The Influence of Sensory Memory and Attention on The Context Effect in Talker Normalization

Guo Li 1, Gang Peng 2,3
1 Department of Linguistics and Modern Languages, CUHK
2 Department of Linguistics and Modern Languages, and Joint Research Centre for Language and Human Complexity, CUHK;
3 Shenzhen Institutes of Advanced Technology, CAS
liguo.vivian@gmail.com, gpengjack@gmail.com

Abstract

Studies on talker normalization have reported that contexts would affect how a vowel/consonant/tone is perceived, which suggests that listeners use cues in the context as a reference for speech perception. However, it is unclear how the cues of the context are encoded in memory, and to what extent the context effects are influenced by interruption in sensory memory and reduction in attentional resources. To fill in this gap, this study examined the effects of noise interruption and a secondary visual task on the identification of words carrying level tones in context by native Cantonese speakers. Experiment 1 compared the tone identification performance between a block where a 300ms noise was presented immediately after the context to interrupt the normalization effects of the context on the following target word and a block without noise interruption. Experiment 2 compared the tone identification performance between a block where participants had to perform a secondary visual task (picture same/different discrimination) and a block without a secondary visual task. Results suggest that context cues are likely encoded both in sensory memory and in short-term categorical memory and that reduction in attentional resources has marginal influence on tone normalization.

Index Terms: tone normalization, attention, memory

1. Introduction

According to [1], [2], talker normalization refers to the process of how listeners recognize words or phonemes despite great inter- and intra-talker variability in the speech they heard. Previous studies on perception of vowels (e.g. [3]), consonants (e.g. [4]) and lexical tones (e.g. [5]) have reported that the perception of a speech sound is influenced by cues in the context.

Tone normalization refers specifically to the process that listeners recognize tones despite great inter- and intra-talker variability in terms of F0 range and F0 slope (see [5]). Previous studies on the three Cantonese level tones (high, mid, and low level tones) [5]-[7] have shown a contrastive context effect, i.e., a target word carrying mid-level tone would be perceived as a word with high-level tone when embedded in a low F0 context, and as a word with low-level tone when embedded in a high F0 context.

Such context effect indicates that tone normalization does not rely solely on the intrinsic F0 information; rather it relies largely on the relative pitch height of the target word against the context (see [5]-[8]). [9] proposed that listeners may construct a reference about a talker’s F0 range based on context and use the reference to determine the intended word by the speaker. However, it is unclear if the reference-construction process involves sensory memory and attention.

To explore the effects of sensory memory and attention in tone normalization, this study compared native Cantonese speakers’ perception of words carrying level tones in situations with and without a noise interruption between the context and the target word, and in situations with and without a secondary picture discrimination task.

Previous studies [5]-[8] and [10]-[13] disagree as to whether tone normalization is a general auditory process or a speech specific process. By including a novel type of context—reversed speech, which, like nonspeech, contained no semantic meaning but had the same spectral information as speech context, this study will provide further information for the debate.

This study examines if the effects of sensory memory and attention for tone normalization would differ for nonspeech, speech and reversed speech. If sensory memory and attention have an impact on tone normalization, the impact is expected to be greatest for speech stimuli, less for reversed speech, and least for nonspeech.

2. Experiments

2.1. Experiment 1: sensory memory

This experiment aimed to examine how sensory memory may affect context effects on tone normalization. It contrasted tone normalization between a block where a 300ms noise was presented immediately after the context to interrupt the normalization effects of the context on the following target word and a block without noise interruption.

The 300ms noise interruption was designed to refresh the sensory memory. [14] discussed two types of sensory memory: unanalyzed raw sensory data, and analyzed sensory memory which may encode phonetic information. The lifetime of the first type of sensory memory is around 300ms but the lifetime of the second type of sensory memory may last for seconds. [14] hypothesized that perceptual categorization (e.g. mapping a tone of given pitch height to a tone category) takes place in short-term memory.

