Cumulative effects of phonetic context on speech perception

Caicai Zhang[1,2], Gang Peng[2,3], Xiao Wang[3], and William Shi-Yuan Wang[2,3]

[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[3]JRCLHC, The Chinese University of Hong Kong, Hong Kong SAR
caicai.zhang@polyu.edu.hk, gpeng@cuhk.edu.hk, w.joycewang.x@gmail.com, wsywang@ee.cuhk.edu.hk

ABSTRACT

Different speakers produce speech sounds differently. The phonetic context is known to facilitate the recovering of phonological categories from productions with talker variation. However, whether the context effect originates from central auditory processing or speech-related processing remains debated. It is worth noting that the context effect may be a combined effect, contributed by both auditory and speech-related processing. To investigate this question, we compared the effect of four types of contexts with incrementally more cues (nonspeech, reversed speech, meaningless speech and meaningful speech) on perception of Cantonese level tones. Results indicate that the context effect is a product of multiple levels of processing, with the primary contribution from phonological cues (meaningless speech context). The contribution of auditory cues is negligible, and that of phonetic cues and semantic+syntactic cues is both moderate. Phonological cues likely enable listeners to calibrate the acoustic-to-phonological mapping of speech sounds for each talker, facilitating the categorization.

Keywords: Context effect, talker normalization, speech perception, lexical tone, Cantonese

1. INTRODUCTION

Different speakers produce their speech sounds differently. Despite such acoustic variability, most people are able to perceptually categorize speech sounds accurately. An important way that the brain tackles talker variation is via reliance on a context [4, 10]. Listeners can adapt to a particular talker's phonetic space via the context (i.e., what a speaker produced earlier), which serves as a reference for normalizing talker variation. The phonetic context effect has been widely documented on the perception of *consonants* [5, 6, 15], *vowels* [12, 16-19] and *lexical tones* [3, 7-8, 13, 23, 24]. Moreover, the effect is contrastive, such that raising an acoustic cue in the context, e.g., raising the pitch, lowers the perceived pitch of a sound and vice versa.

It remains debated where the cognitive locus of the phonetic context effect is [2, 9, 11, 20, 22]. A number of researchers have found normalization effects induced by both speech and nonspeech contexts [14, 21, 22]. They argue that the context effects of both speech and nonspeech contexts originate from central auditory processing. However, unequal effects of speech and nonspeech contexts have also been reported [3, 18, 23-24]. These researchers have disputed the similarity of nonspeech and speech contexts, arguing that the influence of nonspeech contexts may originate from general auditory processing, but that the effect of speech contexts mostly originates from speech-related levels of processing (cf. [2, 20], suggesting a gestural basis).

It is worth noting that the context effect may have multiple cognitive loci, combining the effect of both general auditory and speech-related processing (e.g., phonetic, phonological, semantic and syntactic processing). Therefore, it is important to disentangle the relative contribution of different levels of processing to the overall context effect.

To this end, we examined the effect of four types of context on the perception of Cantonese level tones in two experiments (see Table 1). These four contexts form a continuum from nonspeech context to meaningful speech context, with incrementally more cues. We subtract the effect of each context from one another to estimate the relative contribution of auditory, phonetic, phonological, and semantic+syntactic cues to the context effect.

Table 1: Summary of context manipulations in Experiment 1 and 2.

Context	Type of cue	Experiment
Nonspeech	Auditory only	1, 2
Reversed speech	Auditory+Phonetic	2
Meaningless speech	Auditory+Phonetic +Phonological	1
Meaningful speech	Auditory+Phonetic +Phonological +Semantic+Syntactic	1, 2

2. EXPERIMENT 1

We examined the nonspeech, meaningless speech and meaningful speech contexts in this experiment.

2.1. Method

2.1.1. Participants

Sixteen native speakers of Hong Kong Cantonese (9F, 7M; age=21±1.2yr) were paid to participate. All subjects reported normal hearing, no musical training and no neurological illness. Informed written consent was obtained in compliance with the experiment protocol approved by the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee.

