TAL2018, Sixth International Symposium on Tonal Aspects of Languages
18-20 June 2018, Berlin, Germany

Extrinsic normalization of lexical tones and vowels：Beyond a simple contrastive general auditory mechanism

Kaile Zhang[1], *Matthias J. Sjerps*[2], *Caicai Zhang*[1,3], *Gang Peng*[1,3]

[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong
[2]Neurobiology of Language Department, Radboud University, Netherlands
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

kaile.k.zhang@connect.polyu.hk, m.j.sjerps@gmail.com, caicai.zhang@polyu.edu.hk, gpeng@polyu.edu.hk

Abstract

Spoken context provides valuable information for listeners to accommodate speech variability. One example of this influence is extrinsic normalization: the finding that formant and tone ranges in preceding context constrain the interpretation of subsequent tone and formant cues. One dominant hypothesis has been that contrastive general auditory processes play an important role in normalization. A contrastive general auditory mechanism suggests that speech and non-speech contexts should have similar contrastive influences on speech perception. The present study tests this prediction across segmental (formants) and suprasegmental (tone) speech cues. Participants listened to target stimuli that were preceded by either speech or non-speech contexts. Importantly, the cues that distinguished target stimuli were contrastively related to their context. The results demonstrate that speech contexts, but not non-speech context, induced significant contrastive effects on the perception of both lexical tones and formants. In addition, we observed considerable individual difference in the size and direction of context effects. Some listeners reliably demonstrated contrastive context effects while others demonstrated assimilative effects. These results suggest that the underlying mechanism of speech normalization is more complicated than simply contrastive general auditory processes.

Index Terms: speech normalization, lexical tones, vowels

1. Introduction

Speech signals vary a lot, both across speakers and within a speaker. Such variation leads to considerable overlap between spoken instances of different phonological categories. For example, a female speaker's low tone can be higher than a male speaker's high tone in absolute fundamental frequency (F0) [1]. Similarly, vowels of different phonological categories may have similar absolute formant values when they are produced by different individuals [2]. Interestingly, however, listeners can typically accommodate the speech variation, relying on both acoustic and visual information [3], [4], for a range of different speech cues.

Cue distributions in spoken context has been demonstrated to affect how listeners interpret cues in subsequent target words, a process that has been known as "extrinsic normalization" [5]. Speech context is generally reported to affect perception of currently incoming sounds in a contrastive fashion. That is, listeners tend to perceive an identical lexical tone as a high tone if the pitch of the context is relatively low, and as a low tone if the pitch of the context is relatively high [6]. A similar pattern has been demonstrated for the perception of segmental components. An ambiguous vowel [$^{I}_{\varepsilon}$] is more often perceived as the vowel /I/, which has a low first formant (F1), if the preceding sentence has a relatively high F1, and more frequently reported as the vowel /ɛ/ (which has a high F1) when its precursor sentence has a relatively low F1 [5].

On a number of occasions, non-speech analogs have also been shown to induce contrastive context effects on the normalization of lexical tones [7], vowels [8] and consonants [9],[10]. Moreover, contrastive perceptual effects with speech sounds have also been demonstrated in birds [11]. Hence, it has been suggested that normalization effects are largely the result of a general auditory mechanism that affects context-target pairs in a similar way, regardless of whether they involve speech or non-speech signals [11]. One prominent operating characteristic of this mechanism has been suggested to rely on *spectral contrast.* The core property of this mechanism is that while perceiving the target speech, listeners will be more sensitive to those spectral properties that have been relatively suppressed in the context signal [12]. Hence, it has been suggested that normalization relies on the buildup of something similar to a Long Term Average Spectrum (LTAS) of context, and that the subsequent perception of the target is similar to an inverse filtering of the target speech with the properties of the preceding speech's LTAS [13].

However, although statistically significant contrastive context effects have often been observed with non-speech contexts the effect sizes are typically reduced compared to those obtained with speech contexts [8]. Such unequal effect of speech and non-speech contexts indicates that context effects might at least partially rely on speech-specific mechanisms as well [6]. In addition, not only contrastive effects but also integrative effects have been observed, with lexical tones[14], vowels [15],[16], and consonants [17]. Hence the total context effect that is observed may in fact be the net result of a combination of both contrastive and integrative effects.

To re-examine this controversy (speech vs. non-speech context; contrastive vs. integrative effect), the present study tested context effects in relatively comprehensive conditions (tone-/formant- based distinctions; speech/non-speech contexts). In all conditions, the contexts and the targets were manipulated to contain contrastive properties in the main acoustic speech cues (i.e., F0 for lexical tones and F1 for vowels). If extrinsic normalization process does operate via a contrastive general auditory mechanism, normalization should be observed in a contrastive direction (i.e., more low responses in high contexts) regardless of the nature of context

10.21437/TAL.2018-46

(i.e., speech or non-speech) and the target cues (i.e., pitch or formant).

