NATIVE, NAÏVE, AND EXEMPLAR-BASED PERCEPTION OF STATEMENT AND QUESTION INTONATION IN CANTONESE AND MANDARIN

Una Y. Chow[1], Stephen J. Winters[2]

[1]University of British Columbia, [2]University of Calgary
una.chow@ubc.ca, swinters@ucalgary.ca

ABSTRACT

This study investigated whether an exemplar model of speech perception could account for the intonation-based classification of statements and questions in Cantonese and Mandarin. Both native and naïve listeners of each language performed a sentence-type identification task. They were presented with gated forms of 80 pairs of statements and questions that ended in all the lexical tones in each language. An exemplar-based model also simulated the listening task by classifying the same tokens based on the Euclidean distance of the F0 values between new and previously presented tokens. Results showed that the naïve listeners and the model performed worse than native listeners on whole-utterance stimuli, but all groups performed similarly well on gated utterances that comprised the final syllable only. Both naïve listeners and the model performed at similar above-chance levels, suggesting that a "naïve" exemplar-based model could account for the naïve listeners' perception of intonation.

Keywords: intonation perception, exemplar model, statement and question, Cantonese, Mandarin

1. INTRODUCTION

Previous research on speech variability has investigated whether Exemplar Theory could account for the perception of vowels [6], words [7], syllables [20], pitch accents [21], intonation [4], intonation and lexicon [19], as well as dialects [2]. Each of these studies focused primarily on native speakers of a language. Few studies have applied Exemplar Theory to non-native speakers' perception of speech [10]. To better understand exemplar effects on non-native speech perception, this study compared the perceptual sensitivity of native listeners, naïve listeners, and an exemplar-based model on the identification of statement and question intonation in Cantonese and Mandarin.

Cantonese and Mandarin differ in both tonal and intonation systems; thus, they provide two different test cases for this study. Cantonese has six lexical tones [1, 3]: high-level /55/, high-rising /25/, mid-level /33/, low-falling /21/, low-rising /23/, and low-level /22/. Mandarin, on the other hand, has four lexical tones [13, 16]: high-level /55/, rising /35/, low-falling(-rising) /21(4)/, and falling /51/.

This study used declarative questions that are yes/no questions seeking confirmation from the listener. In Cantonese, declarative questions end in a high F0 rise [8] or a high boundary tone [23] regardless of the tone of the final syllable. In Mandarin, however, they exhibit a gradual increase in F0 towards the end of the utterance [14] as well as an overall higher pitch level than statements [24]. They also retain the tonal shape of the final syllable [5]. Similarly, statements in both Cantonese and Mandarin also retain the final tonal contour.

Our study was designed to address the following research questions: (1) How well can naïve listeners correctly identify statement and question intonation in Cantonese and Mandarin, compared to the native listeners? (2) How well would an exemplar-based model perform on the same task, compared to both the native and naïve listeners? (3) Are there cross-linguistic differences in all three groups' performances on the task?

2. EXEMPLAR-BASED MODEL

We propose an exemplar-based model that uses a simplified version of the algorithm from [11] and [17]. It categorizes statements and questions based on intonation, without normalization of fundamental frequency (F0) for each speaker. Since F0 is a salient acoustic correlate of intonation for Cantonese and Mandarin, it was used as an auditory property to calculate the auditory similarity between the compared sentences. The auditory distance d_{ij} between a new token i and a previously experienced token j was determined by the Euclidean distance of the F0s at eleven equidistant timepoints [$t_0..t_{10}$] of i and j, as shown in (1).

$$(1) \qquad d_{ij} = \sqrt{\textstyle\sum_{t=0}^{10}\left(F0_{it} - F0_{jt}\right)^2}$$

The auditory distance d_{ij} was then applied to the exponential function e^{-x} to derive the auditory similarity s_{ij} between i and j. This function enables auditorily close exemplars to have greater influence in the calculation of auditory similarity. We took the overall similarity between a new token *i* and a category to be the sum of the auditory similarity values between *i* and every token *j* in that category. The model then assigned *i* to the category (i.e., 'statement' or 'question') that had the higher overall similarity value with *i*.