2.1.2. Stimuli

Materials were recorded from four native speakers of Hong Kong Cantonese (2 F, 2 M) with different F0 height (see Figure 1). From each speaker, a meaningful sentence, 呢個字係意 /li55 ko33 tsi22 hɐi22 ji33/ ('This word is meaning') and a meaningless sentence, 呢錯視幣意 /li55 tsʰo33 si22 pɐi22 ji33/ ('This mistake sees money meaning') were recorded. In both sentences, the final word was the target word, which carries the mid level tone. Meaningful and meaningless contexts were matched in rhymes and tones. The meaningful sentence was semantically neutral, to minimize the effect of semantic expectation on the target. Each syllable in the meaningless context was a morpheme, but their combination had no coherent semantic content.

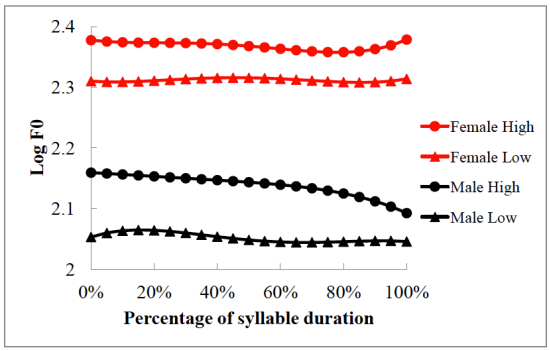
Each talker repeated these two sentences six times. One typical token of the target word (i.e., close to the average F0 of six repetitions) was selected for each talker. Each target word was normalized in duration to 450 ms and in peak intensity level to 55 dB using Praat [1]. One token of the meaningful context and the meaningless context matched in statistical properties of the F0 (mean, minimal and maximal F0) were selected for each talker. Average intensity level of each context was normalized to 55 dB.

We manipulated the F0 of the context to shift the perception of the target word (mid level tone) to the high level tone, mid level tone and low level tone in Cantonese, respectively. For that purpose, the overall F0 trajectory of each context was lowered by three semitones, kept unshifted and raised by three semitones, giving rise to three contextual F0 heights.

Nonspeech contexts were modeled after the F0 and intensity profiles of meaningful speech contexts, and synthesized using a triangle wave that has a different harmonic structure from speech sounds. Average intensity level of the nonspeech context was 75 dB. The target was attached to the end of nonspeech, meaningless contexts after a jittered interval of 300-500 ms for each talker.

In addition, four filler sentences were included. A meaningful sentence, 請留心聽意 /tsʰiŋ25 ləu21 sɐm55 tʰiŋ55 ji33/ 'Please carefully listen to meaning' was recorded from two talkers, and a second sentence, 我以家讀意 /ŋo23 ji21 ka55 tuk2 ji33/ 'Now I read meaning' from the other two talkers. Two meaningless sentences 頂留金青意 /tiŋ25 ləu21 kɐm55 tsiŋ55 ji33/ and 我時花俗意 /ŋo23 si21 fa55 tsuk2 ji33/ were recorded accordingly. Nonspeech counterparts of fillers were synthesized for two meaningful contexts. The ratio of test and filler sentences was 3:1.

Figure 1: F0 contour of the target word produced by four talkers.



2.1.3. Procedure

Stimulus presentation was blocked by the context, with each block comprising stimuli of one context condition. Within a block, 16 trials ((3 test sentences + 1 filler) × 4 talkers) were presented randomly and repeated nine times. Presentation order of three blocks was counterbalanced across the subjects. One practice block with meaningful speech sentences recorded from two additional talkers was presented first to familiarize subjects with the procedures.

Subjects were instructed to identify the target word as any of the three Cantonese words, 醫 (/ji55/ 'doctor'), 意 (/ji33/ 'meaning'), and 二 (/ji22/ 'two') by pressing labelled buttons on a computer keyboard within two seconds. These three words differ exclusively in tone.

2.1.4. Analysis

The target word is expected to be identified as /ji22/ (low level tone, 'two') in the raised F0 condition, as /ji33/ (mid level tone, 'meaning') in the unshifted F0 condition, and as /ji55/ (high level tone, 'doctor') in the lowered F0 condition [3, 23, 24]. Rate of expected word responses was calculated per context condition, per F0 shift and per talker. If a context condition facilitates talker normalization, then the rate of expected responses should be significantly

higher than the chance level (0.33) for each F0 shift and for each talker condition. To this end, one-sample t-tests were conducted to compare the response rates with the chance level for each condition. Repeated measures ANOVAs were conducted but not reported due to space limit.