2. Methods

2.1. Subjects

Nineteen native Hong Kong Cantonese speakers (nine males) were paid to participate in the experiment. These participants were undergraduates in Hong Kong Polytechnic University. None reported hearing impairment. Before the experiment, informed written consents were obtained from every participant in compliance with the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University.

2.2. Stimuli

The stimulus manipulation and synthesis procedures were very similar to those reported in [18]. Each trial consisted of two parts: the target (/bo55/, /bo33/, /bo$^{55}{}_{33}$/, /bu55/, or /b$^{u}{}_{o}$55/) and the context (/p^{h}a21 tsi25/ 琶紫 "guitar purple"). A native male Cantonese speaker was recruited to read three nonsense sequences /fo55 p^{h}a21 tsi25/, /fu55 p^{h}a21 tsi25/, and /fo33 p^{h}a21 tsi25/ 25 to 30 times in a soundproof booth. The average F0 and F1 to F5 trajectories were calculated for each syllable (i.e., /fo55/, /fo33/, /fu55/, /p^{h}a21/ and /tsi25/) based on all these repetitions. The neutral version of each syllable (Table 1) was synthesized based on the average pitch and formant frequencies.

Table 1: *The F0 and F1 of each neutral syllable.*

	/fo55/	**/fo33/**	**/fu55/**	**/p^{h}a21/**	**/tsi25/**
F0	168	128	171	94	116
F1	554	552	379	816	307

Notes: The unit is Hz. The F0 value is the average value across the whole pitch contour, while the F1 value is measured at the midpoint of each formant trajectory.

The initial fricatives of /fo55/, /fu55/, and /fo33/ were excised[1]. Theoretically, this manipulation should have resulted in /o55/, /u55/, and /o33/. However, native Cantonese speakers reported that these syllables sound like /bo55/, /bu55/, and /bo33/, respectively, which might be caused by the remaining formant transitions. A 17-step tone continuum changing from /bo33/ to /bo55/ (F0 range: 118-180Hz, F1= 540Hz) and a 17-step vowel continuum changing from /bu55/ to /bo55/ (F0 = 170Hz, F1 range: 292-641 Hz) were generated by manipulating either the F0 or F1 trajectory of the neutral targets in Praat [19]. The duration of each target was normalized to 200 milliseconds (ms) and the intensity was equalized across the targets.

To induce contrastive acoustic properties between contexts and targets, the F0 trajectory of the neutral context / p^{h}a21 tsi25/ was either raised or lowered by 20 Hz and the F1 trajectory of the neutral context was shifted either 100 Hz up or down. The degrees of shifts (i.e., 20 Hz for F0 and 100 Hz for F1) were a tradeoff between the largest shifts possible and the sounds' naturalness. This manipulation formed four types of speech context: high F0-neutral F1 context, low F0 -neutral F1 context, neutral F0-high F1 context, and neutral F0-low F1 context. For each speech context, we also generated a non-speech counterpart. Non-speech signals consisted of iterated rippled noise (IRN). IRN was manipulated to have the same F0 trajectory and overall LTAS as its speech counterpart, yielding four non-speech contexts. The duration of each context was normalized to 565 ms and the intensity was equalized across contexts.

2.3. Procedures

The experiment contained two parts. In phase one, a 10-minute categorical perception task was carried out to identify the most ambiguous targets among the 17-step continuum for each participant, since context effects can be most reliably observed for ambiguous items [17]. The second, main, phase consisted of a word identification task which was designed to evaluate the extrinsic normalization effect of interest, with the selected stimulus set for each subject. Because the stimuli used in the present study are not naturally produced /bo55/, /bo33/, and /bu55/, subjects were familiarized with the three syllables before the experiment. For that familiarization, the endpoints of the lexical tone and vowel continua were embedded in a neutral version of the context. Subjects were asked to listen to the stimuli until they thought they were familiar with them.

2.3.1. Phase one: The categorical perception task

Only stimuli with odd numbers in the vowel or tone contimuum (i.e., Step 1, Step 3…Step 17) were used as the targets in this task to shorten the experimental time. They were embeded in a speech context with both neutral F0 and neutral F1. Stimulus presentation was controlled through Praat [19]. On each trial, a target stimulus was played and then the context was played after 500 ms silence. A window with two choices (/bo55/ and /bo33/ in the tone perception task; /bo55/ and /bu55/ in the vowel perception task) was then shown on the screen after the audio stimuli. Participants were asked to click the button to indicate their choice. Each target was repeated five times and played in a random order. Based on this part, the most ambiguous target items was selected for each participant.

