Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context

Matthias J. Sjerps[a,b], Caicai Zhang[c,d], Gang Peng[c,d1]

[a] *Neurobiology of Language Department, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands*

[b] *Department of Linguistics, University of California, Berkeley, CA, USA*

[c] *Department of Chinese and Bilingual Studies, the Hong Kong Polytechnic University, Hong Kong*

[d] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*

[1] Corresponding author.

E-mail address: gpengjack@gmail.com

Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context

Abstract (max 200 words)

Important speech cues such as lexical tone and vowel quality are perceptually contrasted to the distribution of those same cues in surrounding contexts. However, it is unclear whether preceding and following contexts have similar influences, and to what extent those influences are modulated by the auditory history of previous trials. To investigate this, Cantonese participants labeled sounds from 1) a tone continuum (mid- to high-level), presented with a context that had raised or lowered F0 values and 2) a vowel quality continuum (/u/ to /o/), where the context had raised or lowered F1 values. Contexts with high or low F0/F1 were presented in separate blocks or intermixed in one block. Contexts were presented following (Experiment 1) or preceding the target continuum (Experiment 2). Contrastive effects were found for both tone and vowel quality (e.g., decreased F0 values in contexts lead to more high tone target judgements and vice versa). Importantly, however, lexical tone was only influenced by F0 in immediately preceding and following contexts. Vowel quality was only influenced by the F1 in preceding contexts, but this extended to contexts from preceding trials. Contextual influences on tone and vowel quality are qualitatively different, which has important implications for understanding the mechanism of context effects in speech perception.

Keywords

Speech perception; lexical tone; vowel quality; normalization; context effects

Public significance statement

Speech perception is highly context dependent. This study investigates the extent of contextual influences on the perception of lexical tone and vowel quality in a number of ways. Perception of lexical tone was found to be influenced by locally preceding and following contexts, while vowel quality was only

influenced by the preceding context, but that context extends further in time. These patterns demonstrate that the temporal scope of perceptual influences of context is speech cue specific.

Introduction

Human languages rely on a multitude of acoustic cues to express differences in lexical meaning. Speech elements, for example, can distinguish lexical meaning through differences in spectral quality, pitch, and duration. For each of these cues their perception is highly dependent on acoustic-phonetic properties of the context in which they appear (e.g., Francis, Ciocca, Wong, Leung & Chu, 2006; Ladefoged & Broadbent, 1957; Newman & Sawusch, 2009; Reinisch & Sjerps, 2013; Zhang, Peng & Wang, 2012, 2013). Such influences are in many ways similar to contextual influences in other perceptual domains. That is, the same lukewarm bath feels hot when you have just walked in the snow; but may feel rather cold when you just spent 15 minutes in a sauna (see e.g., Kluender, Coady & Kiefte for further discussion of the generality of such perceptual effects). Importantly, in the domain of speech perception, such contrastive processes aid listeners in resolving variability in how speech sounds are realized by different speakers. That is, they help to "normalize" different speakers' utterances. However, it remains largely unclear what the "scope" of such normalization processes is. Such scope concerns two aspects: 1) whether immediately preceding and following contexts affect perception similarly; and 2) whether speech sounds in further preceding phrases (or trials, in an experimental setting) also induce contrastive effects. Determining the scope of normalization is relevant for multiple speech cues, and one could thus ask to what extent the scope of two cues may or may not differ. The current study simultaneously investigated the normalization of two important distinctive cues in Cantonese, lexical tone and vowel quality, to address these issues.

The realization of F0 and formant values is highly variable even within different realizations of the same speech sounds. Important sources of this variability are the between-speaker differences related to the length of the vocal tract and vocal folds. Additional influences may arise through influences of

locally surrounding speech sounds as demonstrated in coarticulation (e.g., Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967; Peterson & Barney, 1952; Xu, 1997). Moreover, even repetitions of the same words by the same speaker may differ in the exact F0 and formant values. Such values change across the course of the day and may be influenced by factors such as affective states, speaking style and intended interlocutor (Garrett & Healey, 1987; Heald & Nusbaum, 2015; Protopapas & Lieberman, 1997; Johnson, Flemming & Wright, 1993; Kuhl, et al., 1997; Smiljanić & Bradlow 2005; Picheny, Durlach & Braida, 1986, 1989; Krause & Braida, 2004). Importantly, however, these combined influences typically affect the realization of surrounding speech sounds in a similar, predictable, way. That is, a speaker with a high F1 and/or F0 in the production of the vowel /u/ in a word like "boot" will typically also produce a relatively high F1 and/or F0 for the production of the vowel /æ/ "bat" and /i/ in "beet". And this is especially so in adjacent speech. Hence, when interpreting a given speech sound, the acoustic-phonetic properties of its context can provide a listener with very useful information.

Indeed, the perception of vowel quality is strongly influenced its context. As one of the earliest demonstrations of this effect, Ladefoged and Broadbent (1957) asked participants to listen to repetitions of words that were synthesized to sound like "bit" and "bet" and asked them to indicate what they heard. Critically, these words were preceded by utterances in which the F1 was either generally shifted to lower frequencies, or shifted to higher frequencies. Results demonstrated that after a high F1 precursor, listeners gave more responses of the vowel with low F1 (i.e., /ɪ/ or "bit") than after a low F1 precursor. That is, in a high-F1 context they seem to have "shifted" their category boundary towards the higher-F1 vowel /ɛ/ (thus more /ɪ/ responses). Similar effects have since been reported with various experimental designs (e.g., Assgari & Stilp, 2015; Johnson, Strand & D'Imperio, 1999; Mitterer, 2006; Nearey, 1989; Sjerps, Mitterer & McQueen, 2012; 2013; Sjerps & Smiljanić, 2013; Stilp, Anderson, & Winn 2015; Watkins, 1991; Watkins & Makin, 1994). In addition, these influences are at least partly driven by general principles of auditory contrast enhancement (Kluender, Coady & Kiefte, 2003; Sjerps et al., 2011, 2012; Stilp, Alexander, Kiefte & Kluender, 2010; Watkins, 1991; Watkins & Makin 1996).

In line with the suggested general nature of such effects, the same type of context dependent shifts in perception occur for lexical tone (e.g., Francis et al., 2006; Huang & Holt, 2009; 2011; Zhang, et al., 2012, 2013, Zhang & Chen, 2016). Zhang and Chen (2016), for example, asked participants to identify Cantonese words that are minimally distinguished based on tone. When these words were presented in the context of a sentence with a raised overall F0, participants gave more low-level tone responses (i.e., shifting the perceptual boundary of the target words toward the high-level tones). In contrast, lowering the overall F0 in the same context resulted in more high-level tone responses. Although the same contrastive effect also occurs in non-linguistic pitch perception, those effects are notably smaller than those with speech (Zhang, Wang & Peng, 2017; but see Huang and Holt, 2009).

These combined results demonstrate that normalization is not specific to any particular speech cue. Instead normalization seems to be the result of a general perceptual tendency to perceive cues relative to their local context. Indeed, context also influences the perception of the difference between long and short vowels in Dutch (Reinisch & Sjerps, 2013), the perception of VOT in English (e.g., Newman & Sawusch, 2009; Toscano & McMurray, 2015), and a number of other important speech cues (see e.g., Holt, 2005; Kiefte & Kluender, 2008; Miller & Liberman 1979; Stilp & Assgari, 2017 for detail). Given the fact that the interpretation of most -if not all- speech cues is highly context dependent, the question arises to what extent such normalization processes operate in a qualitatively similar way. One way to address this question is to investigate whether different cues are normalized over a similar temporal scope.