2.2. Results

For the nonspeech context, the expected response rates are only significantly higher than the chance level in three conditions, i.e., in female high and male low talker conditions in the unshifted F0 condition ($t(15)=5.67$, $p<0.001$; $t(15)=3.42$, $p=0.004$), and female low talker in the raised F0 condition ($t(15)=4.73$, $p<0.001$).

The expected response rates are significantly higher than the chance level in all conditions for both meaningless speech context ($ps<0.01$) and meaningful speech contexts ($ps<0.01$). It suggests that the nonspeech and two speech contexts have unequal effects on the perception of Cantonese level tones. The effect of the nonspeech context is influenced by talker and F0 shift changes, whereas the effects of two speech contexts are significant irrespective of the talker and F0 shift manipulations.

3. EXPERIMENT 2

We examined the nonspeech, reversed speech and meaningful speech contexts in this experiment.

3.1. Method

3.1.1. Participants

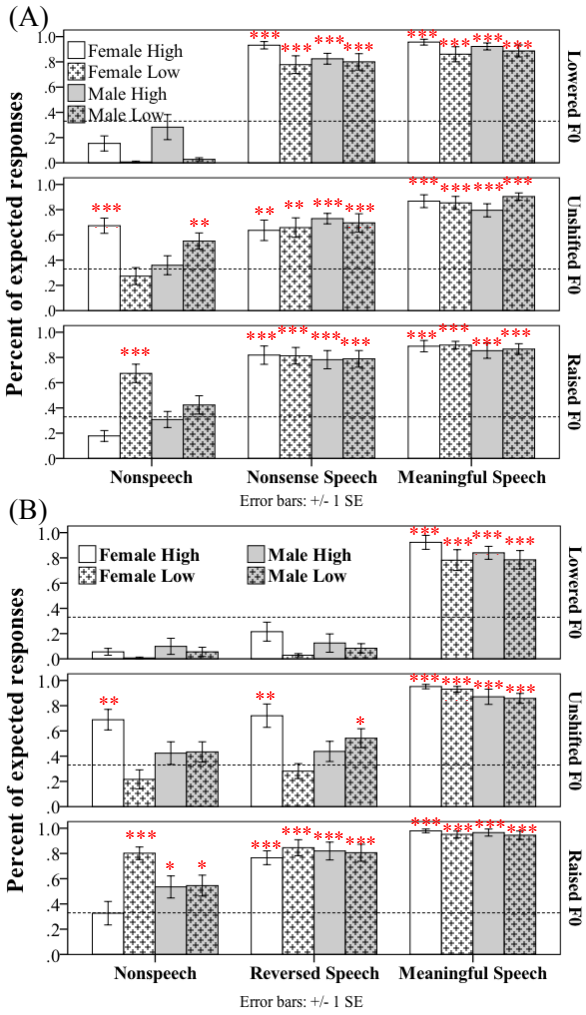
Another group of 16 native speakers of Hong Kong Cantonese (8F, 8M; age=21.3±2.0yr) were paid to participate. All subjects reported normal hearing, no musical training and no history of neurological illness. Informed written consent was obtained in compliance with the experiment protocol approved by the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee.

3.1.2. Stimuli, procedure and analysis

Nonspeech and meaningful speech contexts were identical to those used in Experiment 1. The reversed speech context was generated by time-reversing the meaningful speech contexts. Reversed speech context contain some phonetic information, but it is difficult to map those sounds to Cantonese consonants, vowels, lexical tones or syllables.

Procedure and data analysis were identical to those of Experiment 1.

Figure 2: Rate of expected responses for each context, each F0 shift and each talker condition. (A) Experiment 1. (B) Experiment 2. One-sample t-test: *, $p<0.05$, **, $p<0.01$; ***, $p<0.001$. Dotted lines indicate the chance level (0.33).



3.2. Results

For the nonspeech context, the expected response rates are only occasionally significantly higher than the chance level, i.e., for female high talker in the unshifted F0 condition ($t(15)=4.41$, $p=0.001$), and for female low talker ($t(15)=9.47$, $p<0.001$), male high talker ($t(15)=2.34$, $p=0.033$) and male low talker ($t(15)=2.6$, $p=0.02$) in the raised F0 condition.