2.3.2. Phase two: The word identification task

The word identification task employed a blocked design. In each block, the context was kept constant, resulting in a total of eight blocks. The eight blocks were presented in a counter-balanced order across subjects. Each block consisted of five types of targets: two endpoints of the tone (or vowel) continuum (Step1 and Step17; 10 repetitions each), the most ambiguous target chosen in phase one (Step X; 60 repetitions), the stimulus before the most ambiguous target in the continuum (Step X-1; 20 repetitions) and the one after the most ambiguous target in the continuum (Step X+1; 20 repetitions). The resulting 120 trials were presented in a random order. On each trial, the target was played bilaterally to subjects via a headphone. After a silence (randomly jittered between 400 ms to 600 ms), the context was played. This was then followed by a question mark on the screen. The blocked design allows for a paradigm where context stimuli occur *after* the target, since the context of one trial *precedes* the target of the next. Subjects were told to pay attention to the audio

[1] The electroencephalographic data was also collected at the same time. The initial fricatives affected the timing accuracy and thus were excised.

stimuli they heard and choose which word the first syllable was once they saw the question mark. Participants responded by pressing the corresponding buttons with a mouse. The maximum allowed response time was 1250 ms.

2.4. Data analysis

A contrastive context effect should be visible as an increase in high tone responses (i.e., /bo55/) in the low F0 context compared to the high F0 context. Similarly, for vowel normalization we expected that more /bo55/ (a vowel with high F1) should be observed in the low F1 contexts compared with the high F1 contexts. The size of the context effect was hereby defined as the proportion of /bo55/ responses in low F0/F1 context minus the proportions of /bo55/ response in high F0/F1 context. If the difference obtained was a positive number, then the context affected speech perception in a contrastive way; otherwise, the context affects the perception of a target in an assimilative fashion.

3. Results

Figure 1 displays the overall proportions of /bo55/ responses in the different context conditions. The size of the context effects was indexed by the distance between the two lines [18]. As can be seen, the speech context induced more notable contrastive context effect than the non-speech context in both lexical tone and vowel (formant) materials. This effect was especially notable on the ambiguous (middle) steps.

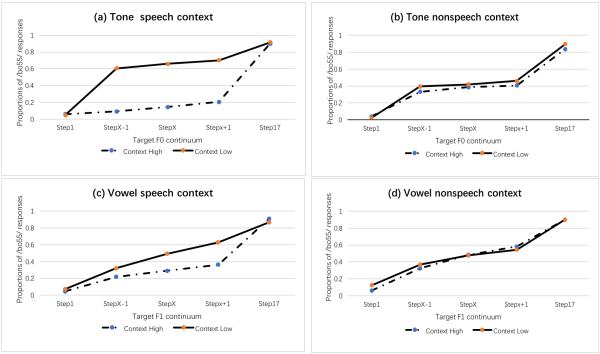

Figure 1: *The proportions of /bo55/ response in different contexts.*

Further analyses only included the ambiguous targets (i.e., Step X-1, Step X, Step X+1), since endpoints stimuli are typically less affected by their context in such normalization designs (presumably due to ceiling effects). Next, the size of the context effect was calculated for each participant and was averaged across three ambiguous targets to represent the overall effect size in a certain context type. A two-way repeated measures ANOVA was conducted on the size of the context effect, with *cue* (lexical tone and vowel) and *context type* (speech and non-speech) as the within-subject factors. The analysis revealed a significant main effect of *cue*, $F(1, 18) = 8.563$, $p<0.05$ and *context type*, $F(1, 18) = 17.803$, $p < 0.05$. The size of the context effect was larger for the perception of tones (M=27.9%, SE=5.3%) than the perception of vowels (M=9.6%, SE=5.9%). Moreover, speech contexts induced normalization (M=34.9%, SE=5%), while non-speech contexts generated no reliable context effect (M=2.6%, SE=6.8%). The *cue* by *context type* interaction was not significant (p=0.087).

Interestingly, a large variation in effect sizes was observed among participants (Figure 2). This variation suggested notable individual differences. The overall performance averaged across subjects (see the black bars in Figure 2) again confirm that non-speech contexts failed to elicit contrastive context effects for the perception of lexical tone (M=5.1%, SE=9.1%) and vowels (M=0.1%, SE=8.8%). Interestingly, however, the figure also suggests that nearly half of the subjects did perceive the lexical tone (11 subjects) or vowel (nine subjects) targets in a contrastive way, even when the context was non-speech. The other half, however, appeared to show an influence of context in an assimilative fashion.