Although research on these specific questions is scarce, there exists some evidence that lexical tone and vowel quality may be normalized over a rather different temporal scope. For example, that normalization for lexical tone seems to occurs mostly over local context (Wong and Diehl, 2003, Zhang et al 2012). For the perception of formant-based contrasts, on the other hand, effects induced by immediate and more distal context are rather similar (e.g., Holt, 2006). Furthermore, for tone normalization context effects occur with both following and with preceding contexts (Francis et al., 2006; Lin & Wang, 1984;

Leather, 1983; Moore & Jongman, 1997; Huang & Holt, 2009, 2011; Zhang et al. 2013; Zhang & Chen, 2016). For normalization of vowel quality, on the other hand, results seem inconclusive or even contradictory (e.g., van Bergem, Pols & Koopmans-van Beinum, 1988; Johnson & Strange, 1982). The combined reports described above thus suggest that there may be interesting differences in the normalization for tone and vowel quality. However, they are based on separate reports and the designs and stimuli of these experiments were often very different, precluding any strong claims so far.

The current study compared the temporal scope of both normalization effects in a group of Cantonese listeners. We adopted a *speech cue* (tone vs. vowel quality) × *presentation mode* (context F0/F1 was blocked vs. intermixed across trials) × *context position* (preceding context vs. following context) design. Target stimulus continua for tone and vowel quality were always presented in separate blocks. In the tone materials the targets consisted of an instance of /fo/ that carried a tone continuum ranging between mid-level and high-level tone in Cantonese. In the vowel quality materials the vowels ranged between the vowels /o/ and /u/, while the F0 was fixed to indicate a high-level tone. A disyllabic context (/p^ha21 tsi25/)[1] was manipulated to reflect someone speaking with either a high F0 or a low F0 (for the tone materials), or to reflect a speaker with either a high F1 or a low F1 (for the vowel quality materials). Based on previous research the context manipulations were expected to induce normalization effects.

The stimuli with high and low contextual manipulations were either intermixed within the same block (i.e., where a trial with a high F0/F1 context could be followed by a trial with a low F0/F1 context), or presented in separate blocks (i.e., a fully context-immersed presentation mode). This should allow for the estimation of whether contexts in further preceding trials affect target judgements. In a blocked context, subsequent trials have the same context type, and as such normalization effects should be additive. In a mixed presentation mode, contexts of adjacent trials are equally likely to be either opposing or additive. Hence, if further preceding trials have no influence on target judgement on the current trial, then blocked and mixed presentation modes should result in the same effect sizes. Furthermore, the

context sequences /p^{h}a21 tsi25/ were either presented right after the target (Experiment 1; /fV p^{h}a21 tsi25/) or right before the target (Experiment 2; /p^{h}a21 tsi25 fV/). In both experiments, participants could only respond once the entire stimulus had been presented (i.e., both target and context). This approach ensured that the context (preceding or following) could be integrated with the target in all conditions. As discussed above, based on previous reports we expected that the temporal scope of normalization for vowel quality and tone may be different.

Experiment 1: Speaker normalization with following context

Method

Participants Eighteen native speakers of Cantonese (nine female; all right handed; average age of 21.7 years) were recruited among students at the Chinese University of Hong Kong. They received a monetary reward for their participation. None of the participants reported a language impairment, hearing disorder, or uncorrected visual impairment. Informed written consent was obtained from each participant in compliance with the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee. Data from two participants were discarded due to non-categorical responses for the endpoint stimuli (see below).

Power analysis As described above, the existing literature investigating normalization of Tone and Vowel Quality, and especially those investigating their scope, have resulted in rather mixed results. Because of this variability, we did not have specific predictions about effect sizes, and hence, the design should be considered *exploratory*. Importantly, however, the experiments hinge on the ability to detect an effect of Context (high vs Low F0/F1) along with differences in the size of a Context effect between at least two levels of another condition. Hence, the power analysis focuses on the number of participants that are needed for enough power to detect a main effect and such interactions. To calculate power, simulations were performed based on the data of 72 participants reported in Sjerps and Smiljanić (2013; data available on osf.io/yf2pq). That study was in a number of ways similar to the current design,

(although it only tested effects of vowel quality, it did rely on an /u/-/o/ target continuum, and investigated the influence of increased or decreased F1 values in contexts).

To simulate a main effect for Context and an interaction between the factor Context with another factor we used the Sjerps and Smiljanić's dataset and semi-randomly assigned a condition level to each observation (e.g., Vowel Quality versus Tone; with balanced number of trials). Then, for one of the conditions (say, Tone), for each participant, and among all trials that *only* differed with respect to their level of Context, all responses were shuffled (to simulate a null-effect for the factor Context in that condition). Simulations were performed for increasing numbers (*n*, ranging from 5:16) of (randomly drawn) participants from the Sjerps and Smiljanić's dataset. For each *n*, 1000 simulations were carried out. For each such simulation, a generalized linear mixed effect model was fitted to the existing data with the same contrast coding scheme as used in the current study (see results section for detail on predictors and their coding). Hypothetical responses for our current design matrix were predicted from the models (using the simulate() function in the stats package in R [R Core Team, 2014]). The predicted data was then analyzed and the significance (with $\alpha = 0.05$, two tailed) of the interaction between the factors Context and a second factor was recorded for each simulation. The proportion of significant tests across each *n* was calculated. The results of these simulations indicated that our design led to enough power ($(1 - \beta) > 0.90$ with ≥ 10 participants) to detect both a main effect of Context and an interaction involving Context.

Materials Details about the construction of the speech materials can be found in Online Supplement A. The procedures for manipulations were similar to those applied in Sjerps and Smiljanić (2013; for Vowel Quality), and to those in Zhang et al., (2013; for Tone). Figure 1 provides a visualization of the synthesis parameters that were applied to create an F0 (for the tone materials) and an F1 (for the vowel quality materials) continuum, and to create context syllables that had increased or decreased F0 (tone) or an increased or decreased F1 (vowel quality).

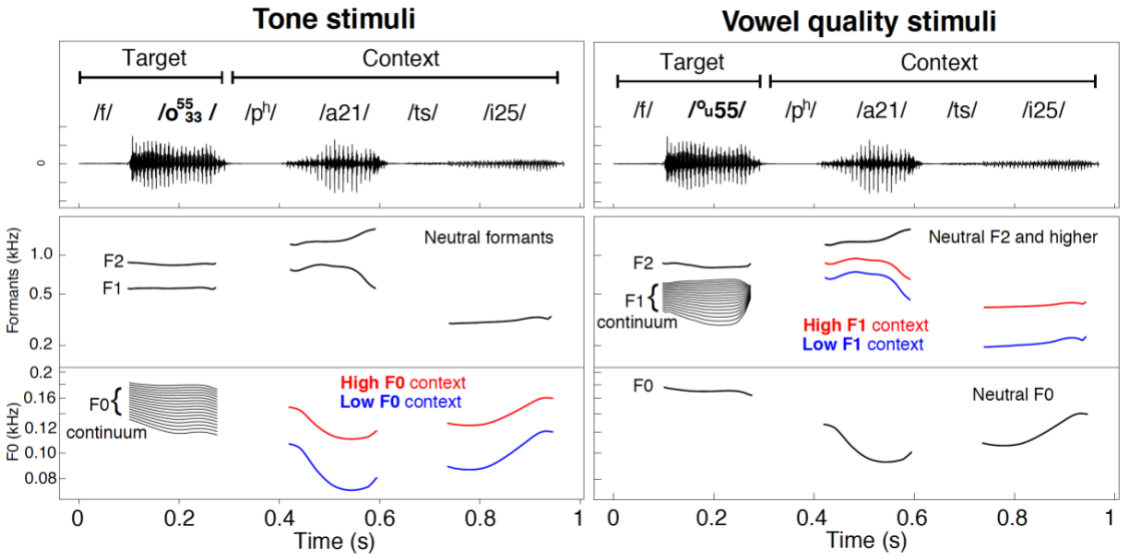

Figure 1. Parameters for the synthesis of the tone and vowel quality stimuli. Top left and right: Annotated waveform of an example stimulus for visualization of temporal stimulus characteristics. Middle left: in the tone materials, formant frequencies are identical (at a neutral value) across the two speaker conditions. Bottom left: an F0 continuum is created for the target vowels. The F0 of the context vowels is shifted up (high-F0 speaker condition) or shifted down (low-F0 speaker condition). Middle right: for the vowel quality materials, an F1 continuum is created for the target vowels. The F1 of the context vowels is shifted up (high-F1 speaker condition) or shifted down (low-F1 speaker condition). Formants higher than F1 and F0 (latter in bottom right panel) are left identical (at a neutral value) across speaker conditions. For the middle and bottom panels, y-axes are logarithmically scaled.

