Perceptual learning of novel intonation patterns by Cantonese and Wu speakers

Man Yan Priscilla Lam, Wenwei Xu, Yiya Chen

Leiden University Centre for Linguistics, Leiden University
m.y.p.lam@hum.leidenuniv.nl; w.xu@hum.leidenuniv.nl; yiya.chen@hum.leidenuniv.nl

ABSTRACT

This study explores how tonal language speakers categorize novel intonation patterns. The perceptual intonation learning paradigm constructed in [1] was tested on native speakers of two Sinitic languages, Cantonese and Wu. The results were compared with those from the English native speakers in [1]. No difference was found in the success or rejection rates between the Sinitic language speakers and English speakers in various conditions. However, further analysis of reaction time revealed subtle differences between the speaker groups, such as a longer reaction time observed for English speakers when processing some of the intonation patterns. The results suggest that while the ability to categorize simple intonation contours may be universal among tonal and non-tonal language speakers, listeners' sensitivity to different cues can vary depending on their linguistic background.

Keywords: intonation, perceptual learning, categorization, pitch aptitude

1. INTRODUCTION

1.1. Pitch variation and intonation processing

Pitch variation is a vital part of the speech signal in human languages. Across spoken languages of the world, in any utterance containing sonorant segments, there exists pitch variation. The possible (para)linguistic functions of pitch variation are wide-ranging. In lexical tone, for example, which is estimated to be present in 60–70% of the world's languages [2], pitch variation can occur at the syllable level to mark lexical meaning. In contrast, in intonation, which is a universal feature of all languages [3], pitch variation occurs at the utterance level to convey discourse-level meaning such as sentence mode and information structure [4].

That many communicative functions can be encoded by pitch variation leaves curious implications on how speakers of typologically different languages process pitch signals. For successful speech communication, intonation language speakers need to attend to pitch variation at the utterance level, while tonal language speakers need to additionally attend to pitch variation at a much lower level such as syllables and morphemes. Given the continuously varying nature of pitch in an utterance, what has remained open is how, during intonation processing, listeners of tonal languages extract the relevant pitch cues for discourse-level information.

In intonation languages, a popular view regarding intonation processing is that listeners can parameterize an intonation contour into major intonational events such as pitch accents and boundary tones [4]. Previous studies on English online intonation processing provided evidence corroborating this claim [5,6].

The ability of intonation language listeners to pay attention to major intonational events and use them as cues for discourse comprehension likely makes intonation learning more effective and less cognitively demanding. Kapatsinski and colleagues [1] conducted a perceptual intonation learning experiment with English children and adults, involving stimuli derived from three intonation pattern prototypes: flat, final fall, and double peak. They found that both children and adults were able to learn new intonation contours quickly and formed relatively abstract representations of the contours. Adults were also found to attend more to the end of a contour than the beginning.

To our knowledge, there is no research on intonation learning by speakers of tonal languages. Considering tonal speakers need to constantly detect syllable- or morpheme-level pitch variation to determine lexical meaning, it is reasonable to assume that they use higher temporal resolution than intonation language speakers during pitch perception. Such an assumption is also in line with the PENTA model of intonation [7] where pitch targets are specified per syllable. A possible consequence is that they may find it difficult to only attend to major intonational events when learning temporally extended patterns, and thus perform worse than intonation language speakers in this respect.

1.2 Overview of the present study

In this project, we set out to investigate how tonal language speakers process pitch information to learn intonation patterns and whether their learning patterns differ from those of non-tonal speakers. We adopted the perceptual learning paradigm to address our research question. In perceptual learning, it is believed that experienced categories could enhance one’s future perceptual processing ability. In [1], a perceptual intonation learning experiment was conducted with speakers of English. We aimed to extend this line of research with speakers of two Sinitic languages: Hong Kong Cantonese (Yue; ‘HKC’) and Kunshan Wu (Wu; ‘KS’).

Within the Sinitic language family, Cantonese and Wu constitute an interesting pair for comparison due to radical structural differences between their tonal systems. In Cantonese, each syllable, regardless of its position, is specified for a lexical tone. On the contrary, in Wu, there is obligatory tone sandhi in prosodic domains larger than a syllable. In the case of KS, not all syllables in a sandhi domain are specified for tone. Regarding the density of tone specification above the syllable level, HKC and KS represent two opposite ends of the spectrum within the Sinitic languages, thus providing an ideal testing ground for whether the effect of L1 tonality on intonation learning is further modulated by the typological characteristics of the L1 tone system.

