

Categorical Perception of Cantonese Tones in Context: a Cross-Linguistic Study

Hongying Zheng¹, Peter WM. Tsang², William S-Y. Wang¹

¹ Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China

² Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China

hyzheng@ee.cuhk.edu.hk, eewmtsan@cityu.edu.hk, wyswang@ee.cuhk.edu.hk

Abstract

When human beings perceive speech sounds, they categorize the sounds into one or another phonemic category. The mechanism which is responsible for this phenomenon remains unknown. Is it influenced by listeners' long term language experience or does it reflect some general psychoacoustic aspects of processing? Previous study shows Cantonese level tones are perceived continuously in citation forms [1]. However, they are perceived categorically in the presence of context [6]. The work in [6] does not provide enough evidence to support the hypothesis of long term language influence. In this study, we compare Mandarin and Cantonese speaker's perception on Cantonese level tones in context, and Cantonese speaker's perception on speech and nonspeech (analogous complex harmonics) in the same context as well. Results show evidence of categorical perception on speech stimuli for Cantonese speakers only. These findings support the hypothesis of long term language influence.

Index Terms: speech perception, categorical perception, tone, nonspeech, cross-linguistic

1. Introduction

50 years ago, Liberman and his colleagues found that when listeners comprehend a continuum of synthetic speech sounds, which varies along some acoustic dimensions, they categorize the continuum into 'one or another of the phoneme categories that his language allows' [3]. Even more, listeners are easier to discriminate two equally separated stimuli when they are located across a phonemic boundary than when they are located within the same phonemic category. This phenomenon is called categorical perception [CP]. The most controversial aspect about CP is whether this phenomenon reflects some aspects of language experience specific processing or it is a product of general psychophysical processing.

Lexical tones in tonal language, which are mainly determined by pitch contours, play the same role in lexical meaning as consonants and vowels do. Since perception of pitch contours reflects general psychophysical processing and language specific processing in tonal language, more and more CP studies now focus on lexical tones [1][5][6]. Tonal languages have different tone inventories. For example, Cantonese has 3 level tones: high-level (tone 1), mid-level (tone 3) and low-level tones (tone 6); while in Mandarin and Fuzhou dialect, there is only one level tone: high-level (tone 1). Comparison of perception performance on the same tone pattern by different tonal language speakers will provide more evidence on effects of language experience.

Up to the present, most of the CP studies have been done with the citation form of target syllable [TS]. In the daily conversation, speech is continuous and made up of many

syllables. When the TS is embedded in context (with neighbor syllables), perception of the TS not only depends on changes on the syllable itself but also depends on changes in the context. We call this behavior relative perception.

Previous studies have shown that when Cantonese level tones are presented in context, the degree of CP is increased [1][6]. However, since those studies obtain the results only from native speakers, it is unknown whether if the increase of CP degree is only an artifact in the presence of context or whether it is related to language specific influence. In these experiments we explore the mechanism of contextual influence on CP by two comparisons. One is the comparison of perception on the same speech stimuli by Cantonese and Mandarin speakers. The other is the comparison of perception on different TSs (real speech and nonspeech) by Cantonese speakers. If the increasing degree of CP is an artifact in the presence of context, the perception performance by Cantonese and Mandarin speakers will be the same and there will be no performance difference between perceptions on different TSs by Cantonese speakers as well. However, if there are differences in the above conditions, the mechanism which is responsible for the increase of CP degree is not simply an artifact but rather a language related factor.

2. Experiments

Three experiments were carried out: Cantonese speakers listened to the real speech in the experiment 1 (CS: Cantonese + Speech); Mandarin speakers listened to the same stimuli in the experiment 2 (MS: Mandarin + Speech) and Cantonese speakers listened to the nonspeech in the experiment 3 (CN: Cantonese + Nonspeech). The nonspeech stimuli were constructed by replacing the TS with analogous complex tones. So that they had the same pitch contours, amplitude and duration parameters as the speech sound. Both identification and discrimination tasks were carried out in the experiment 1. Due to the lack of corresponding phonemic labels in Mandarin and nonspeech, only the discrimination task was carried out in the experiments 2&3. In each experiment, two conditions were tested: the left context [LC] condition, where neighbor syllables preceded the TS; the right context [RC] condition, where neighbor syllables followed the TS.