Procedure The experiment consisted of two phases. Phase 1 was used to select participant-specific tokens that would span across an ambiguous range, to be used in phase 2. For phase 1 participants were first presented with 6 steps from the vowel quality continuum (steps 1, 4, 7, 10, 13, 16) presented in a neutral F1 context. Next, participants were presented with 6 steps from the tone continuum (steps 1, 4, 7, 10, 13, 16) presented in a neutral F0 context. The task was to identify each stimulus as /o/ or /u/ for the vowel quality continuum, and as /55/ or /33/ for the tone continuum. For both pretests, each step was

repeated 10 times. Only a subset of the target ranges was used for phase 1 in the interest of testing time and participant fatigue. The on-site experimenter then used the categorization results to identify a set of stimuli that would be likely to result in ambiguous performance on the midpoints of the continuum.

In phase 2, participants were presented with five steps from the vowel quality and tone continua. For each participant, these five individually-selected steps (as determined in phase 1) consisted of two endpoint tokens that had been reliably categorized as one or the other target sounds by the participant in phase 1, and three consecutive tokens in the ambiguous region (the latter three were adjacent in terms of the 16 resynthesis steps). Context effects are often strongest for the most ambiguous items, and this approach ensured a dense sampling of the ambiguous region, while the endpoints allowed us to judge whether the participants displayed reliable categorization. Appendix B reports the across-participant average F0 and F1 values of the steps used in phase 2 (along with their variability). The endpoint stimuli were presented 20 times, and the ambiguous items were presented 40 times, generating a total of 160 items. The tone and vowel quality stimuli were presented in separate blocks. For each type of cue, the stimuli were presented once in a blocked design (i.e., high and low contexts were presented in separate blocks), and once in a mixed design (i.e., high and low contexts were intermixed in a single block). In total, phase 2 consisted of the following 6 sub-parts, with the total number of trials per block type indicated in brackets: 1) Tone - Low F0 context (160); 2) Tone - High F0 context (160); 3) Vowel quality - Low F1 context (160); 4) Vowel quality - High F1 context (160); 5) Tone - Low & High F0 context (320); 6) Vowel quality - Low & High F1 context (320). While the first four blocks were blocked conditions, the remaining two blocks were mixed conditions. The orders of the conditions were counterbalanced across participants. The orders of the trials within blocks were randomized. At the start of each block participants were told what button to press for the different response options (e.g., for the tone blocks they would see: press left arrow when you hear "fo33 (課)" as the first of the three sounds, and press right

arrow when you hear "fo55 (科)" as the first of the three sounds [for vowel quality blocks this would be left arrow for "fu55 (呼)" and right arrow for "fo55 (科)"]. After presenting a stimulus a question mark appeared on screen, prompting the response. Participants could only respond after the complete stimulus was presented (i.e., after both target and context had been heard). After the offset of an audio stimulus, there was a period of 1500 ms for responding. Responses being earlier than the offset or after the response period were recorded as "error" and omitted from further analyses. Participants could take a self paused break after every 80 trials (and between the blocks).

Results

Data from individual participants were removed from an experimental block if the participant did not demonstrate reliable categorization (i.e., no numerical increase in /fo55/ responses between step 1 and 5 on the continua, unless at floor/ceiling). For two participants all data were discarded because the majority of their conditions did not conform to this criterion. For the remaining data, the number of participants contributing to any individual block type was minimally 14 participants and maximally 16 participants. Figure 2 displays the overall categorization behavior across the conditions. Context effects are revealed as a separation between the high-F0/F1 and the low-F0/F1 lines. The separate panels display the categorization behavior in the different conditions (see the labels or legend of Fig. 2). The results suggest that with a following context, the normalization effect with respect to tone appears to be present in both the mixed and the blocked presentation conditions. For the normalization of vowel quality with following context, however, only an effect seems to be present in the blocked presentation condition.

To statistically assess these patterns, the results were analyzed using generalized linear mixed effects models in R (version 3.1.1; The R foundation for statistical computing) as provided in the lme4 package

(Bates Mächler, Bolker, & Walker, 2014). For the dichotomous dependent variable of categorization responses (i.e., /fo55/ = 1 vs. /fo33/ [in the tone experiment] or /fu55/ [in the vowel quality experiment] = 0), a logit linking function was used. All fixed factors were centered around zero. These were Step (with the levels −2, −1, 0, 1, 2) reflecting the step on the F0 or F1 continua in the target syllable; Context (with the levels Low−context = −1 vs. High−context = 1), indicating the high/low F0 or F1 manipulation in the context; Cue (with the levels Tone = −1 vs. Vowel quality = 1), indicating what the target continuum was; Mode (with the levels Blocked = −1 vs. Mixed = 1), indicating whether the stimuli were presented in a blocked manner or a mixed manner. A full structure of (uncorrelated) random effects was included (Barr, Levy, Scheepers & Tily, 2013). As the number of possible interactions in this four-way design was large we adopted a directed analysis approach testing only specifically for the three-way interaction between Context, Cue and Mode, along with its lower-level interactions and main effects. Only the two-way interactions involving Step were included, as we had no a-priori expectations about its involvement in the three or four way interactions. Only significant effects are reported below.

The analysis revealed a main effect of Step (B = 1.16, z = 12.49, $p < 0.001$), reflecting the observation that, overall, more /fo55/ responses were given for stimuli on the /fo55/ end of the continuum. In addition, an effect of Context was observed (B = -1.07, z = -5.27, $p < 0.001$) as more /fo55/ responses were given for stimuli with the low-F0/F1 context. A significant effect was observed for the factor Mode (B = -0.21, z = -2.26, p = 0.02) as there was a tendency for overall fewer /fo55/ responses in the mixed presentation mode than in the blocked mode (we have no clear interpretation of this observation, but it is orthogonal to the effects of interest). A significant interaction was observed between the factors Step and

Mode (B = -0.26, z = -6.26, p < 0.001) as the slopes of the categorization curves were shallower in the mixed presentation condition. An interaction was observed between the factors Context and Mode (B = 0.50, z = 3.05, p = 0.002) as the overall effect of Context was reduced in the mixed presentation condition. An interaction was also observed between the factors Context and Cue (B = 0.64, z = 4.35, p < 0.001) as the effect of Context was larger in the Tone materials than in the Vowel quality materials.

