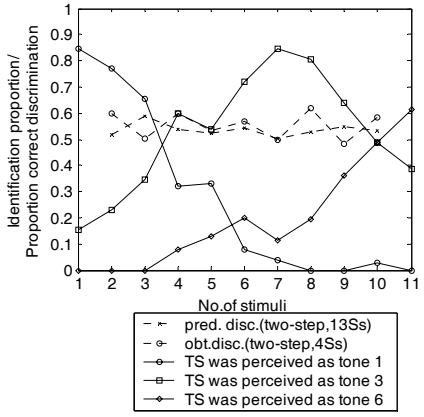

Figure 2: *Identification and discrimination functions of TSs in resynthesized sentences based on natural speech with F_0 contour of LC syllables changed.*

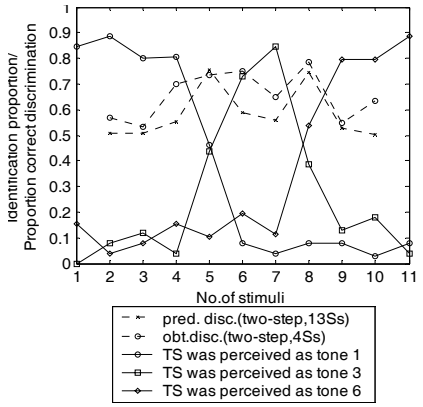

Figure 3: *Identification and discrimination functions of TSs in resynthesized sentences based on natural speech with F_0 contour of RC syllables changed.*

2.2. Experiment II

2.2.1. Subjects

11 native Cantonese speakers* ， university students in Hong Kong, 7 male and 4 female (aged 25-33), with no reported history of speaking or hearing disability, participated in this experiment. None of them participated in experiment I.

2.2.2. Stimuli

Stimuli of experiment II were resynthesized sentences based on synthesized speech generated from CUTALK (a text-to-speech engine) < http://dsp.ee.cuhk.edu.hk/speech/cutalk>. The LC was presented by the sentence: / tau^4 sin^3 nei^5 teng^1 TS/ (You have just heard the word TS). The RC was presented by the sentence: / TS teng^1 dak^1 hou^2 hou^2/ (The word of TS is heard quite clearly). The TSs were: /si^1/(poem), /si^3/(hobby) and /si^6/ (thing).

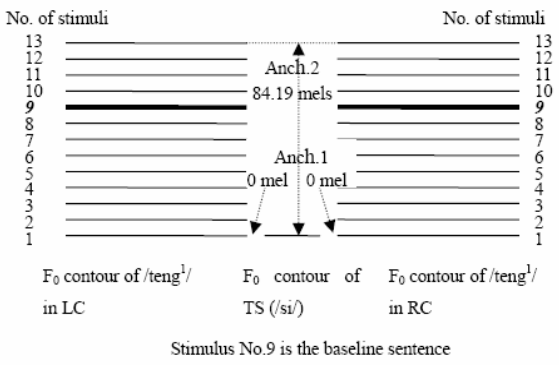

Figure 4: *Stimuli continua structure of experiment II. The step in the continua is 7.0 mels. We anticipate TS in stimulus No.1 to be perceived as /si^1/ and TS in stimulus No.13 to be perceived as /si^6/.*

Table 2: F_0 distribution of stimuli based on synthesized sentence in LC and RC continua

No. of Stim.	F0(Hz) of /teng^1/	Step (Hz)	Dist. (mel)	Dist. (Hz)
13	197.6	5.6	-84.2	-64.6
12	192.0	5.5	-77.2	-59.0
11	186.5	5.5	-70.2	-53.5
10	181.0	5.5	-63.1	-48.0
9	***175.5***	***0***	***-56.1***	***-42.5***
8	170.1	-5.4	-49.1	-37.1
7	164.7	-5.4	-42.1	-31.7
6	159.3	-5.4	-35.1	-26.3
5	154.0	-5.3	-28.1	-21.0
4	148.7	-5.3	-21.1	-15.7
3	143.4	-5.3	-14.0	-10.4
2	138.27	-5.2	-7.0	-5.2
1	133.0	-5.2	0	0

The calculation of Dist. was shown in Eq.(1), and $F_{0/TS/}$=133Hz was kept constant.

Three sentences of LC and RC each, with three TSs, were generated from CUTALK. The engine generated the identical sound for /teng^1/ and three TSs for LC and RC sentences. The three TSs differed only in F_0 value. Similarly, 2 anchors for LC and RC sentences were obtained by measuring the F_0 difference between /teng^1/ and TSs. The distance between the 2 anchors was larger than that in experiment I, so mel scale was chosen for step unit. 13 stimuli spaced by 7.0 mels step were constructed as shown in Tab. 2 and Fig. 4 with the same procedure as in experiment I. $F_{0/TS/}$ and difference of F_0 between /teng^1/ and TS were the same in RC and LC continua.

2.2.3. Procedure

Questionnaires in soft copy, embedded with the speech stimuli in wave file format, were distributed to subjects. All subjects listened to the speech files by headphone.

13 tokens in each of the RC and LC continua were randomized into a block with 1.5s ITI. Each of these two subtasks contained such 9 blocks. The first block was a practice block, which was not counted into the response. Subjects followed the same procedure as in experiment Ia.

* 20 subjects participated in this experiment in total. However, 11 among those were chosen, who respond correctly above chance level (50%) at the two anchors.