The reversed speech context is largely similar to the nonspeech context, except that two more conditions reached significance. Significant effects were found in the following six conditions: female high talker ($t(15)=4.23$, $p=0.001$) and male low talker ($t(15)=2.83$, $p=0.013$) in the unshifted F0 condition; female high talker ($t(15)=7.93$, $p<0.001$), female low talker ($t(15)=8.13$, $p<0.001$), male high talker ($t(15)=6.9$, $p<0.001$) and male low talker ($t(15)=7.03$, $p<0.001$) in the raised F0 condition.

For the meaningful speech context, the expected response rates are significantly higher than the chance level in all conditions (ps<0.001).

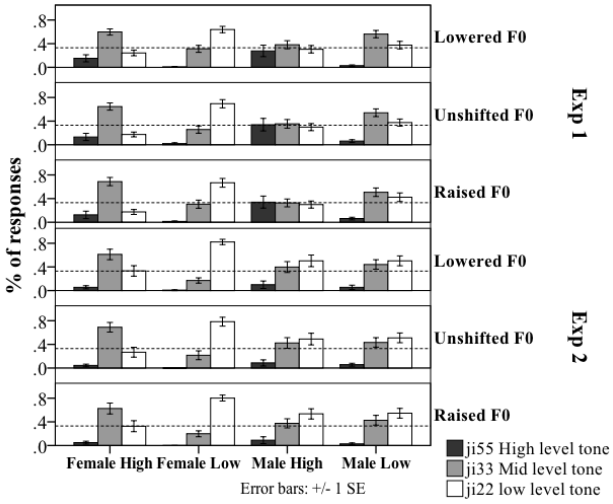
4. OVERALL ANALYSIS

In this analysis, we pooled Experiment 1 and 2 together to estimate the relative contribution of different cues, by subtracting the effect of each type of context from one another (see Table 2).

Table 2: Relative contribution of each type of cues to the overall context effect.

Type of cue	Context subtraction	Effect
Auditory	Nonspeech (congruent F0 shift – incongruent F0 shift)	0.6%~ 0.8%
Phonetic	Reversed speech − Nonspeech	13.5%
Phonological	Meaningless speech – Reversed speech	29.9%
Semantic+ Syntactic	Meaningful speech – Meaningless speech	11.8%

Figure 3: Rate of three word responses in the nonspeech context for each F0 shift and each talker in Experiment 1and 2. Dotted lines indicate the chance level (0.33).



A question is how to estimate the effect size of a context with auditory cues only. One way is to subtract the expected response rates in the nonspeech context from the chance level (33.3%). The other way is to compare the rate of a word response in a congruent F0 shift condition (where that response is expected) with the rate of that same response in two incongruent F0 shift conditions (where it is not expected) (see Figure 3). For example, the mid level tone response is expected in the unshifted F0 condition (congruent), but not in the other two conditions (incongruent). In incongruent F0 shift conditions, the rate of a response is likely random. Rate of that response would be increased by auditory cues of a congruent F0 shift condition. Both analyses reveal similar results. Expected response rates in congruent F0 shift conditions (mean=33.9%, SD=34.7%) are 0.6% higher than the chance level (33.3%), or 0.8% higher than those in incongruent F0 shift conditions (mean=33.1%, SD=33.6%).

The most prominent increase of the context effect is induced by the addition of phonological cues (29.9%). Addition of phonetic cues and semantic+syntactic cues mildly increase the context effect by 13.5% and 11.8% respectively.

5. GENERAL DISCUSSION

In this study, we have found that the phonetic context effect is a product of multiple levels of processing. The contribution of auditory cues is negligible. The contribution of phonetic cues and semantic+syntactic cues is both moderate. The most prominent contribution to the context effect is phonological cues.

Different types of contexts provide different information to facilitate the perception of Cantonese level tones. The reversed speech context likely provides information of a particular talker's pith range via the maximal and minimal pitch deflections in the contour. The meaningless speech context, on the other hand, provides information about the mapping of pitch and tone category for each talker. For example, high level tone as in /li55/ (initial word in the context) can be associated with a mean F0 of 311 Hz in the Female High talker's production. It appears that listeners can rely on phonetic information of a talker's pitch range to categorize level tones, but information about the acoustic-to-phonological mapping for each talker is much more important. Presence of semantic+syntactic cues likely further facilitates the acoustic-to-phonological mapping for each talker and corrects inaccurate mapping via lexical and syntactic constraints. It mildly increases the context effect further.

A remaining question is whether the carrying syllables influence the acoustic-to-phonological mapping. Future studies may compare the effects of a context with native and non-native syllables carrying identical Cantonese tones on the perception of Cantonese level tones.