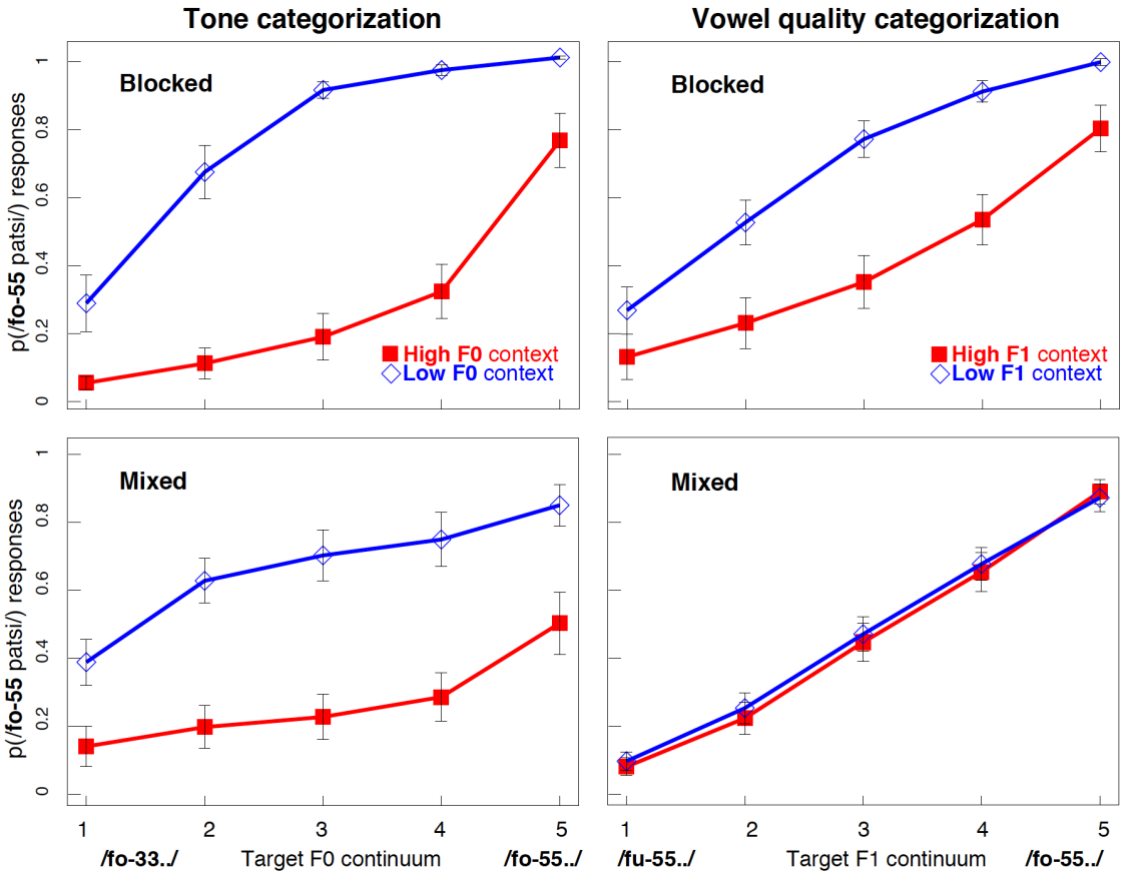

Figure 2. Experiment 1 categorization behavior with following context. The four panels represent the data split over two factors: 1) Tone continuum stimuli (left hand panels) versus Vowel quality stimuli (right hand panels); 2) Blocked presentation (top panels) or mixed presentation (bottom panels). Within each panel, proportions of /fo55../ responses are displayed for the five stimulus steps, split out by context condition. Target sounds presented in the context of High F0 and high F1 properties are displayed in red. Target sounds presented in the context of Low F0 and Low F1 properties are displayed in blue. Error bars reflect standard errors of the mean. See Figure 4 for visualization of by-participant effects of Context.

The primary difference between blocked and mixed presentation modes is that in mixed presentation, contexts on *previous trials* were sometimes of the same condition (i.e., high or low F0/F1) as the current trial and sometimes from a different condition. To probe the influence of previous trials, as a final analysis, we investigated to what extent the following context part of a previous trial affected categorization. The same model approach (including Step, Context, and Cue) was taken as above, except that we only included data from the mixed condition (as subsequent trials were always of the same speaker condition in the blocked condition). In addition, a factor Previous trial−context was added (with the levels Low previous trial−context = −1 vs. High previous trial−context = 1). The analysis revealed a main effect on the Intercept (B = -0.30, z = -2.50, p = 0.01). As in the analysis described above, significant main effects were observed for Step (B = 0.94, z = 10.03, p < 0.001) and (following) Context (B = -0.79, z = -3.87, p < 0.001), along with the interaction between Context and Cue (B = 0.75, z = 4.07, p < 0.001). More interestingly, a trend towards a main effect was also observed for Previous trial-context (B = -0.10, z = -1.80, p = 0.07), indicating that speaker information of a previous trial may have a small effect on the interpretation of the target sound on subsequent trials. Importantly, in addition, an interaction was observed between the factors Previous trial-context and Cue (B = -0.10, z = -2.13, p = 0.03), reflecting the fact that the effect of previous trial-context was larger for Vowel quality than for Tone.

Discussion

The results of this experiment demonstrated that context effects on the perception of lexical tone and those on vowel quality were qualitatively different. That is, for tone stimuli we observed large normalization effects in both the blocked and in the mixed presentation conditions. For vowel quality, on the other hand, effects were only observed in the blocked presentation mode. These results suggested that

for vowel quality, normalization in the blocked condition must have been a result of the contexts that were presented on previous trials. Indeed, a subsequent analysis demonstrated that the context in the previous trials had a reliable effect on categorization. This was not the case for lexical tone.

These results are thus generally in line with previous observations described in the Introduction. In other words, normalization for tone and for vowel quality may operate over a qualitatively different scope. However, Experiment 1 only consisted of a following context. Most previous normalization experiments used preceding contexts. As such it may be that for preceding contexts the scope of normalization for tone and vowel quality is more similar. Furthermore, for vowel quality normalization, there was a reliable effect of the context of previous trials. This also invites the investigation of the effect of preceding contexts on the current trial, and whether it differs between vowel quality and tone normalization. To further investigate this issue we carried out a second Experiment. For Experiment 2 we used the same context as for Experiment 1, but this context was now prepended to the target vowels so that on each trial the target was immediately preceded by context material.

Experiment 2: Speaker normalization with preceding context

Method

Participants Sixteen new native speakers of Cantonese were recruited among students at the Chinese University of Hong Kong (9 female, all but one right handed, average age 20.5 years). The same selection, reimbursement, and consent procedures were used as for Experiment 1.

Materials & Procedure The same stimuli were used as for Experiment 1, except that the initial (target) consonant-vowel sequences were excised from the materials and appended at the end of the files. In other words, instead of /fV p^ha21 tsi25/), the stimuli had the structure of /p^ha21 tsi25 fV/. The downside of this approach is that the resulting stimuli inevitably deviate from natural coarticulatory patterns. Furthermore, phrases typically contain an F0 downdrift which may also be violated by appending the initial /fV/ to the end of the context. The advantage, though, is that it keeps the stimuli

maximally similar across the two experiments. Furthermore, note that the F1 and F0 target ranges in phase 1 were extrapolated beyond the natural recordings, and as a result participants could thus settle on a generally lower F0 target range in case that resulted in perceptually more ambiguous sounds. Indeed, based on the results of phase 1 of Experiment 2, the across-participant average midpoint of the ambiguous region for F0 was somewhat lower than for Experiment 1 (see Appendix B for detail). Apart from these changes all other stimulus properties and procedures were identical to Experiment 1.

Results

Following the data-inclusion criteria explained for Experiment 1, no participants were excluded overall. Subsets of block-types were excluded for some participants, however. The number of participants contributing to any individual block type was minimally 13 participants and maximally 16 participants. Figure 3 displays the overall categorization behavior across the conditions. The same model structure was applied as for Experiment 1. The analysis revealed a main effect of Step ($B = 1.54$, $z = 11.36$, $p < 0.001$), reflecting the observation that more /fo55/ responses were given for stimuli on the /fo55/ end of the continuum. An effect of Context was observed ($B = -2.20$, $z = -7.29$, $p < 0.001$) as more /fo55/ responses were given for stimuli with the low F0/F1 context. An interaction was observed between the factors Step and Mode ($B = -0.10$, $z = -2.54$, $p = 0.01$), as categorization curves were generally more shallow in the mixed presentation conditions. An interaction was observed between the factors Context and Cue ($B = 1.32$, $z = 6.39$, $p < 0.001$) as the effect of Context was larger in the Tone materials than in the Vowel quality materials. A small interaction was also observed between the factors Cue and Mode ($B = 0.17$, $z = 2.30$, $p = 0.02$). There was a tendency for fewer /fo55/ responses in the mixed speaker condition, which was stronger for the tone materials (i.e., especially in this condition one can observe that the red line is closer to 0 than the blue line is to 1). Finally, there was a significant three-way interaction between the factors Context, Cue, and Mode ($B = 0.37$, $z = 3.06$, $p = 0.002$). Especially in the mixed presentation condition, there was a difference in the normalization effect between the tone and vowel quality materials.