2. METHODS

2.1. Experimental design

We adapted the design of the intonation learning experiment in [1] and kept the audio stimuli identical to those published in [1]. Each stimulus is a string of fifteen segmentally identical syllables (/mi/), with an intonation contour superimposed on it. The intonation contours are random distortions of three prototypical pitch contours: “flat” (a monotone), “final fall” (a pitch drop over the last four syllables), and multifeature (hereafter “M”; two peaks at the fourth and eleventh syllable respectively). The distortion level is set to be one, three, or five semitones. An additional type of exemplar is “M” distractors, which are manipulations of the “M” prototype with 1-semitone distortion, such that the first peak, the second peak, or the valley between the two peaks is eliminated; they are referred to as “late peak”, “early peak” and “hat” respectively. For details of how the stimuli were constructed, readers are referred to [1].

In the learning task, participants were first exposed to 36 exemplars (generated from low-level distortions of the three prototypical contours). Each intonation category was associated with one of three aliens represented by a name and a picture. In the next phase, different from the original design in [1], participants heard the same training exemplars but had to identify which of the three aliens each exemplar corresponds to. Then, regardless of the correctness of their response, they received feedback with the correct alien displayed and the exemplar replayed. This feedback phase was one novelty of our experiment and enabled us to know the effect of the training before the final test phase. In the test phase, participants were presented with a total of 72 exemplars, which consists of the training exemplars, prototype exemplars, novel exemplars with three levels of distortion, and “M” distractors. In this phase, participants had to either pick one of the three aliens, or choose “None of the above” if they judge the exemplar as not belonging to the three categories.

Another novelty of the current study is that two tasks were added to account for possible extralinguistic effects on intonation learning: an n-back task and a musical background questionnaire developed from Edinburgh Lifetime Musical Experience Questionnaire [8]. They served to examine the participants’ working memory capacity [1,9,10] and indirectly estimate their pitch aptitude [11,12,13] respectively. Also, a linguistic background questionnaire developed from LEAP-Questionnaire [14] and the Lexical Test for Advanced Learners of English (LexTALE) [15] were included to serve as indicators of the participants’ language background and proficiency in English, which is an additional language of all participants.

The experiment was administered online using the *Gorilla Experiment Builder* (www.gorilla.sc). The working language of the experiment was Chinese except that LexTALE was administered in English.

2.2. Participants

A total of 62 participants completed the experiment (age: 21–32), including 31 HKC speakers (16F, 14M, 1 non-binary gender; age: *M=23.3*, *SD=1.32*) and 31 KS speakers (16F, 15M; age: *M=26.3*, *SD=1.47*). Participants were recruited through online social media platforms, and they received monetary compensation for taking part in the study.

2.3. Analysis

The responses in the intonation learning task were transformed into three types, *success*, *confusion* and *rejection*, for assessment of the participants’ performance. *Confusion* corresponds to wrong identification of the intonation categories, while *rejection* includes all the instances of the “None of the above” option chosen. The full dataset included the data collected in this experiment and the data on English native adult speakers published by [1]. Logistic regression and ordinary regression were fitted for the responses and the reaction time, respectively, with the R package *lme4* [16]. The baseline logit model contained a by-subject random intercept and by-speaker random slopes for distortion level and intonation category. The following fixed effects were included if model fit was improved: distortion level (*1-semitone, 3-semitone, 5-semitone*), intonation category (*flat, final fall, M*) and language group (*English, HKC, KS*). The other effects, including gender, working memory, musical experience, English proficiency, and exposure to English (living in English-speaking countries for at least one year), were only tested on the data collected for this experiment, given that they were not conducted in [1]. Post-hoc comparison was performed where necessary with Holm-Bonferroni correction.

3. RESULTS

3.1 Learner profile

Regarding the participants in this experiment only, no effect was found for gender, musical experience, and exposure to English on their performance based on their success rates (*success* vs. the rest). Working memory showed a significantly positive effect (p = 0.007) for categorizing low-level (1-semitone) distortions. English proficiency showed a marginally significant effect (p = 0.021) in that the participants with higher English proficiency were less likely to correctly identify 3-semitone distortions.

3.2 Categorization performance

Two measures were specifically analysed to compare the participants' performance between language groups. "M" distractors were excluded for this part of analysis.

First, the success rate (*success* vs. the rest) was analysed to directly compare the listeners' ability to identify an exemplar for the correct intonation category rather than make a mistake or judge the exemplar as not belonging to the three learned intonation categories. Both distortion level and category, as well as their interaction, significantly improved the logit model fit. However, including the language group did not improve the model fit, nor did its two-way interaction with distortion level or category respectively.

Then the rejection rate (*rejection* vs. the rest) was analysed as an indirect measure of a listener's ability to attend to multiple features. It has been argued that the more features a listener can attend to at a time, the more likely they can distinguish exemplars with decreasing perturbation distance from the training items [1]. This may be reflected by them judging the exemplar as not belonging to the three learned categories. Distortion level, intonation category and their interactions all significantly improved the logit model fit, but language group and its interactions again did not.