If contextual cues for tone normalization are encoded in sensory memory, context effects are expected to drop in situations with a noise interruption, and reversed speech is expected to contribute to tone normalization in a way similar to speech. If contextual cues are encoded only in short-term memory, context effects is expected not to differ much between situations with and without a noise interruption, and reversed speech is expected to contribute to tone normalization in a way similar to nonspeech.
2.1.1. Participants
A group of 16 right-handed native speakers of Hong Kong Cantonese from the Chinese University of Hong Kong (8 female, 8 male; mean age=21 years, SD=2, aged 19-26) were paid to participate in the experiments. All participants had normal hearing and none was from music or linguistic majors. The experimental procedures were approved by the Survey and Behavioral Research Ethics Committee of The Chinese University of Hong Kong. Informed written consent was obtained from each participant in compliance with the experiment protocol.

2.1.2. Stimuli and experiment design
Stimuli and experiment design largely followed [7] and [8]. Two female and two male native speakers of Hong Kong Cantonese with varying pitch ranges produced the target word (意/ji33/ ‘meaning’) and the carrier sentence (i.e., the context stimuli) (呢個字係…jii55 ko33 tsi22 ha22 ‘This word is…’). Additionally, one female and one male produced the filler carrier sentence 請留心聽…/vi25 lou21 sm55 ti55 / ‘Please carefully listen to …’. All speakers were asked to produce six recordings and the one whose mean F0 closest to the average of recordings of all six times was selected for use.

Tokens of the target word (意/ji33/ ‘meaning’) from the four speakers were normalized to 450ms in duration and 55dB in peak intensity in Praat. Tokens of the carrier sentences were normalized to 1000ms in duration and 55dB in peak intensity.

A triangle wave was used to synthesize nonspeech stimuli. Each speech stimuli (including fillers) had a corresponding nonspeech stimuli matched in terms of F0 trajectory. The average intensity of nonspeech stimuli was set to 75dB to match with the loudness level of target. Reversed speech stimuli were synthesized by reversing the speech stimuli in Praat (both segmental and F0 contour went through reversion).

All three kinds of context stimuli (Speech, Nonspeech, Reversed speech) were then adjusted in overall F0 to generate three types of F0 range. The original stimuli composed the contexts with ‘unshifted’ F0 range. Stimuli generated by raising the original overall F0 trajectory by three semitones composed the contexts with ‘raised’ F0 range. Stimuli generated by lowering the original overall F0 trajectory by three semitones composed the contexts with ‘lowered’ F0 range. Target stimuli with raised or lowered overall F0 range were synthesized to be used with filler contexts. Context stimuli and target stimuli in each trial in the experiment came from the same speaker.

Altogether, there were 16 context stimuli (3 F0 ranges × 4 speaker × 1 filler × 4 speaker) for each context type (Nonspeech, speech, reversed speech). For all the test stimuli, target stimuli with unshifted F0 range were used. For all filler stimuli, target stimuli with raised or lowered F0 range were used.

2.1.3. Procedure
The experiment consisted of two blocks: the block with interruption had noise presented between context and the target stimuli; the block without interruption had no such interruption between the context and the target. Presentation order for the two blocks were counterbalanced between participants. Within each block, the 48 stimuli (3 context types × 3 F0 ranges × 4 speakers + 3 context types × 1 filler × 4 speakers) were presented in random order for 7 times.

All stimuli were presented binaurally to the participants via a Sennheiser HD 380 pro headphone. All participants were seated in a quiet enclosure and were given a written instruction before the experiment. They were instructed to pay close attention to everything they heard and identify the target word at the end of each trial as any of the three Cantonese words, 医 (ji55/ ‘doctor’), 意 (ji33/ ‘meaning’), or 二 (ji22/ ‘two’). Before each experiment block, each participant would run a practice block to get familiar with the procedure. Stimuli used in the practice block were recorded from one additional female speaker and one additional male speaker and were not used during test blocks.

Participants’ identification responses were recorded. In the analysis, the responses were transcribed into numerical values to reflect perceptual height. Specifically, 医 (ji55/ ‘doctor’), responses were transcribed as 6, indicating that the target was perceived to have high F0. Similarly, 意 (ji33/ ‘meaning’) responses were transcribed as 3 and 二 (ji22/ ‘two’) responses were transcribed as 1.