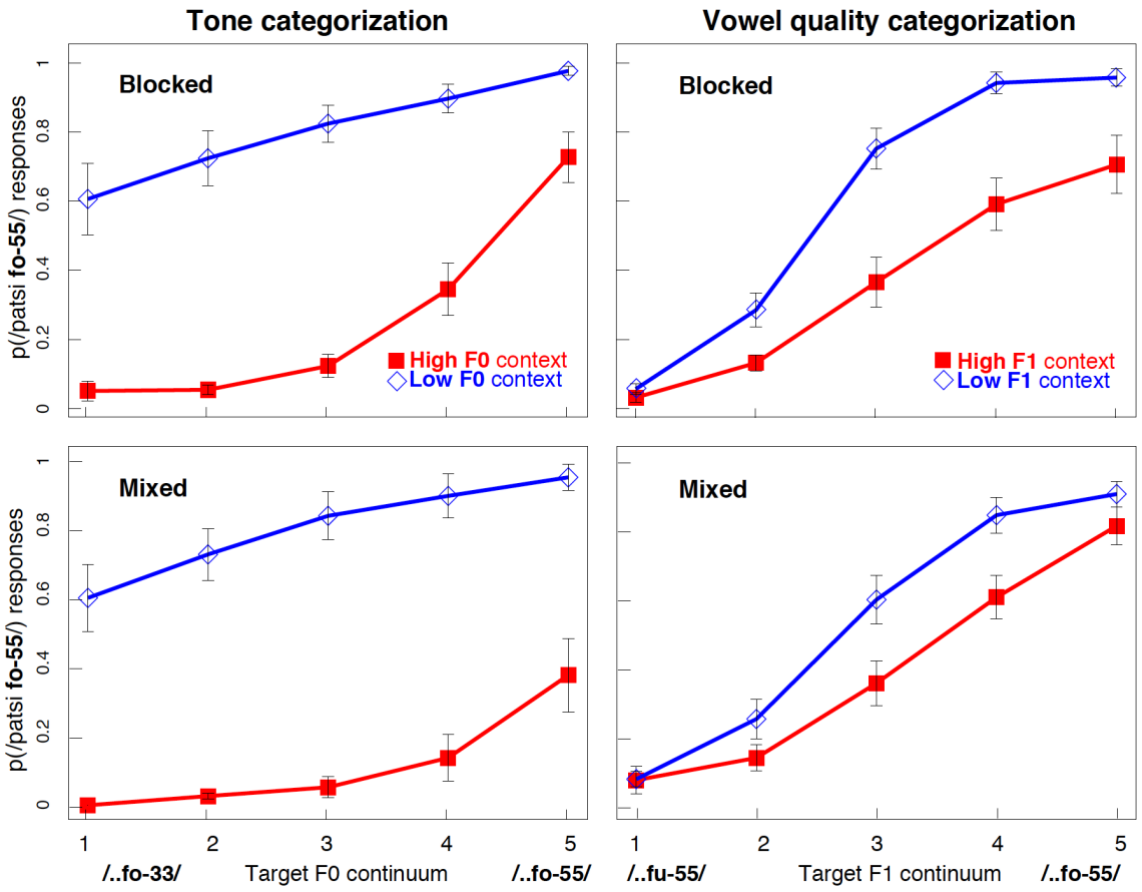

Figure 3. Experiment 2 categorization behavior with preceding context. See figure 2 legend for detail.

Finally, an investigation of effects induced by the previous trial-context in the mixed presentation conditions was performed. The same model approach (including Step; Context; Previous trial-context; and Cue) was taken as in Experiment 1 for this analysis. As in the analysis presented in the previous paragraph, significant main effects were observed for Step (B = 1.41, z = 10.78, $p < 0.001$) and Context (B = -2.11, z = -8.78, $p < 0.001$), and an interaction was observed between Context and Cue (B = 1.52, z = 7.07, $p < 0.001$). Interestingly, a main effect was observed for Previous trial-context (B = -0.14, z = -2.69, $p = 0.007$), indicating that speaker information of a previous trial had effects on the interpretation of the target sound on subsequent trials.

Summary: Visualization of across-experiment effect sizes

To ease visual comparison and discussion of context effects across the different conditions in the two experiments, Figure 4 visualizes the distribution of the context effects in the two experiments by plotting the by-participant context effects (i.e., the proportion of /fo55/ responses in the low F0/F1 condition minus the proportion in the high F0/F1 condition).

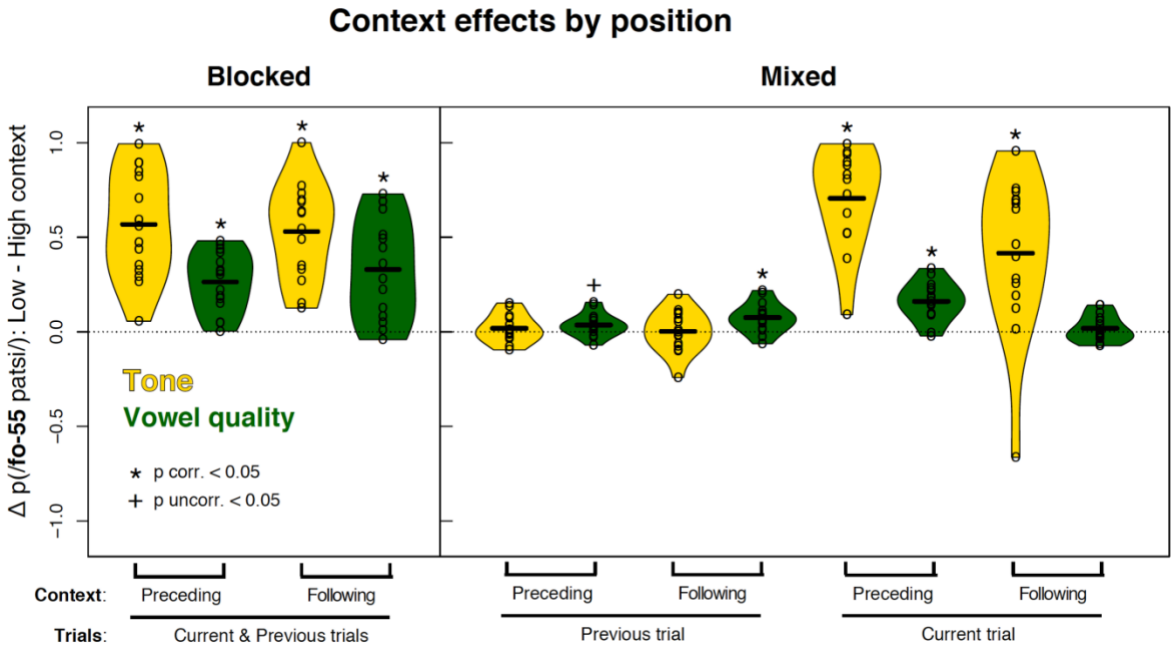

Figure 4. Violin plots of context effects of Experiments 1 and 2 combined. Context effects are calculated as the difference between the proportion of /fo55 p^{h}a21 tsi25/ responses in the high and low contexts. Yellow distributions display Tone data, green distributions Vowel quality data. Circles indicate individual subject context effects. For each distribution the horizontal black line indicates the mean. Left panel: Context effects of pre- and post-cursors, induced in a blocked presentation condition; Right panel: Context effects of pre- and post-cursors, induced in a mixed presentation condition. Data are displayed for both effects induced by context in the previous trial and those induced by the current trial. Significance marks indicate: '' for Bonferroni corrected significance at p < 0.05; '+' for uncorrected significance at p < 0.05. See Appendix C for detailed statistics.*