Therefore, the learning results were comparable across the two language groups. As shown in Figure 1, across intonation categories, exemplars were less likely to be correctly identified and more likely to be rejected as the distortions become larger ($p < 0.001$ within each category between distortion levels). Within low-level (1-semitone) distortions, "final fall" was more likely to be correctly identified than the other categories ($p < 0.001$), while there was no significant difference between "M" and "flat" (p = 0.250). "Flat" was more likely to be rejected than the other categories ($p < 0.001$), while there was no significant difference between "final fall" and "M" (p = 0.058). Within high-level (3-semitone or 5-semitone) distortions, there was no significant difference between categories regarding both success and rejection ($p > 0.069$), except that in 5-semitone distortions, "M" was more likely to be rejected than "final fall" (p = 0.016).

For low-level distortions of "M" and "M" distractors only, the rejection rate was analysed as it reflects the participants' tolerance of the absence of some of the necessary features. Again, language group did not improve the model fit. Across language groups, all three types of distractors were more likely to be rejected than "M" ($p < 0.001$), among which "late peak" was the least likely to be rejected ($p < 0.001$), and "hat" the most (p = 0.002 between "hat" and "early peak").

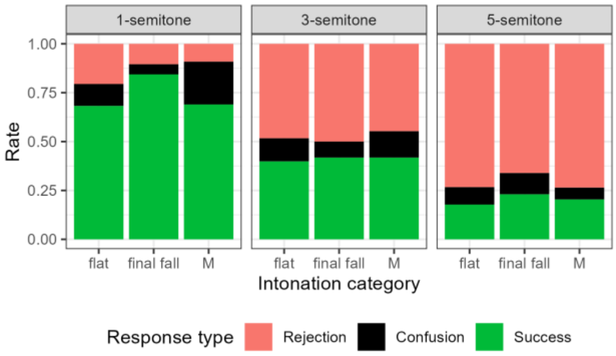

Figure 1: The success, confusion, and rejection rates by intonation category and distortion level.

3.2 Reaction time

For reaction time (RT), the predictor variables response type, distortion level, intonation category, language group, and their interactions were tested in sequence, and a three-way interaction model was first established between the first three variables. Including language group and its interaction with intonation category improved the model fit, which was also significant in the final model (language: $p < 0.011$; language × category: $p < 0.001$). Post-hoc comparison showed that it was within the "flat" category that the English listeners showed a longer RT than both HKC ($p < 0.001$) and KS ($p = 0.001$) listeners, and within the "M" category that HKC listeners showed a shorter RT than the English listeners ($p = 0.001$). Given that there are up to three-way interactions, raw RTs are visualised in decile plots (mean RT within each decile, see Figure 2) for each intonation category between language groups.

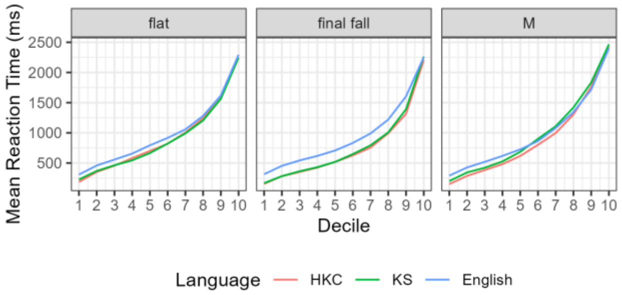

Figure 2: Decile plots of reaction time.

4. DISCUSSION & CONCLUSIONS

The categorization performance and RT patterns presented above have offered complementary perspectives on how tonal language speakers categorize intonation contours compared with non-tonal language speakers.

First, the Sinitic language speakers' performance shows that they have comparable categorization ability to the English speakers, at least for the paradigm and intonation categories tested in this experiment. The same patterns between distortion levels and intonation categories were observed across the language groups: (i) exemplars with lower-level distortions were easier to learn and categorize; (ii) within low-level distortions, "final fall" was an easier category than "flat" and "M". As discussed earlier, it is reasonable to speculate that one's ability to categorize intonation patterns is modulated by the tonality in their L1. The performance results of this study do not corroborate this idea, nor the idea that tonal language speakers, who are believed to use higher temporal resolution for pitch, would find it harder to only attend to major intonational events. However, it is worth noting that all KS and HKC participants in this study were exposed to English since childhood, and some (36%) have lived in an English-speaking country for at least one year. It is possible that the considerable exposure to English has affected the way these tonal language speakers process intonation contours, resembling the behaviour of English native speakers.

Regarding the importance of different features within an intonation contour, the results concerning "M" distractors demonstrate a finality advantage: among the three distractor types, "late peak" is least likely to be rejected, and thus probably deemed more acceptable as a member of "M". This corroborates Kapatsinski and colleagues' [1] findings, for which they proposed the following explanation: since the end of an utterance is often the most informative, it is likely to be associated with pitch peaks; consequently, listeners pay more attention to the end of an intonation contour. Our results suggest that the final prominence effect may be universal across languages.

