INTER-TALKER VARIATION AS A SOURCE OF CONFUSION IN CANTONESE TONE PERCEPTION

Caicai Zhang[a], *Gang Peng*[a,b] *& William S-Y. Wang*[a]

[a]Language Engineering Laboratory, The Chinese University of Hong Kong, Hong Kong;
[b]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
yzcelia@gmail.com; gpeng@ee.cuhk.edu.hk; wsywang@ee.cuhk.edu.hk

ABSTRACT

This study examined the effect of inter-talker variations on the perceptual confusion of Cantonese tones. We found that: (1) identification accuracy of six unchecked tones in Cantonese was influenced by inter-talker variations in a way that high tones like T1 and T2 were identified more accurately in high voices whereas low falling tone T4 were identified more accurately in low voices; (2) across the board, inter-talker variations had a relatively limited effect on the identification of low rising tone T5 and low level tone T6; (3) the confusion patterns across these six tones revealed that inter-talker variations resulted in perceptual confusion among tones with a similar F0 contour, but not those with different F0 contours. Findings of this study imply that inter-talker variations could be a driving force for Cantonese tone merger.

Keywords: tone perception, inter-talker variations, Cantonese, tone merger

1. INTRODUCTION

This study investigates the effect of inter-talker variations on the tone identification in Hong Kong Cantonese. Specifically speaking, this study aims to examine (1) whether inter-talker variations which introduce acoustic overlapping in fundamental frequency (F0) between tones would consequently result in perceptual confusion; (2) whether inter-talker variations are a possible driving force for the tone merger in Cantonese.

Cantonese contrasts six unchecked tones, e.g. T(one) 1 /i/55 醫 ‘doctor’, T2 /i/25 倚 ‘to lean’, T3 /i/33 意 ‘meaning’, T4 /i/21 兒 ‘son’, T5 /i/23 耳 ‘ear’, and T6 /i/22 二 ‘two’ [1]. T1, T3 and T6 are described as level tones which are mainly distinctive from each other in F0 height. Two rising tones, T2 and T5 share a similar F0 onset and differ only in the magnitude of slope. Due to inter-talker variations, T3 produced by one speaker could be overlapping in F0 with T1 or T6 produced by another speaker. Similarly, T2 produced by a speaker could be overlapping with T5 produced by another. Therefore, it is likely that inter-talker variations result in confusion in tone perception.

Exploring the mechanism of perceptual confusion has important implications for understanding the tone merger in Cantonese. It has been widely reported that several Cantonese tones are undergoing the merging process, for example, T3 and T6, and T2 and T5 [1, 10]. Given that inter-talker variations may give rise to perceptual confusion of tones, it is possible that inter-talker variations serve as a driving or accelerating force for the tone merger in Cantonese.

Many previous studies address the issue of individual variations from the production point of view (cf. [1, 7, 8]). For those studies investigating tones from perceptual perspective (cf. [3]), few of them incorporated inter-talker variations into the design. Wong and Diehl [9] studied the effect of inter-talker variations on the identification of Cantonese level tones via a comparison of mixed-talker design (more than one talker in a block) and blocked-talker design (only one talker in a block). They found significantly higher identification accuracy in the blocked-talker design than the mixed-talker design. It implicitly demonstrated that the unexpected inter-talker variations within a block resulted in perceptual confusion, and therefore, lowered the identification accuracy. But this study did not report F0 ranges of talkers tested in the experiment, making it difficult to evaluate how different these voices were from each other. Moreover, only male voices were used.

The present study aims to complement previous perceptual studies by explicitly examining the effect of inter-talker variations on perceptual confusion through a controlled and balanced design of talker’s variations.

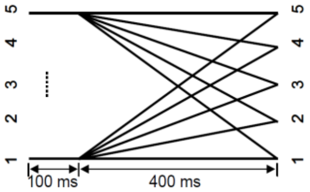
2. METHODOLOGY

16 native speakers of Hong Kong Cantonese (8 M, 8 F; mean age = 20.4 yr, s.d. = 0.78) were recruited.

2.1. Stimuli

25 pitch stimuli were designed in terms of Chao's tone letters [2]. As there is no concave or convex tone in Cantonese, we only used pitch stimuli that can be described by two tone letters. Allowing both tone letters to be any of the five levels that Chao proposed gives rise to a total combination of 25 stimuli (5 × 5 = 25), i.e., 11, 12, 13, 14, 15, 21, 22, … 51, 52, 53, 54 and 55. As shown in Figure 1, F0 contour of each stimulus is a 500 ms linear ramp, with F0 kept constant within the first 100 ms.

Figure 1: F0 trajectories of the pitch stimuli.