2.1. Subjects

17 paid subjects with no reported history of speaking or hearing disability, participated in the experiments. 8 native Cantonese speakers (4 male and 4 female, aged 18-21), university students in Hong Kong, participated in the experiments 1&3. 9 Mandarin speakers (7 male and 2 female, aged 18-26), university students in Fuzhou participated in the experiment 2.

2.2. Stimuli

Stimuli in the experiments were the same as those used in [6]. They were resynthesized sentences based on the natural speech templates, which were recorded from a native male Cantonese speaker. The LC sentence was /ni¹ go³ zi⁶ hai⁶ TS^V (This word is TS) and the RC sentence was /TS hai⁶ mat¹ ji³ si¹/ (What's the meaning of TS). There were three templates with TSs: /fan¹/ (divide), /fan³/ (sleep) and /fan⁶/ (share) for each LC and RC sentences. These three TSs differed from each other only in the pitch value.

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

No. of Stim.	LC sentence $F_{0/TS}=127\text{Hz}$		RC sentence $F_{0/TS}=136\text{Hz}$	
	$F_0(\text{Hz})$ of /hai ⁶ /	Dist. (Hz)	$F_0(\text{Hz})$ 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

The difference in mean F_0 between the /hai⁶/ and /fan¹/ was measured, which we called 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 /hai⁶/ and /fan³/ . They were -6 Hz and -4Hz in LC and RC sentences respectively. LC and RC sentences with the TS of /fan³/ 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 the reference (No.1- 7 in Tab. 1). Similarly, 4 reference points (No.8-11) were obtained between the anchor 2 and the baseline. Totally there were 11 points including 2 endpoints. No.7 was the baseline point.

The F_0 contours of the LC syllables (/ni¹ go³ zi⁶ hai⁶ /) and the RC syllables (/hai⁶ mat¹ ji³ si¹/) in the baseline sentences were adjusted using PRAAT (a program for doing phonetics by computer <<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 $F_{0/TS}=127\text{Hz}$ in LC and $F_{0/TS}=136\text{Hz}$ in RC. Distribution of F_0 distance between TSs and their immediate neighbors (/hai⁶/) is shown in Tab. 1 and stimuli structure is shown in Fig. 1. Distance in Tab. 1 was calculated according to Eq. (1):

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

The above LC& RC continua were used in the experiments 1&2. In the experiment 3, the TSs were replaced by the nonspeech which differed from speech stimuli only in spectral components but had the same suprasegmental parameters (F_0 contours, intensity profiles and duration). The nonspeech stimuli were constructed by concatenating the fricative portion and a voicing portion. The fricative portion was extracted from the real speech's TS. The voicing portion was composed of six equal-amplitude harmonics (1,3, 6, 7,8, 12) of the F_0 in the voicing part (cf. [5]). Harmonics 2, 4, 5, 9,

10 and 11 were omitted to increase perceptual dissimilarity with the speech stimuli.

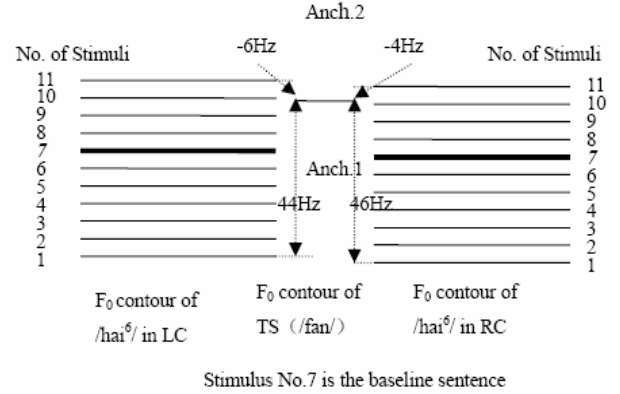


Figure 1: Stimuli continua structure in the experiments. The step in the continua is 5Hz. We anticipate the real speech TS in stimulus No.1 to be perceived as /fan¹/ and that in stimulus No.11 to be perceived as /fan⁶/.

2.3. Procedures

Stimuli were binaurally presented to subjects over a SONY headphone (MDR CD-777) in a quiet room.

2.3.1. Identification task

The identification task was only carried out in the experiment 1. There were 9 blocks of 11 trials for each of LC and RC conditions. Each of the 11 stimuli only was appeared once within a block. The 1st block was treated as practice and was not scored, although the subjects were not aware of this. The presentation order within one block was random. Trials were separated by 4s silence gap (Inter-trial interval, ITI). After a trial, subjects were asked to identify which TS they had heard by circling one from the three given choices, even if they were not sure about the answer. The three choices were the TSs mentioned above appearing in Chinese characters.