3. EXPERIMENT 1: CANTONESE

3.1. Method

3.1.1. Participants

Three groups of listeners participated in the experiment: (1) native, (2) naïve, and (3) the exemplar-based model. The native listeners (10 male, 10 female) originated from Guangdong, China (n=7), Hong Kong (n=7), and Canada (n=6). These listeners performed the identification task (for a larger cross-linguistic study of intonation perception) prior to the naïve listeners. Since there was no significant difference in the results between genders (ANOVA: $p > .05$), only the 10 female listeners (age in years: 18-28, *M*=22.00, *SD*=2.36) were analyzed in this study.

Counterbalancing the 10 native female listeners were 10 naïve female listeners (age in years: 18-28, *M*=20.80, *SD*=3.05) who had no knowledge of Cantonese. Both the native and naïve listeners were fluent English speakers. They were recruited from the University of Calgary and reported no visual, speech, or hearing disorders.

The exemplar-based model simulated 10 listeners as it performed 10 separate classifications of the Cantonese tokens.

3.1.2. Stimuli

To create the stimuli, 10 native Cantonese speakers who originated from Hong Kong (5 male, 5 female; age in years: 18-35, *M*=23.00, *SD*=1.49) each read 20 pairs of statements and declarative questions. These sentence-type pairs were identical syntactically and lexically but differed in their intonation contours (e.g., *Wong*^{55} *Ji*^{22} *gaau*^{33} *lik*^{22} *si*^{25}. *Wong*^{55} *Ji*^{22} *gaau*^{33} *lik*^{22} *si*^{25}? 'Wong Ji teaches history'). The speakers reported no visual, speech, or hearing impairments and were recorded individually in a sound-attenuated booth at the University of Calgary using high-quality equipment at 44.1 kHz.

The recorded sentences of four randomly selected speakers (2 male, 2 female), 80 pairs in total, were then used as stimuli for the listening experiment. To determine the effect of final tone on the identification of the sentence type, the original recordings were gated in three forms: (1) the whole sentence (WHOLE, e.g., *Wong*^{55} *Ji*^{22} *gaau*^{33} *lik*^{22} *si*^{25}.), (2) the final syllable (FINAL, e.g., *si*^{25}), and (3) the non-final portion of the utterance (NON-FINAL, e.g., *Wong*^{55} *Ji*^{22} *gaau*^{33} *lik*^{22}).

3.1.3. Procedure

The identification task comprised two sessions that were conducted one to seven days apart. Half of the 80 pairs of sentences were used for training and the remaining half were used for testing. The training and test tokens were reversed between sessions. Since this study focused on the naïve listeners' first experience with the test language, only the results from session one were reported. Table 1 lists the five phases in session one of the identification task.

Table 1: Session one of the identification task.

Part	Phase	# of Trials	Stimulus Type
I	Practice	4	WHOLE
II	Training	80	WHOLE
III	Testing	80	WHOLE
IV	Practice	8	NON-FINAL, FINAL
V	Testing	160	NON-FINAL, FINAL

The training and test tokens were also randomized and counter-balanced between listeners. They were presented in ten different orders, one for each listener of the three listener groups in the study.

The listeners were presented with the stimuli through headphones one token at a time. In each trial, they responded whether the stimulus that they had just heard was (part of) a statement or question by pressing the appropriate key on a keyboard. The brief practice exercises, which were intended to familiarize the listeners on how to do the task, used sentences that differed from the sentences that were used in training/testing. The speaker who produced these sentences also differed from the four speakers who produced the training/test stimuli. During practice and training, the correct sentence type was displayed after each response. During testing, however, only the number of correct responses was displayed after every 10 trials.

Similarly, the model was first trained on the training data, which became exemplars 'in memory'. During the categorization process, it compared each test token with the statement and question exemplars in memory (using the algorithm described above). Once the test token had been categorized, it became another experienced token in memory and was used to categorize subsequent tokens.

3.1.4. Analysis

For the perceptual analysis, we converted the listeners' responses to measures of sensitivity (d-prime: d') [15]. A d' value of zero means performance at chance level. We then ran a three-way ANOVA with d' as the dependent measure and with listener group (native, model, and naïve), stimulus type (WHOLE, NON-FINAL, and FINAL), and lexical tone as independent factors. Since this study examines perception across listener groups, we will only report significant effects or interactions involving listener groups (at $\alpha = .05$).

3.2. Results

Figure 1: Sensitivity (d') by listener group and stimulus type for Cantonese.