Four voice ranges (2M, 2F) were defined based on a Cantonese speech database [5]. First, we obtained average F0 ranges for both male and female talkers in the database. Then we defined two marginal voice ranges, a higher-than-average female voice and a lower-than-average male voice (difference between average and marginal voices is 2 s.d.). Here is the F0 range for each voice: *Female High voice* (FH), 240 ~ 350 Hz; *Female Average voice* (FA): 200 ~ 290 Hz; *Male Average voice* (MA): 110 ~ 160 Hz; *Male Low voice* (ML): 85 ~ 125 Hz. F0 range of each voice covers around 0.54 Octave. Four voices form a continuum from high to low F0, with some overlap between two voices of the same gender.

Four talkers compatible with the above ranges were selected from the database. F0 range of each talker was equally divided into five levels in log scale (1-5), with the upper range aligned to tone letter 5 and lower range to 1. Then 25 pitch stimuli were generated for each voice range according to the F0 contours shown in Figure 1. For each talker, one sample of syllable /i/ was selected and pitch stimuli were superimposed on it. Intensity profile was kept constant across these four voices.

2.2. Task

Pitch stimuli from each voice range were presented in separated blocks (blocked-talker design). Within each block, all 25 pitch stimuli (serving as a sub-block) were randomized and repeated nine times. Each stimulus was presented in isolation; after the stimulus was presented, the subject had 3 seconds to freely identify the heard stimulus as any of the six words (醫 T1, 倚 T2, 意 T3, 兒 T4, 耳 T5, 二 T6) by pressing labeled buttons on a keyboard. Subjects were instructed to respond as quickly and accurately as possible. A practice block containing a voice not occurring in the test blocks was presented first. The order of four test blocks was counterbalanced across the subjects.

3. RESULTS

The purpose of this study being to examine the effect of inter-talker variations on perceptual confusion, we only report the identification of six pitch stimuli that correspond to prototypical Cantonese tones. Based on tone descriptions in [1], the following six stimuli, /i/55 (T1), /i/25 (T2), /i/33 (T3), /i/21 (T4), /i/23 (T5), /i/22 (T6) were defined as prototypical stimuli in this study.

3.1. Identification results

Table 1 shows the mean accuracy and reaction time of six Cantonese tones. Accuracy (ACC) was calculated as the percentage that a pitch stimulus was correctly identified as its corresponding tone. Reaction time (RT) was obtained from correct identification only. RT of incorrect responses is not shown here due to space limit.

Table 1: Accuracy (ACC) and reaction time (RT) of identification of six Cantonese tones. ACCs higher than 0.5 were marked with italic and boldface.

ACC	T1	T2	T3	T4	T5	T6
FH	***0.72***	***0.62***	***0.54***	0.37	***0.63***	***0.67***
FA	***0.71***	***0.52***	0.35	0.31	***0.74***	***0.83***
MA	0.38	0.39	0.43	0.43	***0.61***	***0.68***
ML	0.18	0.35	0.22	***0.55***	***0.63***	***0.56***
Mean	***0.50***	0.47	0.38	0.41	***0.65***	***0.68***
RT	T1	T2	T3	T4	T5	T6
FH	1424.3	1499.4	1445.0	1666.1	1506.2	1381.7
FA	1560.5	1532.3	1384.8	1716.3	1496.5	1327.2
MA	1566.0	1568.1	1471.1	1649.1	1600.1	1531.1
ML	1391.3	1576.2	1447.6	1482.1	1672.4	1326.9
Mean	1485.5	1544.0	1437.1	1628.4	1568.8	1391.7

A *tone* × *voice* repeated measures ANOVA was conducted on ACC and RT separately. For ACC, the statistical analysis revealed significant main effects of *tone*, (F(1, 3)=13.154, p<0.001) and *voice* (F(1, 2.818)=5.682, p=0.003), and a significant interaction of *tone* by *voice* (F(1, 6.382)=7.909, p<0.001). It means that ACC differed among these tones, and was influenced by inter-talker variations. The interaction of *tone* and *voice* suggests that ACC of different tones varied as a function of talker's voices. High tones such as high level tone T1 and high rising tone T2 were identified more accurately in high voices (FH & FA) than low voices (MA & ML), whereas low falling tone T4 was identified more accurately in the low voices (see Table 1).

A one-way ANOVA was conducted to further examine the main effects of *voice* and *tone*. For *voice*, post-hoc Tukey test showed that the overall tone identification was significantly more accurate in FH and FA than in ML (p<0.01 in both cases). Given these results, no conclusive remark about voice typicality can be drawn (i.e. FA and MA are supposedly more typical than FH and ML, but the result does not show that accuracy in FA and MA is significantly higher than FH and ML).