6. ACKNOWLEDGEMENTS

This work was partially supported by grants from Research Grant Council of Hong Kong (GRF 448413), and National Natural Science Foundation of China (11474300).

7. REFERENCES

[1] Boersma, P., Weenink, D. 2012. Praat: Doing phonetics by computer (Version 5.3.23) [Computer program], http://www.praat.org (Last viewed August 7, 2012).

[2] Fowler, C. A. 2006. Compensation for coarticulation reflects gesture perception, not spectral contrast. *Percept. Psychophys.* 68, 161-177.

[3] Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., Chu, P. C. Y. 2006. Extrinsic context affects perceptual normalization of lexical tone. *J. Acoust. Soc. Am.* 119, 1712-1726.

[4] Gerstman, L. 1968. Classification of self-normalized vowels. *IEEE Trans. Acoust.* AU-16, 78-80.

[5] Holt, L. L. 2005. Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychol. Sci.* 16, 305-312.

[6] Holt, L. L. 2006. Speech categorization in context: Joint effects of nonspeech and speech precursors. *J. Acoust. Soc. Am.* 119, 4016-4026.

[7] Huang, J., Holt, L. L. 2009. General perceptual contributions to lexical tone normalization. *J. Acoust. Soc. Am.* 125, 3983-3994.

[8] Huang, J., Holt, L. L. 2011. Evidence for the central origin of lexical tone normalization (L). *J. Acoust. Soc. Am.* 129, 1145-1148.

[9] Huang, J., Holt, L. L. 2012. Listening for the norm: adaptive coding in speech categorization. *Front. Psychol*. 3, 10.

[10] Joos, M. 1948. *Acoustic Phonetics*. Baltimore: Linguistic Society of America.

[11] Kluender, K. R., Kiefte, M. J. 2006. Speech perception within a biologically realistic information-theoretic framework. In Gernsbacher, M. A., Traxler, M. (eds.), *Handbook of Psycholinguistics*. London: Elsevier, 153-199.

[12] Ladefoged, P., Broadbent, D. E. 1957. Information conveyed by vowels. *J. Acoust. Soc. Am.* 29, 98-104.

[13] Lin, T., Wang, W. S-Y. 1984. Shengdiao ganzhi wenti (Tone perception). *Zhongguo Yuyan Xuebao*, 2, 59-69.

[14] Lotto, A. J., Sullivan, S. C., Holt, L. L. 2003. Central locus for nonspeech context effects on phonetic identification (L). *J. Acoust. Soc. Am.* 113, 53-56.

[15] Mann, V. A. 1980. Influence of preceding liquid on stop-consonant perception. *Percept. Psychophys.* 28, 407-412.

[16] Sjerps, M. J., Mitterer, H., McQueen, J. M. 2011a. Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia*, 49, 3831-3846.

[17] Sjerps, M. J., Mitterer, H., McQueen, J. M. 2011b. Constraints on the processes responsible for the extrinsic normalization of vowels. *Atten. Percept. Psychophys.* 73, 1195-1215.

[18] Sjerps, M. J., Mitterer, H., McQueen, J. M. 2012. Hemispheric differences in the effects of context on vowel perception. *Brain Lang.* 120, 401-405.

[19] Sjerps, M. J., Smiljanić, R. 2013. Compensation for vocal tract characteristics across native and non-native languages. *J. Phon.* 41, 145-155.

[20] Viswanathan, N., Magnuson, J. S., Fowler, C. A. 2013. Similar Response Patterns Do Not Imply Identical Origins: An Energetic Masking Account of Nonspeech Effects in Compensation for Coarticulation. *J. Exp. Psychol. Hum. Percept. Perform.* 39, 1181-1192.

[21] Watkins, A. J. 1991. Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *J. Acoust. Soc. Am.* 90, 2942-2955.

[22] Watkins, A. J., Makin, S. J. 1996. Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *J. Acoust. Soc. Am.* 99, 3749-3757.

[23] Zhang, C., Peng, G., Wang, W. S-Y. 2012. Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones. *J. Acoust. Soc. Am.* 132, 1088-1099.

[24] Zhang, C., Peng, G., Wang, W. S-Y. 2013. Achieving constancy in spoken word identification: Time course of talker normalization. *Brain Lang*. 126, 193-202.