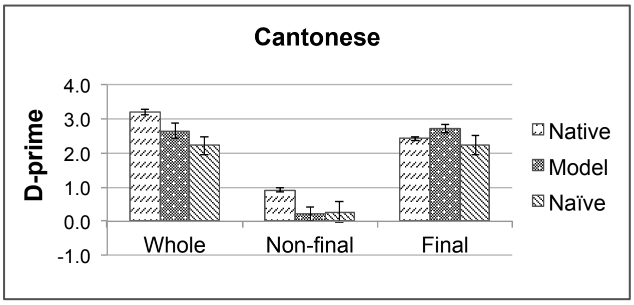


Fig. 1 shows that all three listener groups correctly identified the sentences at an above-chance level. A three-way ANOVA on d' indicated a significant main effect of listener group [$F(2, 27) = 8.34$, $p < .01$] and a significant interaction between listener group and stimulus type [$F(4, 54) = 8.98$, $p < .001$]. However, there was no significant interaction among listener group, stimulus type, and tone. A post-hoc Tukey HSD test revealed that the native listeners were significantly more sensitive to the statement-question distinction than both the model and the naïve listeners ($p < .001$; mean difference $\bar{d} = .31$ and $\bar{d} = .60$, respectively), while the model was significantly more sensitive than the naïve listeners ($p < .001$; $\bar{d} = .29$).

Specifically, on WHOLE stimuli, the native listeners performed significantly better than both the model ($p < .01$; $\bar{d} = .54$) and the naïve listeners ($p < .001$; $\bar{d} = .97$), while the model performed significantly better than the naïve listeners ($p = .04$; $\bar{d} = .43$). On NON-FINAL stimuli, the native listeners also performed significantly better than both the model ($p < .001$; $\bar{d} = .70$) and the naïve listeners ($p < .001$; $\bar{d} = .65$), but there was no significant difference between the model and the naïve listeners. On FINAL stimuli, there was no significant difference between the native listeners and the model or the naïve listeners, but the model performed significantly better than the naïve listeners ($p < .01$; $\bar{d} = .49$).

4. EXPERIMENT 2: MANDARIN

Experiment 2 replicated Experiment 1 in design, procedure, and analysis, but used Mandarin instead of Cantonese listeners and stimuli.

4.1. Method

4.1.1. Participants

Similar to Experiment 1, three groups of listeners participated in the experiment: (1) native, (2) naïve, and (3) the exemplar-based model. The native listeners (10 male, 10 female) originated from China but not Hong Kong. Since there was no significant difference in d' between genders (ANOVA: $p > .05$), only the 10 female listeners (age in years: 18-28, M=23.70, SD=2.45) were analyzed. These native listeners were counterbalanced with 10 naïve female listeners (age in years: 18-28, M=21.20, SD=2.35) who had no knowledge of Mandarin. Both the native and naïve listeners were fluent English speakers. They were recruited from the University of Calgary and reported no visual, speech, or hearing impairments. Ten runs of the exemplar-based model simulated 10 listeners of Mandarin.

4.1.2. Stimuli

The stimuli were recorded and gated in the same manner as the stimuli for the Cantonese experiment. Sixteen native Mandarin speakers (8 male, 8 female; age in years: 18-35, M=24.94, SD=4.80), who originated from China but not Hong Kong, produced the Mandarin statements and declarative questions. These sentences ended in all four of the Mandarin tones.

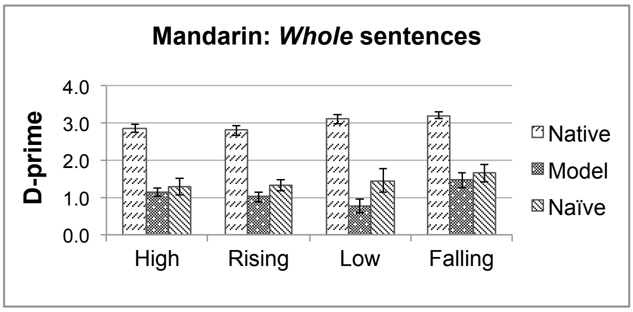
4.2. Results

All three listener groups correctly identified the sentence types at an above-chance level for all three stimulus types. A three-way ANOVA on d' showed a significant main effect of listener group [$F(2, 27) = 64.82$, $p < .001$]. It also indicated significant interactions between listener group and stimulus type [$F(4, 54) = 13.40$, $p < .001$] and among listener group, stimulus type, and tone [$F(12, 162) = 2.47$, $p < .01$]. A post-hoc Tukey HSD test revealed that, overall, the native listeners were significantly more sensitive to the statement-question distinction than both the model ($p < .001$; $\bar{d} = 1.20$) and the naïve listeners ($p < .001$; $\bar{d} = 1.03$).