2.3.2. Discrimination task

The discrimination task was carried out in all the three experiments. The stimuli for this task consisted of all pairwise combinations of individual stimuli separated by zero or two tokens along the continuum, with a 500ms ISI (e.g. 1-3, 3-1, 2-2, 1-1... et al). There were 29 such pairs for each of LC and RC continua. The 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.

2.4. Data analysis

2.4.1. Obtained discrimination scores

The score of discriminations for each pair was calculated following the descriptions in [1][5]. The proportion of correct discrimination for a stimuli pair included two parts. The first part was the proportion of different responds for the different pairs. The other part was the proportion of same responds for the same pairs. For example, the proportion of correct discrimination for pair 1-3 was the average of the proportion

of the different responds for pair 1-3 and 3-1 and the proportion of same responds for pair 1-1 and 3-3. Results of obtained discrimination scores (dotted lines with open circles) in Exp. 1 were shown in Fig.2. A higher value in the function indicated that the subjects can more easily detect the difference in the pair.

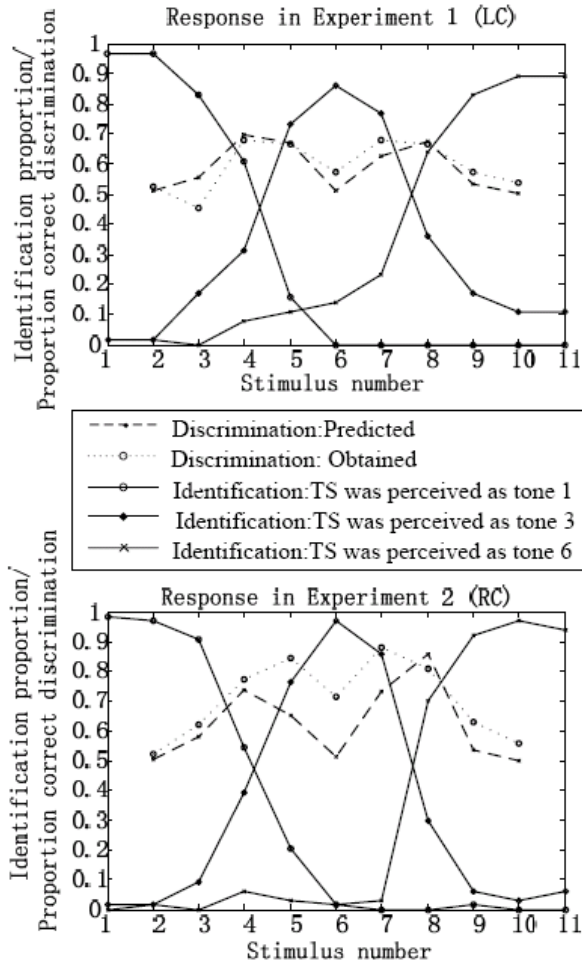


Figure 2: Identification and discrimination functions of TSS in experiment 1 (CS).

2.4.2. Predicted discrimination scores

To compare the identification with discrimination performance, discrimination function predicted from the identification results was introduced. The predicted discrimination function for each two-step pair was superimposed on the identification figure (dashed line) in Fig. 2. The predicted discrimination curve was calculated by Eq. (2), taken from [1]:

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

Where $p(disc_{ij})$ was the predicted probability of discrimination on stimuli pair No.i and No.j (e.g. 3 and 5 in 3-5 pair). $p_{LL}(i)$ was the measured proportion of low-level responses to stimulus No.i, and so forth.

2.4.3. Grouping of discrimination scores

To better illustrate the difference between the cross phonemic category pairs [AC] and the within phonemic category pairs

[WC], the obtained discrimination scores of the 9 pairs of stimuli were grouped into 2 categories according to identification and discrimination results. The cross-over of tone 1 and tone 3 was located in stimulus pair 3-5 and 4-6; the cross-over of tone 3 and tone 6 was located in stimulus pair 6-8 and 7-9. So, AC score was calculated by averaging the discrimination scores of pairs 3-5, 4-6, 6-8 and 7-9 and WC score was calculated by averaging the discrimination scores of pairs 1-3, 2-4, 5-7, 8-10 and 9-11. After the grouping, four averaging discrimination scores (AC in LC, WC in LC, AC in RC and WC in RC) were obtained for each subject in each experiment. The grouped discrimination scores in three experiments (CS=Exp.1; MS=Exp.2; CN=Exp.3) were shown in Fig. 3.