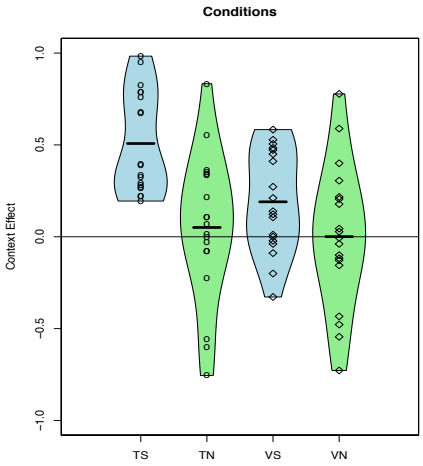

Figure 2: *The context effect size in different contexts. The black bar represents the group mean and each point represents one subject's result. TS stands for the lexical tone speech context, TN for the lexical tone non-speech context, VS for the vowel speech context, and VN for the vowel non-speech context.*

Table 2. *The correlation between the odd and even trials.*

Odd-even pairs	Correlation	p-values
T_S_Odd & T_S_Even	0.971	<0.001
T_NS_Odd & T_NS_Even	0.978	<0.001
V_S_Odd & V_S_Even	0.978	<0.001
V_NS_Odd & V_NS_Even	0.96	<0.001

Notes: T stands for tone, V for vowel, S for speech and NS for non-speech.

To assess whether these contrastive and integrative effects were reliable, we performed a correlation analysis of the by-participant context effect. Each participant's data was separated into two parts based on their odd/even trial numbering. The context effect size was calculated for these separate sets in the four contextual conditions. Pearson's correlation coefficients were calculated for each condition (See table 2). The analyses revealed significant correlations between odd and even trials in all four contextual conditions (p<0.001), suggesting that the contrastive or integrative effect was reliable within an individual.

4. Discussion

4.1. Spectral contrast operates in speech contexts: Support for a partly speech-specific mechanism

To assess whether speech normalization is the result of a contrastive general auditory mechanism, the present study tested the normalization of segmental (F1) and suprasegmental (tone) components in both speech and non-speech contexts. Based on a group average, only speech contexts induced a contrastive context effect for the normalization of both segmental and suprasegmental components. The non-speech contexts were manipulated to have the same LTAS (for the vowel conditions) or the same pitch (through IRN, for the tone conditions). The targets were identical across the conditions. The relation between targets and contexts was thus similar across conditions and could, in principle, have induced contrastive effects in all conditions. Hence, our results are not consistent with a strict contrastive general auditory mechanism as the sole explanation for speech sound normalization [10]. Although the processes of lexical tone and vowel differ in many ways, such as hemispherical lateralization [20], with respect to extrinsic normalization, both lexical tone and vowel normalization exhibited stronger effects in the speech context condition than in the non-speech context condition. This suggests that speech normalization may at least partly operate through speech-specific mechanisms [6].

4.2. A mixed perceptual pattern in non-speech context

The size of normalization effects with non-speech contexts on speech targets has been debated. For example, [8] reported statistically significant contrastive context effects of non-speech contexts, although the effect size was reduced compared to the effects observed with speech contexts. Furthermore, [21] failed to find the contrastive context effect of non-speech contexts altogether. [16] and [17] reported that context effects could even be observed in an assimilative direction. This inconsistency may be partially related to differences in experimental designs across these studies. However, in the present study, i.e., within the same experiment paradigm, participants displayed a mixed pattern for non-speech contexts. Around half of the participants perceived the lexical tones or vowels in a contrastive fashion, whereas, the other half demonstrated an assimilative effect. The considerable between-subject differences might be one of the reasons why the results of previous studies may have been inconsistent. Most studies (e.g. [21], [22]) interpreted their data based on the averaged results across all the participants. It is possible that in some studies, comparatively more participants demonstrated contrastive context effects. Therefore, the average results could demonstrate significant contrastive context effects with non-speech contexts. An important question, then, is why contrastive or assimilative effects may emerge. Further studies are needed to further clarify this question.

5. Conclusion

The results of the present study suggest that spectral contrast between a precursor and target does not always lead to contrastive context effects in perception. Speech context was, in most cases, found to affect listeners' perception of targets in a contrastive way. Non-speech contexts failed to induce contrastive effects on the group level. This observation applied to both the extrinsic normalization of segmental and suprasegmental speech cues. The different results in speech and non-speech contexts suggest that extrinsic normalization may be only partly the result of general auditory processing. In addition, the overall normalization effect in non-speech contexts displayed large individual differences. A number of participants showed contrastive perceptual results, while another group displayed effects in an assimilative way. Further studies need to be carried out to explore these notable individual differences in speech normalization.