The conditional effect of English language proficiency is also intriguing. That speakers with higher proficiency perform worse than those with lower proficiency, crucially only in the 3-semitone distortion level but not in the 1-semitone level, suggests that greater English proficiency may reduce intonation category breadth, which means they are better at attending to multiple features.

Whereas there is no apparent difference between the performance of different language groups, their RT reveals some subtle differences. L1 background is shown to have an effect on RT in the categorization of some of the intonation categories: HKC and KS speakers process "final fall" faster than the English speakers, and HKC speakers process "M" faster than the English speakers. Given that this is the only difference between the two tonal language groups, it is unlikely a consequence of different experimental settings. One possible explanation for the Sinitic speakers' processing advantage (compared to English speakers) concerns the language-specific functional load of these intonation contours. For example, in both HKC and KS, dubitative questions can be constructed with sentence-final falling intonation, which potentially leads to a high functional load for similar contours. The higher functional load may, in turn, cause HKC and KS speakers to devote more attention to utterance-final contours. In contrast, in English, a falling intonation typically marks a declarative statement, which can be considered the 'default' setting.

To conclude, this study shows that while language background, specifically L1 tonality, does not necessarily modulate one's intonation categorization performance, it could nonetheless pose a processing advantage for certain intonational contours. Finally, to enable a clearer examination of the L1 tonality effect, it would be worthwhile to conduct further experiments with native speakers of tonal languages who do not have any non-tonal language experience. Future studies should also consider a wider range of intonation contours.

5. ACKNOWLEDGEMENTS

We wish to thank Kapatsinski and colleagues [1] for making available the stimuli and dataset of their study. This research is supported by an NWO VICI grant (vi.c.181.040) to Yiya Chen.

6. REFERENCES

[1] V. Kapatsinski, P. Olejarczuk, and M. A. Redford, “Perceptual Learning of Intonation Contour Categories in Adults and 9-to 11-Year-Old Children: Adults Are More Narrow-Minded,” *Cognitive Science*, vol. 41, no. 2, pp. 383–415, 2017.

[2] M. Yip, *Tone*. Cambridge: Cambridge University Press, 2002.

[3] C. Gussenhoven, The Phonology of Tone and Intonation. Cambridge: Cambridge University Press, 2004.

[4] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 2008.

[5] K. Ito and S. R. Speer, “Anticipatory effects of intonation: Eye movements during instructed visual search,” *Journal of Memory and Language*, vol. 58, no. 2, pp. 541–573, 2008.

[6] W. F. L. Heeren, S. A. Bibyk, C. Gunlogson, and M. K. Tanenhaus, “Asking or telling – real-time processing of prosodically distinguished questions and statements,” *Language and Speech*, vol. 58, no. 4, pp. 474–501, 2015.

[7] Y. Xu, “Speech melody as articulatorily implemented communicative functions,” *Speech communication*, vol. 46, no. 3–4, pp. 220–251, 2005.

[8] J. A. Okely, I. J. Deary, and K. Overy, “The Edinburgh Lifetime Musical Experience Questionnaire (ELMEQ): Responses and non-musical correlates in the Lothian Birth Cohort 1936,” *PloS one*, vol. 16, no. 7, p. e0254176, 2021.

[9] T. J. Laméris and B. Post, “The combined effects of L1-specific and extralinguistic factors on individual performance in a tone categorization and word identification task by English-L1 and Mandarin-L1 speakers,” *Second Language Research*, p. 02676583221090068, 2022.

[10] S. Goss, “Exploring variation in nonnative Japanese learners’ perception of lexical pitch accent: The roles of processing resources and learning context,” *Applied Psycholinguistics*, vol. 41, no. 1, pp. 25–49, 2020.

[11] A. Cooper and Y. Wang, “The influence of linguistic and musical experience on Cantonese word learning,” *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4756–4769, 2012.

[12] P. C. Wong and T. K. Perrachione, “Learning pitch patterns in lexical identification by native English-speaking adults,” *Applied Psycholinguistics*, vol. 28, no. 4, pp. 565–585, 2007.

[13] J. A. Alexander, P. C. Wong, and A. R. Bradlow, “Lexical tone perception in musicians and non-musicians,” in *Ninth european conference on speech communication and technology*, 2005.

[14] V. Marian, H. K. Blumenfeld, and M. Kaushanskaya, “The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals,” *Journal of Speech, Language, and Hearing Research*, vol. 50, pp.9 942–947, *2007*.

[15] K. Lemhöfer and M. Broersma, “Introducing LexTALE: A quick and valid lexical test for advanced learners of English,” *Behavior research methods*, vol. 44, no. 2, pp. 325–343, 2012.

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting Linear Mixed-Effects Models Using lme4,” *J. Stat. Soft.*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.