From this figure three important observations can be made. 1) Focusing on the blocked presentation conditions (presented in the left panel of Figure 4; related to the top panels in Figures 2 and 3): it can be observed that both tone and vowel quality elicit strong and reliable normalization effects. In addition, however, effects of tone normalization were generally larger than effects of vowel quality normalization[2]. Both observations are independent of whether the context materials were presented as preceding or following contexts. 2) For tone normalization, the effects of context in the current trial are of a roughly similar size in the mixed and blocked conditions. That is, across the two experiments, there is no interaction between Mode and Context ($B = -0.04$, $z = -0.27$, $p = 0.78$). Although note that effects of preceding contexts are larger than those of following contexts in mixed presentation (expressed as an interaction between Context and Experiment: $B = -1.42$, $z = -2.86$, $p = 0.004$). The similarity in effect size between blocked and mixed presentation for tone normalization suggests that the robust normalization for tone as observed in blocked presentation may be almost fully attributed to the normalization effects induced by the local trial context. This suggestion is further supported by the fact that the context in the previous trials had no reliable effect on tone perception. For vowel quality a different pattern is observed. In the mixed presentation mode, the size of the normalization effect induced by context on the current trial is substantially smaller than that observed in the blocked (i.e., fully immersive) presentation mode (across-experiment interaction between Mode and Context: $B = 0.25$, $z = 4.28$, $p < 0.001$). This suggests that the effect sizes observed for normalization of vowel quality in blocked presentation are the result of a slower buildup over the previous trial or trials. Indeed, the context in the just preceding trial did have a reliable effect on the perception of vowel quality. It should be noted that these influences of further preceding spectral properties are in line with observations from Holt (2005) who demonstrated influences of context carrying over a 1.3 second silent interval. In the current design, contexts from a preceding trial were found to carry across the 1.5 second response window (see also Benders, Escudero & Sjerps, 2012). 3) Finally, in the mixed speaker presentation mode, the perception of tone was found to be reliably influenced by both preceding and following context of the current trial, whereas no influence of the

following context was observed for the perception of vowel quality. Note that for both tone and vowel quality stimuli, participants could only make their response when the complete stimulus was presented, so that the following context was always available to the listener for integration. Still, for judgements of vowel quality listeners did not use the following context for normalization.

General Discussion

Two experiments were designed to investigate the scope over which the perception of tone and vowel quality is normalized to their context. In Experiment 1, listeners heard target sounds that were followed by a disyllabic context. The contexts had either a high or low F0 (for the tone continuum), or a high or low F1 (for the vowel quality continuum). High and low context stimuli were either intermixed within the same block, or the different contexts were presented in separate blocks. Experiment 2 was a replication of Experiment 1 except that the context was presented immediately *before* the target sound instead of after it. In both experiments, tone and vowel quality stimuli were always presented in separate blocks.

The results demonstrated considerable qualitative and quantitative differences in the scope of the context effects between the two speech cues. The contextual influences of surrounding tone distributions came from both preceding and following contexts. Those influences could be mostly attributed to local information (i.e., from the context in the current trial). For vowel quality, on the other hand, only preceding but not following contexts influenced target perception. Moreover, stronger normalization effects were observed for blocked as compared to mixed presentation modes, indicating that normalization for vowel quality extends further back in time compared to that of tone.

Given these findings, one may ask why tone and vowel quality are integrated over different scopes? One reasonable consideration is that tone may be most strongly affected by the local context because the F0 information carried by the local utterance is most useful for estimating a speaker's *current* F0 range. The realization of lexical tone is strongly affected by phrase-level information (e.g., phrase

boundary), emphatic stress, sentence-level prosodic information (e.g., statement/question), and natural F0 downdrift. Hence, normalizing F0 based on preceding phrases could in fact result in worse perception. For vowel quality, however, estimates of formant frequencies from preceding phrases may be more informative. That is because there is no general utterance-level decrease in formant values across utterances that is similar to the well-known downshift in F0 across a sentence (Connell, 2001; Poser, 1983; Ohala, 1978; Wong & Diehl, 2003)[3]. This allows formant values in preceding sentences to be relatively more informative for the interpretation of current vowels than for tone. In other words, information about F0 distributions gained from previous phrases or sentences may be of less value than information about formant distributions because of the added variability in F0 from utterance to utterance. As such, it may be especially important to integrate tone within a local utterance, or at least more so than it is for formants.

The findings reported here have a number of implications for our broader understanding of the processing of tone and vowel quality. For vowel quality, or spectral properties more generally, it has been suggested that normalization may, for an important part, be the result of general auditory processes that help to enhance perceptual contrast (e.g., Kluender, Coady & Kiefte, 2003). That is, thinking beyond specific speakers and their vocal-tract properties, stable acoustic properties of the listening channel (be it room acoustics or the filter properties of a telephone line) can aid any sound classification task for which listeners use spectral properties, such as recognizing different instruments in a musical ensemble (Stilp, Alexander, Kiefte & Kluender, 2010). Indeed, spectral properties of non-speech sounds can affect the interpretation of following speech sound in the context of "compensation for coarticulation" (e.g., Holt, 2006), and similar effects induced by nonspeech contexts occur for vowel normalization (e.g., Sjerps et al., 2011, 2012; Watkins, 1991; Watkins & Makin 1996; see also Sjerps & Smiljanić, 2013, for cross-language effects). Given that normalization for vowel quality may be based, for an important part, on such general auditory processes[4] it makes sense that influences of context would not be restricted to the acoustically arbitrary boundary of the current phrase or sentence.

For the perception of tone, on the other hand, influences of F0 in non-speech context seem to be more restricted. Although there is some evidence of non-speech context induced tone normalization in Mandarin (Huang & Holt, 2009), a number of studies have found no, or markedly reduced influences of nonspeech precursors on tone perception. Zhang et al. (2012), for example, compared the effect sizes of tone normalization that were induced by F0-bearing speech and nonspeech precursors. They observed reliable normalization effects for the speech precursors but no, or strongly reduced, normalization for nonspeech precursors. Similarly, no normalization of lexical tone is observed when the context consists of a continuous hummed neutral vowel (Francis et al. 2006). In line with these observations, here we observed that normalization for tone was in fact restricted to the current phrase. That is, it was restricted to a scope that is acoustically arbitrary but linguistically relevant. These results thus suggest that normalization for tone may be relatively speech specific, or at least more so than the normalization for vowel quality.

A further aspect of the results presented here is that for vowel quality, only the preceding context affected perception, while for tone both preceding and following context affected tone perception. Although the observation for vowel quality is in agreement with an early report on normalization for vowel quality in Dutch (van Bergem, Pols & Koopmans-van Beinum, 1988) one may still ask why listeners would fail to use information that, in everyday life, is highly informative of the formant range of a current speaker? And why would this property differ between two speech cues? Although the current results cannot provide a definitive answer to this question, a number of previous studies have reported that perceptual decisions on tonal information may remain open for reinterpretation over a longer time window than perceptual decisions about segmental cues such as vowel quality (Cutler & Chen, 1997; Ye & Connine, 1999). As such, a final perceptual decision on tonal information may arise at a later time point when compared to vowel quality. As long as the listener has not committed to a decision about tonal information, their ultimate decision may also remain susceptible to influences of following context.

To conclude, the results presented here demonstrate that the normalization of vowel quality and

tone operate over a qualitatively different temporal scope. Tone cues are interpreted relative to the local phrase or trial, regardless of whether the context is provided before or after the target sound. Vowel quality is interpreted relative to preceding contexts only, and this preceding context seems to consist of a further extended range of input materials.

Footnotes

[1] Note that we describe the lexical tones here using Chao's tone letters (Chao, 1930). These range on a scale from 1 to 5, with 1 referring to the lowest pitch and 5 referring to the highest pitch. Each lexical tone is defined using two numbers, which describe, in an abstract way, the pitch at the beginning and end of a syllable respectively. Therefore, "55" is a high level tone, "33" a mid level tone, "21" a low falling tone and "25" a high rising tone.

[2] Note that this property is probably not explained as a consequence of the relation between the sizes of the target range and that of the context manipulation, as both had a similar ratio. However, the current experiment was not designed to address overall normalization effect sizes directly, and the fact that the target ranges that were used were tailored to individual participants does not allow for a proper assessment of this issue.