Specifically, on WHOLE stimuli (Fig. 2), the native listeners performed significantly better than both the model ($p < .001$; $\bar{d} = 1.72\text{-}2.33$) and the naïve listeners ($p < .001$; $\bar{d} = 1.47\text{-}1.65$). On NON-

FINAL stimuli that excluded a final high or falling tone, the native listeners also performed significantly better than the model ($p < .01$; $\bar{d} = 1.24$ or 1.13) and the naïve listeners ($p < .01$; $\bar{d} = 1.23$ or 1.17). Furthermore, the native listeners performed significantly better than the naïve listeners only on NON-FINAL stimuli that excluded a final low tone ($p < .05$; $\bar{d} = 1.10$) and better than the model only on FINAL stimuli that carried the rising tone ($p < .01$; $\bar{d} = 1.17$).

Figure 2: Sensitivity (d') of WHOLE stimulus type by listener group and final tone for Mandarin.



5. DISCUSSION

In both Cantonese and Mandarin, the naïve listeners and the exemplar-based model performed at above chance levels. Additionally, they performed (nearly) as well as the native listeners on utterances that were presented with their final syllables only. Both results are unexpected because neither the naïve listeners nor the exemplar-based model had knowledge of the intonation patterns of the test language prior to the identification task. For the naïve listeners, their experience with the rising intonation in English yes/no question [12, 22] likely influenced their performance on the task since the declarative questions in both languages end in a higher pitch than statements (i.e., a final rise in Cantonese [8] and a raise in pitch in Mandarin [18]). There is also a general tendency for some language speakers to perceive utterances with a final rising intonation as questions [9]. On the other hand, the exemplar-based model's above chance performance provides evidence that F0 is a salient cue for declarative questions in Cantonese and Mandarin.

On whole utterances, however, the native listeners outperformed the naïve listeners and the exemplar-based model. This result suggests that, in addition to the salient cue for questions at the end of the utterance, there is distinguishing information between statements and questions in the non-final portion of the utterance to which only the native listeners were sensitive, most likely because of their experience with their native intonation. [10] noted a similar case—presented by Bradlow—in which the native Mandarin listeners and the non-native (English) listeners performed similarly in a discrimination task on Mandarin tones, when the tones were presented in monosyllables. However, when the tones were presented in trisyllables, the non-native listeners performed worse because they "relied more on acoustic similarity between stimuli" whereas the native listeners "could rely on their abstract knowledge of the categories" [10].

Compared with the naïve listeners, the exemplar-based model performed similarly well in Mandarin but better on stimuli that included the final syllables in Cantonese. This performance difference reflects the relatively larger F0 difference towards the end of the statement and question intonation patterns in Cantonese than in Mandarin. Comparing performance across final tones, no differences emerged for the Cantonese stimuli, possibly due to the presence of a high boundary tone cue at the end of questions [23]. The same comparison for Mandarin revealed tonal differences likely because Mandarin retains the shape of the final tone at the end of both statements and questions [5].

6. CONCLUSION

This study used an exemplar-based model to investigate the effects of native tone and intonation on the identification of statements and questions in two tone languages that differ in their intonation patterns. On the one hand, native experience (or more exemplars) helped the native listeners perform better on whole utterances than the naïve listeners and the exemplar-based model. On the other hand, familiarity with similar intonation patterns from another language (or relying primarily on acoustic similarity) could compensate for the lack of native experience, as indicated by all three listener groups' similar performances on the final syllables alone.

The performance of the exemplar-based model on the sentence-type identification task closely paralleled the performance of the naïve listeners rather than the native listeners, suggesting that an exemplar categorization process could in principle account for the naïve listeners' perception of sentence-type intonation, at least for Cantonese and Mandarin. Future modeling work will examine the potential effect of the listener's first language on the identification of intonation patterns in a non-native language.