6. Acknowledgement

This study was supported by a grant from the Research Grant Council of Hong Kong (GRF: 14408914).

7. References

[1] G. Peng, C. Zhang, H. Zheng, J. W. Minett, N. Territories, and C. Mandarin, "The Effect of Intertalker Variations on Acoustic-Perceptual Mapping in Cantonese," *Hear. Res.*, vol. 55, pp. 579–596, 2012.

[2] G. . Peterson and H. . Barney, "Control Methods Used in a Study of the Vowels," *Joual Acoust. Soc. Am.*, vol. 24, no. 2, pp. 175–184, 1952.

[3] K. Johnson, "Speaker Normalization in Speech Perception," in *The Handbook of Speech Perception*, 2006, pp. 363–389.

[4] H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[5] P. Ladefoged and D. E. Broadbent, "Information Conveyed by Vowels N recent years a great deal of research," *J. Acoust. Soc. Am.*, vol. 29, no. 1, pp. 98–104, 1957.

[6] K. Zhang, X. Wang, and G. Peng, "Normalization of lexical tones and nonlinguistic pitch contours: Implications for speech-specific processing mechanism," *J. Acoust. Soc. Am.*, vol. 141, no. 1, pp. 38–49, 2017.

[7] J. Huang and L. L. Holt, "General perceptual contributions to lexical tone normalization," *J. Acoust. Soc. Am.*, vol. 125, no. 6, pp. 3983–3994, 2009.

[8] M. J. Sjerps, H. Mitterer, and J. M. McQueen, "Constraints on the processes responsible for the extrinsic normalization of vowels," *Attention, Perception, Psychophys.*, vol. 73, no. 4, pp. 1195–1215, 2011.

[9] L. L. Holt, "Temporally nonadjacent nonlinguistic sounds affect speech categorization," *Psychol. Sci.*, vol. 16, no. 4, pp. 305–312, 2005.

[10] A. J. Lotto and L. L. Holt, "Putting phonetic context effects into context: A commentary on Fowler (2006)," *Percept. Psychophys.*, vol. 68, no. 2, pp. 178–183, 2006.

[11] L. L. Holt, A. J. Lotto, and K. R. Kluender, "Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement?," *J. Acoust. Soc. Am.*, vol. 109, no. 2, pp. 764–774, 2001.

[12] C. A. Fowler, "Compensation for coarticulation reflects gesture perception, not spectral contrast," *Percept. Psychophys.*, vol. 68, no. 2, pp. 161–177, 2006.

[13] A. J. Watkins and S. J. Makin, "Effects of spectral contrast on perceptual compensation for spectral-envelope distortion," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3749–3757, 1996.

[14] R. A. Fox and Y.-Y. Qi, "Context effects in the perception of lexical tone," *J. Chinese Linguist.*, vol. 18, no. 2, pp. 261–284, 1990.

[15] A. J. Watkins, "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion [a)]," *J. Acoust. Soc. Am.*, vol. 90, no. 6, pp. 2942–2955, 1991.

[16] M. J. Sjerps, H. Mitterer, and J. M. McQueen, "Hemispheric differences in the effects of context on vowel perception," *Brain Lang.*, vol. 120, no. 3, pp. 401–405, 2012.
[17] H. Mitterer, "On the causes of compensation for coarticulation: evidence for phonological mediation.," *Percept. Psychophys.*, vol. 68, no. 7, pp. 1227–1240, 2006.
[18] M. J. Sjerps, C. Zhang, and G. Peng, "Lexical Tone is Perceived Relative to Locally Surrounding Context, Vowel Quality to Preceding Context," *J. Exp. Psychol. Hum. Percept. Perform.*, 2017.
[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 6.0.16, retrieved 10 August 2016 from http://www.praat.org/," 2016.
[20] Q. Zhang and M. F. Damian, "The time course of segment and tone encoding in Chinese spoken production: an event-related potential study," *Neuroscience*, vol. 163, no. 1, pp. 252–265, 2009.
[21] A. L. Francis, V. Ciocca, N. K. Y. Wong, W. H. Y. Leung, and P. C. Y. Chu, "Extrinsic context affects perceptual normalization of lexical tone," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1712–1726, 2006.
[22] C. Zhang, G. Peng, and W. S.-Y. Wang, "Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones," *J. Acoust. Soc. Am.*, vol. 132, no. 2, pp. 1088–1099, 2012.