3. Results and discussion

The identification and discrimination responses averaged across listeners in Experiment 1 were presented in Fig. 2 for the LC and RC continua respectively. The scores of obtained and predicted discriminations were superimposed over identification results. Three solid curves were obtained from the identification tasks. Each curve showed the proportion of the specific category labeled at each stimulus. At a specific stimulus, the sum of the total value in three identification curves equaled to 1. It showed that there were two major crossovers on two phonemic categories boundaries (tone1&3 at pair3-5 and 2-4 and tone3&6 at pair7-9 and 8-10).

To examine the statistical significance of the behavior performance, a set of analyses provided by SPSS were carried out. A 2-way repeated ANOVA was carried out on DIR (2 levels: RC and LC)* STIM (9 levels: 9 pairs of stimuli) for obtained discrimination. There were two significant main effects but no interaction effect in the output (DIR: $F(1,7)=8.195$, $p=0.024$; STIM: $F(8,56)=13.241$, $p<0.001$; DIR*STIM: $F(8,56)=1.387$, $p=0.222$). The significant DIR effect showed that subjects performed differently in LC and RC continua. Discrimination score was higher in RC (mean=0.729) than that in LC (mean=0.610) continuum. The significant STIM effect showed that different stimuli pairs were not equally discriminated. No significant interaction DIR*STIM effect showed that the discrimination for the 9 stimuli pairs in LC and RC were independent. In other words, the discrimination patterns were the same in LC and RC continua. To further determine the nature of the pair difference, a 2-way repeated ANOVA on DIR*CAT (2 levels: AC and WC) for grouped obtained discrimination in Exp.1 was carried out. The statistical test showed a similar pattern as the ungrouped data (DIR: $F(1,7)=7.845$, $p=0.026$; CAT: $F(1,7)=29.835$, $p=0.001$; DIR*CAT: $F(1,7)=2.095$, $p=0.191$). Discrimination performance across phonemic category boundaries (AC in RC: mean=0.8307; AC in LC: mean=0.6849) was significantly higher than that within the phonemic category (WC in RC: mean=0.6267; WC in LC: mean=0.5347). The difference on AC and WC met one of the CP criteria proposed in [2] and [3]. To test if the performance in experiment 1 met the other CP criterion proposed in [2] and [3], a Pearson's R test was done. The test showed a significant correlation of discrimination between predicted and obtained in RC continuum ($R=0.411$, $n=72$, $p<0.001$, two-tailed) and slightly significant correlation in LC ($R=0.249$, $n=72$, $p=0.035$, two-tailed). The significant correlation meant the discrimination can be well predicted from the identification performance. Above three sets of statistical tests revealed that speech continua were categorically perceived by Cantonese speakers. The analyses

also showed that RC continuum was easier to discriminate than LC continuum. The different p value indicated the degree of CP in RC was possibly higher than that in LC.

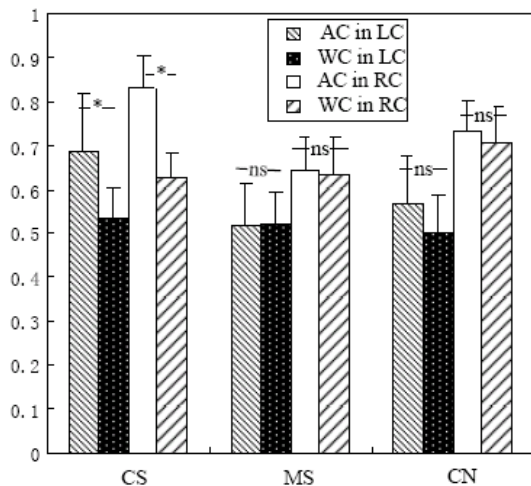


Figure 3: Discrimination performance in three experiments. Four types of bars represent across category and within category in left and right context conditions. Error bars represent standard deviation and '*' means significant difference at 0.01 level, 'ns' means no significant difference.