7. ACKNOWLEDGEMENTS

This research was supported by the Social Sciences and Humanities Research Council of Canada.

8. REFERENCES

[1] Bauer, R. S., Benedict, P. K. 1997. *Modern Cantonese Phonology*. Berlin: Mouton de Gruyter.

[2] Boomershine, A. 2006. Perceiving and processing dialectal variation in Spanish: An exemplar theory approach. *Selected Proc. 8th Hispanic Linguistics Symposium* Somerville, 58–72.

[3] Chan, Y. Y. F. 1974. *A Perceptual Study of Tones in Cantonese*. Hong Kong: HKU Press.

[4] Chow, U. Y., Winters, S. J. 2015. Exemplar-based classification of statements and questions in Cantonese. *Proc. 18th ICPhS* Glasgow. Paper 0987.

[5] Chow, U. Y., Winters, S. J. 2016. The role of the final tone in signaling statements and questions in Mandarin. *Proc. 5th TAL* Buffalo, 167–171.

[6] Ettlinger, M., Johnson, K. 2009. Vowel discrimination by English, French and Turkish speakers: Evidence for an exemplar-based approach to speech perception. *Phonetica* 66(4), 222–242.

[7] Goldinger, S. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. J. *Exp. Psychol. Learn.* 22, 1166–1183.

[8] Gu, W., Hirose, K., Fujisaki, H. 2005. Analysis of the effects of word emphasis and echo questions on F0 contours of Cantonese utterances. *Proc. 6th Interspeech* Lisbon, 1825–1828.

[9] Gussenhoven, C., Chen, A. 2000. Universal and language-specific effects in the perception of question intonation. *ICSLP 6* Beijing, 91–94.

[10] Hazan, V. 2007. Second language acquisition and exemplar theory. *Proc. 16th ICPhS* Saarbrücken, 43–48.

[11] Johnson, K. 1997. Speech perception without speaker normalization: An exemplar model. In: Johnson, K., Mullennix, J. W. (eds), *Talker Variability in Speech Processing*. San Diego: Academic Press, 145–165.

[12] Ladd, D. R. 2008. *Intonational Phonology, 2nd ed.* Cambridge: Cambridge University Press.

[13] Li, C. N., Thompson, S. A. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

[14] Liu, F., Surendran, D., Xu, Y. 2006. Classification of statement and question intonations in Mandarin. *Proc. 3rd Speech Prosody* Dresden. Paper 232.

[15] Macmillan, N. A., Creelman, C. D. 2005. *Detection Theory: A User's Guide, 2nd ed.* New Jersey: Lawrence Erlbaum Associates, Inc.

[16] Norman, J. 1988. *Chinese*. Cambridge: Cambridge University Press.

[17] Nosofsky, R. M. 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. *J. Exp. Psychol.: Learn.* 14, 700–708.

[18] Peng, S.-H., Chan, M. K. M., Tseng, C.-Y., Huang, T., Lee, O. J., Beckman, M. E. 2005. Towards a Pan-Mandarin system for prosodic transcription. In: Jun, S.-A. (ed), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press, 230–270.

[19] Schweitzer, K., Walsh, M., Calhoun, S., Schütze, H., Möbius, B., Schweitzer, A., Dogil, G. 2015. Exploring the relationship between intonation and the lexicon: Evidence for lexicalised storage of intonation. *Speech Communication* 66, 65–81.

[20] Walsh, M., Schütze, H., Möbius, B., Schweitzer, A. 2007. An exemplar-theoretic account of syllable frequency effects. *Proc. 16th ICPhS* Saarbrücken, 481–484.

[21] Walsh, M., Schweitzer, K., Schauffer, N. 2013. Exemplar-based pitch accent categorisation using the Generalized Context Model. *Proc. 14th Interspeech* Lyon, 258–262.

[22] Wells, J. C. 2006. *English Intonation: An Introduction*. Cambridge: Cambridge University Press.

[23] Wong, W. Y. P., Chan, M. K. M., Beckman, M. E. 2005. An autosegmental-metrical analysis and prosodic annotation conventions for Cantonese. In: Jun, S.-A. (ed), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press, 271–300.

[24] Yuan, J., Shih, C., Kochanski, G. P. 2002. Comparison of declarative and interrogative intonation in Chinese. *Proc. Speech Prosody* Aix-en-Provence, 711–714.