[3] Note that sentence-level prosodic information does affect the amount of so-called "centering" of vowels in vowel space (i.e., a shift towards the formant values associated with the neutral vowel /ə/; e.g., Chen, 2008; Chen & Gussenhoven, 2008; Cho, 2004). However, assuming that a typical sentence contains vowels from most of the vowel space there is no overall *decrease* in the mean formant values across the length of the utterance.

[4] It is important to add, however, that vowel quality is also influenced in ways that cannot be attributed to general auditory contrast effects (such as visual effects: Johnson, Strand & D'Imperio, 1999; Hay & Drager, 2010).

Acknowledgements

The project was supported through the Research Grants Council of Hong Kong (GRF project: 14408914). The first author received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7 2007-2013 under REA grant agreement nr. 623072. We would like to thank Neal Fox for his suggestions on parts of the analyses. We would also like to thank two anonymous reviewers and Christian Stilp for useful comments on an earlier submitted versions of this paper. All data and analysis scripts have been made available through the Open Science Framework (https://osf.io/u7xnp/).

References

Assgari, A. A., & Stilp, C. E. (2015). Talker information influences spectral contrast effects in speech categorization. The Journal of the Acoustical Society of America, 138(5), 3023–3032.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D. M., Mächler, M, Bolker, B. M., Walker, S. C. lme4: Linear Mixed-Effects models using eigen and S4 R package. Version 11–7. 2014; http://CRANR-projectorg/package=lme4.

Benders, T., Escudero, P., & Sjerps, M. J. (2012). The interrelation between acoustic context effects and available response categories in speech sound categorization. The Journal of the Acoustical Society of America, 131(4), 3079-3087.

van Bergem, D. R., Pols, L. C., & Koopmans-van Beinum, F. J. (1988). Perceptual normalization of the

vowels of a man and a child in various contexts. Speech Communication, 7(1), 1-20.

Boersma, P. P. G. (2002). Praat, a system for doing phonetics by computer. Glot international, 5.

Chao, Y.-R. (1930). A system of tone letters. Le maiître Phonétique, 45, 24–27.

Chen, Y. (2008). The acoustic realization of Shanghai vowels. Journal of Phonetics 36 (4): 629-748.

Chen, Y. & Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. Journal of Phonetics 36 (4): 724-746.

Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. Journal of Phonetics 32 (2), 141-176.

Connell, B. (2001). Downdrift, Downstep, and Declination. Typology of African Prosodic Systems Workshop, Bielefeld University, Germany.

Cutler, A., & Chen, H. C. (1997). Lexical tone in Cantonese spoken-word processing. Perception & Psychophysics, 59(2), 165-179.

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. The Journal of the Acoustical Society of America, 119(3), 1712-1726.

Garrett, K. L., and Healey, E. C. (1987). An acoustic analysis of fluctuations in the voices of normal adult

speakers across three times of day. Journal of the Acoustical Society of America, 82:58-62.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. Linguistics, 48(4), 865-892.

Heald, S. L., & Nusbaum, H. C. (2015). Variability in Vowel Production within and between Days. PloS one, 10(9), e0136791.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. Psychological Science, 16(4), 305-312.

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. Journal of the Acoustical Society of America, 120:2801-2817.

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. The Journal of the Acoustical Society of America, 125(6), 3983-3994.

Huang, J., & Holt, L. L. (2011). Evidence for the central origin of lexical tone normalization (L). The Journal of the Acoustical Society of America, 129(3), 1145-1148.

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: Phonetic targets are hyper articulated. Language, 505-528.

Johnson, T. L., & Strange, W. (1982). Perceptual constancy of vowels in rapid speech. The Journal of the Acoustical Society of America, 72(6), 1761-1770.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. Journal of Phonetics, 27, 359–384

Kiefte, M., & Kluender, K. R. (2008). Absorption of reliable spectral characteristics in auditory perception. The Journal of the Acoustical Society of America, 123(1), 366-376.

Kluender, K. R., Coady, J. A., & Kiefte, M. (2003). Sensitivity to change in perception of speech. Speech Communication, 41(1), 59-69.

Krause, J. C., & Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. The Journal of the Acoustical Society of America, 115(1), 362-378.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U. & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. Science, 277(5326), 684-686.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. Journal of the Acoustical Society of America, 29(1), 98-104.

Leather, J. (1983). Speaker normalization in perception of lexical tone. Journal of Phonetics.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological review, 74(6), 431.

Lin, T., & Wang, W. S.-Y. (1984). Diao ganzhi wenti [Questions regarding tone perception]. Journal of

Chinese Linguistics, 2, 59–69.

Malins, J. G., & Joanisse, M. F. (2010). The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. Journal of Memory and Language, 62(4), 407–420.

Malins, J. G., & Joanisse, M. F. (2012). Setting the tone: An ERP investigation of the influences of phonological similarity on spoken word recognition in Mandarin Chinese. Neuropsychologia, 50(8), 2032–2043.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. Perception & Psychophysics, 25(6), 457-465.

Mitterer, H. (2006). Is vowel normalization independent of lexical processing? Phonetica, 63, 209–229.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. The Journal of the Acoustical Society of America, 102(3), 1864-1877.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. Journal of the Acoustical Society of America, 85, 2088–2113.

Newman, R. S., & Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. Journal of phonetics, 37(1), 46-65.

Ohala, J. J. (1978). Production of tone. In V. A. Fromkin (Ed.), Tone: A linguistics survey (pp. 5–39). New York: Academic Press.

Peterson G.E. & Barney H.L. (1952). Control methods used in the study of vowels. Journal of the Acoustical Society of America, 24, 175–184.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking Clearly for the Hard of Hearing II. Acoustic Characteristics of Clear and Conversational Speech. Journal of Speech, Language, and Hearing Research, 29(4), 434-446.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1989). Speaking Clearly for the Hard of Hearing III. An Attempt to Determine the Contribution of Speaking Rate to Differences in Intelligibility between Clear and Conversational Speech. Journal of Speech, Language, and Hearing Research, 32(3), 600-603.

Poser, W. J. (1983). On the mechanism of F0 downdrift in Japanese. Journal of the Acoustical Society of America, 74(S1), S89–S89.

Protopapas, A., and Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. Journal of the Acoustical Society of America, 101:2267-2277.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. Journal of Phonetics, 41(2), 101-116.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2012). Hemispheric differences in the effects of context on

vowel perception. Brain and language, 120(3), 401-405.

Sjerps, M. J., McQueen, J. M., & Mitterer, H. (2013). Evidence for precategorical extrinsic vowel normalization. Attention, Perception, & Psychophysics, 75(3), 576-587.

Sjerps, M. J., & Smiljanić, R. (2013). Compensation for vocal tract characteristics across native and non-native languages. Journal of Phonetics, 41(3), 145-155.

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. The Journal of the Acoustical Society of America, 118(3), 1677-1688.

Stilp, C. E., Alexander, J. M., Kiefte, M., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. Attention, Perception, & Psychophysics, 72(2), 470-480.

Stilp, C. E., Anderson, P. W., & Winn, M. B. (2015). Predicting contrast effects following reliable spectral properties in speech perception. The Journal of the Acoustical Society of America, 137(6), 3466-3476.

Stilp, C. E., & Assgari, A. A. (2017). Consonant categorization exhibits a graded influence of surrounding spectral context. The Journal of the Acoustical Society of America, 141(2), EL153-EL158.

Toscano, J. C., & McMurray, B. (2015). The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments. Language, cognition and neuroscience, 30(5), 529-543.

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. Journal of the Acoustical Society of America, 90, 2942–2955.

Watkins, A. J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. The Journal of the Acoustical Society of America, 99(6), 3749-3757.

Wong, P. C., & Diehl, R. L. (2003). Perceptual normalization for inter-and intratalker variation in Cantonese level tones. Journal of Speech, Language, and Hearing Research, 46(2), 413-421.

Xu, Y. (1997). Contextual tonal variations in Mandarin. Journal of Phonetics, 25, 61–83.

Ye, Y., & Connine, C. M. (1999). Processing spoken Chinese: The role of tone information. Language and Cognitive Processes, 14(5-6), 609-630.