Table 2: 2-way ANOVA analyses for obtained discrimination results in three experiments. '*' indicates significant at 0.05 level

		CS	MS	CN
Ungrouped	DIR	P=0.024*	P=0.001*	P=0.004*
	DIR*STIM	P<0.001*	P=0.595	P=0.221
	DIR*STIM	P=0.222	P=0.075	P=0.1
Grouped	DIR	P=0.026*	P<0.001*	P=0.004*
	DIR*CAT	P=0.001*	P=0.901	P=0.199
	DIR*STIM	P=0.191	P=0.784	P=0.072

In the experiment 2&3, since only discrimination task was carried out, Pearson's R test was omitted. 2-way ANOVAs for grouped and ungrouped data showed the similar results which were summarized in Tab.2 and Fig.3. Different from the results in CS, there was no significant AC and WC difference in MS (Exp.2) and CN (Exp.3). These results meant that nonspeech was not categorically perceived by Cantonese speakers and Cantonese speech was not categorically perceived by Mandarin speakers as well. A 3-way mixed ANOVA with LANG (2 levels: Cantonese and Mandarin) as between-subject factor and DIR and CAT (grouped data) or STIM (ungrouped data) as within-subject factors was carried out to examine the performance difference of two groups of subjects. The results showed LANG effect (grouped: $F(1,15)=12.778$, $p=0.003$; ungrouped: $F(1,15)=9.45$, $p=0.008$) and 2 way interaction of CAT*LANG ($F(1,15)=23.572$, $p<0.001$) or STIM*LANG effect ($F(8,120)=5.913$, $p<0.001$), which revealed reliable group difference on discrimination of speech continua. A 3-way repeated ANOVA, with TYPE (2 levels: speech and nonspeech) * DIR* CAT, was carried out to test the performance difference on two set of continua by the Cantonese speakers (CS & CN). Although CAT depended on TYPE (CAT*TYPE: $F(1,7)=20.163$, $p=0.003$), no significant main TYPE effect was obtained. This meant Cantonese speakers discriminate equally well in speech and nonspeech continua but CP degree was different for two sets of continua.

4. Conclusions

The comparison of discrimination performance on CS, MS and CN confirms our hypothesis that language experience plays a role in inducing CP on Cantonese level tones. The fact that only CS shows significant CP indicates that CP in this experimental setting is not a byproduct of domain general processing in the presence of context but rather a language specific processing which relates to long-term training. However, the similar DIR effect and no main TYPE effect in CS and CN show that accurate detection of pitch contours is relative independent of segmental features. The results are comparable to the neural cognitive findings on lexical tones [4], which show the dissociation of the neural basis for lexical tones and vowels. However, the statistical analyses also show difference on degree of CP in speech and nonspeech. This indicates that tone categories can only be retrieved within speech scope. The results in CS further supported the results shown in [1] and [6] that when the target phoneme is embedded in the context, degree of CP increases and when the target phoneme is preceded by the reference context (LC) the discrimination performance is worse than the reverse (RC). Although the results in CS show both CP effect on LC and RC continua, the statistical result indicates the degree of CP in these two continua is slightly different which does not violate the results in [6].

5. Acknowledgements

The experiments were done in City University of Hong Kong, and data analyses were done in the Chinese University of Hong Kong. This work has been supported in part by grants from CUHK RGC1224/02H and 1127/04H, Hong Kong. We also thank anonymous reviewers' valuable comments.

6. References

- [1] Francis, A. L., Ciocca, V. & Ng, B. K. C. 2003. On the (non) categorical perception of lexical tones. *Perception & Psychophysics*. 65(7):1029-1044.
- [2] Harnad, S. (ed.) 1987. *Categorical perception: the groundwork of cognition*. New York: Cambridge University Press.
- [3] Liberman, A. M., Harris, K. S., Hoffman, H. S. and Griffith, B. C. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54(5): 358-368.
- [4] Liu, L., Peng, D., Ding, G., Jin, Z., Zhang, L., Li, K. and Chen, C. 2006. Dissociation in the neural basis underlying Chinese tone and vowel production. *NeuroImage* 29(2): 515-523.
- [5] Xu, Y., Gandour, J. T. and Francis, A. L. 2006. Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America* 120(2): 1063-1074.
- [6] Zheng, H. Y., Peng, G., Tsang, P. W.-M. and Wang, W. S.-Y. 2006. Perception of Cantonese level tones influenced by context position. 3rd International Conference on Speech prosody, Dresden, Germany.

[∇] 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.