For *tone*, results suggested that low rising tone T5 and low level tone T6 which were stably recognized in all voices, were identified significantly more accurately than the remaining four tones (for all cases, p<0.01). Across the board, the identification of T3 was the least accurate.

For RT, repeated measures ANOVA only found a significant main effect of *tone* (F(1, 5)=4.589, p=0.001), suggesting that the responding time is different between these six tones. A one-way ANOVA showed that T6 was identified significantly faster than T4 and T5 (for both cases, p<0.05). Moreover, T3 was identified significantly faster than T4 (F(1, 5)=4.798, p=0.013).

Figure 2: Accuracy and reaction time of tone identification.

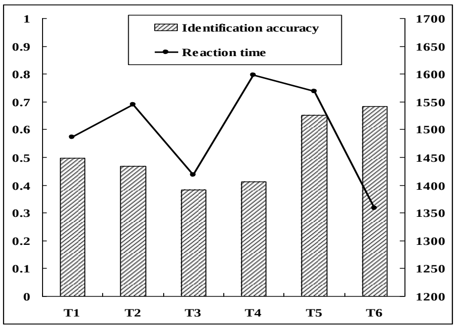


Figure 2 shows the average ACC and RT collapsed across four voices. Examination of this figure suggests a rough match between ACC and RT except for T3, i.e., the higher the accuracy, the faster the response. It possibly indicates that when a listener is confident of his/her choice, he/she tends to respond fast. T3 is exceptional in that it was identified least accurately but its RT was short. We note that when stimulus '33' was misidentified as tones other than T3, its RT was also short (1419.2 ms). It might suggest a speed-accuracy tradeoff in the case of T3, i.e., fast response is achieved at the expense of accuracy. If this is true, it seems to suggest that different responding strategies were adopted for different tones. More research is needed to look into this issue.

3.2. Confusion matrices

Table 2 shows the confusion patterns across six tones. It reveals that inter-talker variations resulted in perceptual confusion between tones with a similar F0 contour (e.g. T1 & T3), but not those with different F0 contours (e.g. T1 & T2).

Table 2: Confusion matrices of tone identification in four voices ranges. Columns represent the pitch stimuli, and rows refer to responses. Identification rates higher than 0.1 were marked with italic and boldface.

FH	55	25	33	21	23	22
T1 (55)	***0.72***	0.02	0	0.01	0.01	0
T2 (25)	0.06	***0.62***	0.05	0.03	0.08	0.03
T3 (33)	***0.1***	0.02	***0.54***	0.04	***0.15***	***0.22***
T4 (21)	0	0.01	0.05	***0.37***	0.01	0.01
T5 (23)	0.02	***0.28***	0.03	0.04	***0.63***	0.03
T6 (22)	0.06	0.02	***0.31***	***0.47***	0.08	***0.67***

FA	55	25	33	21	23	22
T1 (55)	***0.71***	0	0	0.01	0	0.01
T2 (25)	0.01	***0.52***	0.03	0	0.08	0.01
T3 (33)	***0.22***	0.06	***0.35***	***0.1***	0.08	0.08
T4 (21)	0.01	0	0	***0.31***	0	0.01
T5 (23)	0.01	***0.37***	0.05	0.02	***0.74***	0.04
T6 (22)	0.01	0.03	***0.54***	***0.53***	0.06	***0.83***

MA	55	25	33	21	23	22
T1 (55)	***0.38***	0.01	0.01	0.01	0.01	0.01
T2 (25)	0.06	***0.39***	0.01	0.02	0.06	0.03
T3 (33)	***0.41***	0.08	***0.43***	0.04	***0.17***	***0.13***
T4 (21)	0.01	0.05	0.02	***0.43***	0.02	0.04
T5 (23)	0.06	***0.45***	0.03	0.05	***0.61***	0.04
T6 (22)	0.05	0	***0.44***	***0.42***	0.1	***0.68***

ML	55	25	33	21	23	22
T1 (55)	***0.18***	0	0.01	0.02	0.01	0.01
T2 (25)	0.04	***0.35***	0.03	0.01	0.06	0.03
T3 (33)	***0.65***	0.04	***0.22***	0.01	0.06	0.07
T4 (21)	0.01	0.01	0.02	***0.55***	0.05	***0.21***
T5 (23)	0.01	***0.52***	0.03	0.03	***0.63***	0.04
T6 (22)	0.07	0.03	***0.68***	***0.33***	***0.13***	***0.56***

Let us consider three level tones first. In FH, pitch stimulus '55' whose intended tone category is T1 was correctly identified as T1 for 72% of the time (0.72). However, a small portion of this stimulus (0.1) was misperceived as T3. From high voices (FH, FA) to low voices (MA, ML), there was a gradient shift in identification from T1 to T3 so that in the lowest voice (ML), stimulus '55' was predominantly identified as T3 (0.65) instead of T1 (0.18). Similarly, for stimulus '33', its dominant identification was shifted from T3 to T6 in ML. However, stimulus '22' was identified relatively stably as T6 in all 4 voices. Although the ratio that '22' was identified as T6 varies among these 4 voices, there is no turnover in identification dominance from T6 to other tones in any voice.