2.1.4. Results
Figure 1 shows the average perceptual height in each F0 range condition (Raised, Unshifted, Lowered) within each of the three context conditions (Nonspeech, Reversed speech and Speech) during the no interruption block. Figure 2 shows the average perceptual height during the block with interruption. A three-way repeated measures ANOVA was conducted on the perceptual height of three context conditions. Block type (Silence, i.e. no interruption, Noise, i.e. with interruption), Context type (Nonspeech, Speech and Reversed speech), F0 shift (Raised, Unshifted, and Lowered) were three within-subjects factors. Greenhouse–Geisser method was used to correct violations of sphericity where appropriate.

There were significant main effects of context type, F(2, 30) = 13.95, p<0.001 and F0 shift, F(2, 30) = 56.83, p<0.001. Moreover, there was a significant two-way interaction between context type and F0 shift, F(4, 60) = 66.40, p<0.001 and a three-way interaction between block type, context type and F0 shift, F(4, 60) = 4.09, p<0.05. These results indicated that context type and F0 shift had significant effects on modulating tone normalization while in general the presence/absence of a noise interruption did not affect tone normalization.

The three-way interaction effect was broken down to two two-way (Context type and F0 shift) repeated measures ANOVAs for each block separately. As for the silence block, main effects of Context type (F(2, 30)=10.27, p<0.001), F0 shift (F(2,30)=63.0, p<0.001) and a Context type by F0 shift interaction effect (F(4,60)=65.49, p<0.001) were observed.
Simple main effect analyses of F0 shift were conducted with Bonferroni adjustment. For Nonspeech context, perceptual heights in each F0 shift condition did not differ significantly. For Speech context, perceptual height was significantly higher in lowered F0 condition ($M = 4.56, SE = 0.20$) than in unshifted F0 condition ($M = 3.09, SE = 0.13$), $p < 0.001$, which was in turn significantly higher than in raised F0 condition ($M = 1.67, SE = 0.13$), $p < 0.001$. For Reversed speech context, perceptual height in raised F0 condition ($M = 2.27, SE = 0.15$) was significantly lower than in unshifted F0 condition ($M = 2.96, SE = 0.19$), $p < 0.001$, and in lowered F0 condition ($M = 3.29, SE = 0.25$), $p < 0.005$, but perceptual heights did not differ significantly between lowered and unshifted F0 conditions.

The simple main effects of Context type were further analyzed with Bonferroni adjustment. Only for Speech context, the total shift value in Silence block ($M = 2.89, SE = 0.22$) was significantly larger than that in Noise block ($M = 2.29, SE = 0.31$), $p < 0.01$. No significant difference was observed for Nonspeech or Reversed speech.

2.2. Experiment 2: attention
The purpose of this experiment was to examine the effects of attention. It contrasted tone normalization between a block where participants had to perform a secondary visual task (picture same/different discrimination) and a block without a secondary visual task.

The picture same/different discrimination task was designed to compete with tone normalization for attention. As in situations with the picture recalling task, two pictures were presented consecutively during the presentation of context stimuli. Listeners had to remember the two pictures and decide if they were the same or not while listening to the context. They had to store the pictures or their decision in memory when they listened to the target word and performed tone normalization. If tone normalization relies heavily on attention, context effects are expected to drop in the situation with the secondary task. If tone normalization is largely autonomous, context effects are expected not to differ across situations with and without the secondary visual task.

2.2.1. Participants
Another group of 16 right-handed native speakers of Hong Kong Cantonese from the Chinese University of Hong Kong (8 female, 8 male; mean age=20.75 years, SD=1.06, aged 19-22) were paid to participate in the experiments. All participants had normal hearing and none was from music or linguistic majors. The experimental procedures were approved by the Survey and Behavioral Research Ethics Committee of The Chinese University of Hong Kong. Informed written consent was obtained from each participant in compliance with the experiment protocol.