Totally, 1144 (13stimuli*8blocks*11subjects) responses were collected in the experiment for each of RC and LC continua.

2.2.4. Result

Results of listeners' average identification response of TS in the LC and RC continua are similar to experiment I. Three categories of perceived TS (/si/) are obtained along each of the LC and RC continua. The crossovers correspond roughly to the expected location of boundaries between the three tone categories (at the 2-4 and 9-11 pairs in LC continuum and at 1-3 and 6-8 pairs in RC continuum). Category boundaries in RC are sharper than those in LC continuum.

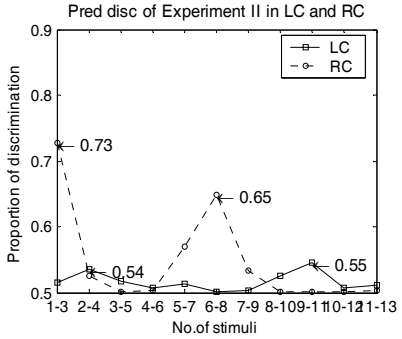

Figure 5: Discrimination (two-step) predicted from identification task in experiment II. Solid line is discrimination curve obtained from LC continuum and dot line is obtained from RC continuum.

The predicted discrimination from average identification response in experiment II is shown in Fig. 5. The predicted discrimination curve is also calculated according to Eq.(2). Similarly, steeper peaks are obtained from the discrimination function in RC continuum than those in LC continuum. The peak values in RC are 0.73 and 0.65; while the peak values in LC are 0.54 and 0.55. The peak values in RC continuum are a little lower than those in experiment I, which indicates the degree of CP is lower than that in experiment I. An ANOVA test shows peaks are significant ($F(10,110)=11.45$, $p<0.001$) in RC, but not significant in LC ($F(10,110)=1.02$, $p=0.434$). 1-3 pair and 6-8 pair are significantly different from other pairs in RC continuum according to post hoc Tukey HSD tests.

Since the identification response obtained from experiment II is similar to experiment I, a similar discrimination result is predicable and not shown here.

3. Discussion

Although the degree of CP is different in the two experiments, comparison of the categorical boundaries obtained from identification performance shows that boundaries in RC are sharper than those in LC continuum. Experimental results from discrimination task show obvious peaks at the category boundaries in RC continuum, but no obvious peaks are observed in LC continuum. It indicates that listeners are able to divide the continuum into categories more sharply in RC. In other words, perception of RC speech continuum is more categorical than that of LC. This context position effect accords with the findings in fricative consonants [7]. A possible explanation for this is a dual processing model [3] for speech perception. When the TS is at the end of a sentence, listeners have set up the reference F_0 range for the sentence. Then, when they perceive the TS, they are more sensitive to the details, which bias them to auditory process. However, when the TS is presented to listeners before they set up the reference F_0 range, they store the most significant information in their short-term memory storage before listening to the next syllable. The pre-categorizing process helps listeners to divide the continuum more sharply. This dual process can also be found in more general cognitive process [4]. Another explanation suggested by the reviewers is that conflicting information in the syllable initial position leads to the gradient perceptual boundary, according to the time structure model of syllable proposed in [9]. Although these two hypotheses are plausible and complementary for each other, more experimental data are needed for further investigation.

4. Conclusion

Identification results of two experiments show that pitch changing in adjacent syllables will influence the perception of the TS. Categorizing tones not only depends on the absolute pitch value of TS but also on the pitch distance between the TS and adjacent syllables, which is consistent with the findings on Mandarin tones [6].

Many conditions were varied in these two experiments; these include the use of a loudspeaker vs. headphones, different pools of subjects, differences in stimuli as well as experimental procedure. Nonetheless, the context position effect is observed across all these conditions though to different extents. This leads us to conclude that the RC produces a greater categoricalness in the perception of Cantonese level tones. We also conclude that CP is an effect that is sensitive to many factors, including the position of the speech context demonstrated here.

5. Acknowledgement

This work is supported by a RGC grant: CUHK-1224/02H. We thank the anonymous reviewers for their constructive comments and suggestions.

6. References

[1] Chao, Y. R. 1947. *Cantonese primer*. Cambridge, MA: Harvard University Press.

[2] Francis, A. L, Ciocca, V. & Ng, B. K. C. 2003. On the (non) categorical perception of lexical tones. *Perception & Psychophysics.* 65(7):1029-1044.

[3] Fujisakj, H. & Kawashima, T. 1970. Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute*.

[4] Harnad, S. *(ed.)* 1987. *Categorical perception: the groundwork of cognition.* New York: Cambridge University Press.

[5] Liberman, A. M., *et al*. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental Psychology: Human Perception and Performance*. 54: 358-368.

[6] Lin, T. & Wang, W. S-Y., (林燾和王士元) 1984. An experiment in tone perception (Trans.) (聲調感知問題). *Journal of Chinese Linguistics (中國語言學報)*. 2: 59-69.

[7] Mann, V. & Soli, S. D. 1991. Perceptual order and the effect of vocalic context on fricative perception. *Perception & Psychophysics*. 49(5): 399-411.

[8] Wang, W. S-Y. 1976. Language change. *Annals of the New York Academy of Sciences*. 280: 61-72.

[9] Xu, Y. & Liu, F. (in press). Tonal alignment, syllable structure and coarticulation: Toward an integrated model. To appear in *Italian Journal of Linguistics*