A similar pattern was found for two rising tones, T2 and T5. Identification of stimulus '25' was changed from T2 to T5 in ML, whereas stimulus '23' was reliably identified as T5 across the voices.

Finally, the identification of pitch stimulus '21' was unstable and tended to be confused with T6. A possible reason is that T4 is traditionally described as either low level or low falling tone. Therefore it is likely to be confused with the low level tone T6.

To summarize, misperception of tones with a similar F0 contour took place when talker's voices changed. Moreover, inter-talker variations seem to have an unequal influence on different tones. For T5 and T6, which were recognized stably across the voices, inter-talker variations may have a limited effect. The remaining four tones, which were often misperceived as other tones, are subject to the influence of inter-talker variations.

4. DISCUSSION AND CONCLUSION

This study examined the effect of inter-talker variations on the perceptual confusion of Cantonese tones. We found that the identification accuracy of Cantonese tones was influenced by inter-talker variations in a way that high tones like T1 and T2 were identified more accurately in high voices whereas low falling tone T4 were identified more accurately in low voices. Across the board, T5 and T6, which were recognized stably across the voices, were relatively resistant to the influence of inter-talker variations. Moreover, the confusion patterns across six tones revealed that inter-talker variations resulted in perceptual confusion among those tones with a similar F0 contour.

To connect findings of this study to tone merger in Cantonese, we hypothesize that inter-talker variations which introduce perceptual confusion among certain tones may give rise to a pool of possible perceptual confusion patterns, (some of) which are later selected in the phonological merging process [4, 6]. Moreover, that inter-talker variations have a relatively limited effect on T5 and T6 may hint on the possible direction of merger, e.g. T2 being merged to T5, and T3 to T6.

Given that inter-talker variations give rise to perceptual confusion of tones, it is possible that inter-talker variations serve as a driving or accelerating force for the tone merger in Cantonese. We suggest inter-talker variations as a driving force, without denying the importance of other factors, such as social, psychological, or linguistic factors, which may contribute to the selection of certain confusing tone pairs in the phonological process. For instance, the confusion matrices showed that T1 was misperceived as T3 in some voices. However, such a confusion trend is not attested in the reported tone mergers. It is also likely that acoustic similarity may interact with inter-talker variations in selecting the candidates for merger. Previous phonetic studies showed that T3 and T6 lie in close adjacency in the acoustic space [1, 7], while both of them remain a distance from T1. Therefore tones with wide acoustic distance (T1 and T3) may be more resistant to confusion-induced sound merging than those of high acoustic similarity (T3 and T6). It is beyond the scope of this study to investigate all factors that contribute to the tone merger in Cantonese.

5. ACKNOWLEDGEMENTS

This work was partially supported by grant from National Science Foundation of China (11074267).

6. REFERENCES

[1] Bauer, R., Cheung, K-H., Cheung, P-M. 2003. Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change* 15, 211-225.

[2] Chao, Y-R. 1930. A system of tone letters. *Le Mai îre Phon étique* 45, 24-27.

[3] Fok Chan, Y-Y. 1984. *A Perceptual Study of Tones in Cantonese*. Hong Kong: Centre of Asian Studies, University of Hong Kong.

[4] Labov, W. 1994. *Principles of Linguistic Change: Social Factors*. Oxford: Blackwell.

[5] Lee, T., Lo, W.K., Ching, P.C., Meng, H. 2002. Spoken language resources for Cantonese speech processing. *Speech Communication* 36, 327–342.

[6] Ohala, J. 1987. Sound change is drawn from a pool of synchronic variation. *Symposium on "The Causes of Language Change, Do We Know Them Yet?"* Norway, 15-17.

[7] Peng, G. 2006. Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of Mandarin and Cantonese. *Journal of Chinese Linguistics* 34, 135-154.

[8] Rose, P. 1996. Cantonese citation tones. In Davis, P. J., Fletcher N. H. (eds.), *Vocal Fold Physiology: Controlling Complexity and Chaos*. San Diego: Singular Pub. Group, 307-324.

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

[10] Yu, A. 2007. Understanding near mergers: The case of morphological tone in Cantonese. *Phonology* 24, 187-214.