2.2.2. Stimuli and experiment design
Audio stimuli used in this experiment were the same as in the previous experiment.

Visual stimuli for the picture recall task were 10 pictures on each of which 14 white squares were organized in various fashions against a black background.

2.2.3. Procedure
The experiment consisted of two blocks: the full attention block where participants were instructed to pay full attention to audio stimuli; the attention-deprived block where participants were instructed to perform a secondary task, picture same/different discrimination, while listening to the context stimuli.

Procedure for the full attention block was the same as that for the no interruption block in Experiment 1. Procedure for the attention-deprived block differed from the full-attention block only in two aspects. First, while the context was presented, two random pictures were presented for 300ms each consecutively. Half of the time the two pictures would be the
same. Second, after the participant identified the target word by pressing a key or after 3000ms, a slide would ask the participant to judge if the two pictures presented during the context were the same or not by pressing designated keys. Participants were given maximally 2000ms to perform this picture discrimination task.

As with Experiment 1, a practice block was run before each test block. Participants’ responses were recorded and transcribed into numerical values also as in Experiment 1.

2.2.4. Results

Figure 3 shows the average perceptual height in each F0 range condition (Raised, Unshifted, Lowered) within each of the three context conditions (Nonspeech, Reversed speech and Speech) during the full attention block. Figure 4 shows the average perceptual height during the attention-deprived block.

The following analyses were conducted only for the block with full attention. A two-way repeated measures ANOVA was conducted on the perceptual with Context type and F0 shift as within-subjects factors. There were significant main effects of Context type, $F(2,30) = 8.39, p<0.005$ and F0 shift, $F(2,30) = 56.24, p<0.001$. Moreover, there was a significant two-way interaction between Context type by F0 shift, $F(4,60) = 51.55, p<0.001$.

Simple main effect analyses of Context type were conducted on Bonferroni adjustment. For Lowered F0 condition, Nonspeech context, perceptual heights in each F0 shift condition did not differ significantly. For Speech context, perceptual height was significantly higher in lowered F0 condition ($M = 4.31, SE = 0.25$) than in unshifted F0 condition ($M = 2.98, SE = 0.18$), which was in turn significantly higher than in raised F0 condition ($M = 1.71, SE = 0.19$), $p<0.001$. For Reversed speech context, perceptual height was significantly higher in lowered F0 condition ($M = 3.07, SE = 0.24$) than in unshifted F0 condition ($M = 2.74, SE = 0.22$), $p<0.05$, which was in turn significantly higher than in raised F0 condition ($M = 2.37, SE = 0.21$), $p<0.01$.

3. Discussion

This study replicated findings in [5]-[8]. In all test blocks, Nonspeech context induced little context effects for tone normalization while Speech context induced significant context effects. Reversed speech did not have semantic meaning but preserved at least partially segmental information [15]. Reversed speech showed a similar pattern as speech context but the context effects were largely reduced, implying that segmental information and semantic meaning may facilitate tone normalization, and that contextual cues for tone normalization may be encoded in both sensory memory and short-term categorization memory. These results supported the view that tone normalization is mostly a speech-specific process and the contribution from general auditory processing is limited.

A significant difference between blocks with and without interruption in Experiment 1 was observed only when the context type is Speech, which suggested listeners were more sensitive to speech contexts than nonspeech contexts. Hence the effects of noise interruption on tone normalization are limited. Also, it further supports the claim that contributions to tone normalization by speech and nonspeech context are unequal. As no significant difference was found between blocks with and without a secondary visual task, the effects of attention reduction on tone normalization are limited. However, given that the secondary task was a visual task, results may differ if it had been an auditory one.

Taken together, this study showed that tone normalization in general was not significantly affected by the presence of a noise interruption between the context and the target word or when full attention was not available, suggesting that contextual cues for tone normalization may be encoded in both sensory memory and short-term categorization memory and that tone normalization might be a highly autonomous process.

4. Acknowledgements

The work described in this paper was partially supported by a grant from National Natural Science Foundation of China (NSFC: 61135003) and a grant from the Research Grant Council of Hong Kong (GRF: 448413). We thank Caicai Zhang for insightful discussions and kind suggestions.
5. References