Zhang, C., Peng, G., & Wang, W. S. (2012). Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones. The Journal of the Acoustical Society of America, 132(2), 1088-1099.

Zhang, C., Peng, G., & Wang, W. S. Y. (2013). Achieving constancy in spoken word identification: Time course of talker normalization. Brain and language, 126(2), 193-202.

Zhang, C., & Chen, S. (2016). Toward an integrative model of talker normalization. Journal of Experimental Psychology: Human Perception and Performance, 42(8), 1252-1268.

Zhang, K., Wang, X., & Peng, G. (2017). "Normalization of lexical tones and nonlinguistic pitch contours: Implications for speech-specific processing mechanism". The Journal of the Acoustical Society of America, 141 (1), 38-49.

Online Supplement A:

Recordings were made from a male native speaker of Cantonese (aged 33). Between 25 and 30 repetitions of the three-syllable nonsense sequences /fo55 p^{h}a21 tsi25/ (科琶紫 "science", "guitar", "purple"), /fu55 p^{h}a21 tsi25/ (夫琶紫 "husband", "guitar", "purple"), and /fo33 p^{h}a21 tsi25/ (課琶紫 "class", "guitar", "purple") were recorded. For Experiment 1, the initial syllable was the target and the two final syllables served as the context. The context was chosen to cover the full range of F1 in the vowel space (/a–i/) and the full pitch range in the tone space (1–5).

Acoustic processing of the stimuli was carried out using Praat software (Boersma & Weenink, 2014). Twenty-five representative instances of each of the tri-syllables were selected. For each of these instances, the three syllables were excised individually. For every syllable, formant and pitch values were estimated at 100 time-points regardless of duration. Maximally 6 formants were estimated between 0 and 5500 Hz for the first syllable items, and 5 formants for /p^{h}a21/ and /tsi25/. Pitch estimation was restricted to the range of 75 - 400 Hz. Then, for every syllable, a single average formant and a single average pitch trajectory was calculated in Praat (one trajectory each for /o33/; /o55/; /u55/; /p^{h}a21/; /tsi25/). The quality of formant and pitch estimation for every token was visually inspected by the first author before allowing it to contribute to the means. Averages for the target sounds were as follows: /o33/ pitch at 128 Hz, F1 at 552 Hz; /o55/ pitch at 168 Hz, F1 at 554 Hz; /u55/ pitch at 171 Hz, F1 at 379 Hz.

For the initial (target) syllables, a continuum was generated that linearly transitioned between the time-varying endpoint tracks to cover the tone and formant ranges. That is, for the tone continuum we linearly extrapolated and interpolated from the endpoint values of /o33/ and /o55/ (left bottom panel of Figure 1). We also applied extrapolation of formant and F0 values because listeners' most preferred values may lie beyond those used by the recorded speaker. A pretest with these materials was used to define a smaller, participant-specific, target range (see Procedure section for detail). The endpoints of the

generated continuum had a pitch at the midpoint of the vowels of 118 Hz and 180 Hz respectively, covering this range in 16 steps (F1 was fixed at 540 Hz). For the vowel quality materials, trajectories of the target vowel were generated that linearly extrapolated and interpolated from the averaged endpoint values (/u55/ and /o55/). This covered a range around and between the respective formant values (right middle panel of Figure 1). The endpoints of this continuum had an F1 at the midpoint of the vowels of 292 Hz and 641 Hz respectively, covering this range in 16 steps (F0 was fixed at 170 Hz). 16 steps were chosen because relatively small step sizes were necessary to be able to select highly ambiguous tokens. For both tone and vowel quality materials, formants higher than F1 were set at a single average formant track across the respective endpoints. For resynthesis, a source signal was extracted from a representative vowel (/o55/) with the Linear Predictive Coding (LPC) procedure in Praat (the resynthesis approach is similar to that taken in, e.g., Sjerps and Smiljanić, 2013 and Sjerps, Mitterer & McQueen, 2012). This source signal was combined with each of the formant filters. The resulting signal was then low-pass filtered with a cutoff frequency at 3000 Hz and then combined with the high frequency portion (>3000 Hz) of the representative vowel that was used for source estimation. The created target stimuli were adjusted so that their amplitude envelope and overall amplitude matched that of the original representative token mentioned above (the top panels of Figure 1 display an example waveform of one of the generated stimuli).With respect to the context, the average first formant value of /p^ha21/ was 816 Hz and the pitch value was 94 Hz (measured at the midpoint); for /tsi25/, the average formant value was 307 Hz and the pitch value was 116 Hz (measured at the midpoint). For both vowels, the pitch contour was either increased or decreased by 20 Hz across the duration of the vowel, thus generating a high F0 and a low F0 context, in addition to a neutral F0 (unchanged) context. Pitch manipulations were implemented in Praat by extracting a pitch contour, modifying the pitch values and resynthesizing with the "overlap-add" method. Similarly, for the formant filters the formant tracks of the first formant values were increased or decreased by 100 Hz across the duration of the vowel, thus generating a high F1 and a low F1 context, in addition to a neutral F1 context. The shift of F0 by 20 Hz and of F1 by 100 Hz were the largest shifts we

could impose while still resulting in relatively naturally sounding stimuli. Formants higher than F1 were set at neutral values for all stimuli. For resynthesis of these contexts the same approach was taken as for the target vowels as described above. Afterwards, the resynthesized contexts were recombined with the different steps of the manipulated target items to create the experimental items. The target syllables were approximately 315 ms in length and the context approximately 646 ms in total. The vocalic portion of the target started 100ms after initial frication onset. The vocalic portion of the context starts approximately 300 ms after the onset of the vocalic portion of the target items.

Online Supplement B: Selected stimuli

Experiment / Cue / Mode	Step 1	2	3	4	5
Exp. 1 Tone					
Blck F0	139(6)	148(4)	151(4)	155(4)	163(7)
Mix F0	138(5)	148(5)	152(5)	155(5)	164(8)
Vowel Quality					
Blck F1	363(41)	404(28)	426(28)	448(28)	492(30)
Mix F1	366(40)	407(26)	429(26)	451(26)	495(28)
Exp. 2 Tone					
Blck F0	137(6)	143(5)	147(5)	151(5)	158(9)
Mix F0	137(7)	144(6)	147(6)	151(6)	159(10)
Vowel Quality					
Blck F1	400(46)	442(35)	464(35)	486(35)	526(49)
Mix F1	401(48)	445(34)	467(34)	489(34)	530(48)

Means and standard deviations (in parentheses) of the selected target continua cue-values across all participants for each experiment and block type.

Online Supplement C: Post-hoc tests

Experiment / Cue / Predictor	Blocked	Mixed
Exp. 1 Tone		
Context	B = -1.61, *z* = -4.98, p < 0.001*;	B = -1.21, *z* = -3.80, p < 0.001*
Previous trial-context		B = 0.01, *z* = 0.12, p = 0.91, ns
Vowel Quality		
Context	B = -0.76, *z* = -3.91, p < 0.001*;	B = -0.03, *z* = -0.94, p = 0.35, ns
Previous trial-context		B = -0.15, *z* = -3.85, p < 0.001*
Exp.2 Tone		
Context	B = -2.11, *z* = -5.16, p < 0.001*;	B = -2.63, *z* = -6.84, p < 0.001*
Previous trial-context		B = -0.14, *z* = -1.61, p = 0.11, ns
Vowel Quality		
Context	B = -0.66, *z* = -4.71, p < 0.001*;	B = -0.36, *z* = -6.43, p < 0.001*
Previous trial-context		B = -0.09, *z* = -2.51, p = 0.012^{+}

Post-hoc analyses were carried out to determine whether normalization effects were present in the individual combinations of Mode and Cue data. Models included only main effects for Context in the Blocked design and main effects for Context and Previous trial context in the Mixed design. The random effects structure included (uncorrelated) by-subject intercepts and slopes for both main effects. $^{+}$ after the reported p values indicates significance at uncorrected p < 0.05; * indicates significance at corrected p < 0.05 (bonferroni method, for 12 tests).