EXEMPLAR-BASED CLASSIFICATION OF STATEMENTS AND QUESTIONS IN CANTONESE

Una Y. Chow, Stephen J. Winters

Department of Linguistics, Languages and Cultures, University of Calgary, Canada
uchow@ucalgary.ca, swinters@ucalgary.ca

ABSTRACT

Previous research has shown that exemplar theories are a promising model of speech perception. Few studies have applied exemplar-based models to intonation perception. This study investigates how intonational exemplars are represented in memory and classified by a computational model. An exemplar-based model was tested with Cantonese statements and echo questions. Each test token was represented with eleven F0 measurements at equidistant time points from the start to the end of the periodicity of the final syllable in the utterance. Three simulations tested the model's performance in classifying these utterances when varying degrees of variation were presented to the model. Results indicate that the model can accurately distinguish between statements and echo questions in Cantonese, based on the intonation of just the final syllable, without normalization for speaker and gender.

Keywords: intonation perception, classification, exemplar theory, tones, Cantonese

1. INTRODUCTION

Variability in speech production [11] poses challenges for speech perception. Sources of variation can be cross-speaker or within-speaker. For example, females on average have shorter vocal folds than males, so in general they produce speech with a higher fundamental frequency (F0) [4]. Individual speakers may produce an utterance with a modified pitch range in one instance to indicate focus [14, 15] but not in another.

Johnson [7] demonstrated how an exemplar-based model could account for human vowel perception despite talker variability. According to exemplar theory, listeners store exemplars of speech that they experience in rich phonetic detail in memory. Since the inherent variability of these exemplars is not filtered out of their representations in memory, listeners can process new tokens without using speaker normalization. A new token is compared with all of the exemplars in memory based on auditory properties and is categorized with the exemplars to which it is most similar overall. [2] showed that an intonation contour that is frequently associated with a particular word or short phrase can be stored with that utterance in memory. [12] also demonstrated that an exemplar-theoretic model can successfully categorize pitch accents (L*H and H*L).

To further investigate how well exemplar models can account for human perception of intonation, our research addresses the question: Can an exemplar-based model correctly classify statements and echo questions in Cantonese?

Cantonese has six contrastive tones: T1 (high-level), T2 (high-rising), T3 (mid-level), T4 (low-falling), T5 (low-rising), and T6 (low-level). Described using Chao's 5-scale tonal system, with 1 the lowest and 5 the highest point in a speaker's F0 range, the tones are 55, 25, 33, 21, 23 and 22, respectively [3]. [6, 13] showed that Cantonese attaches a high boundary tone (H%) to the end of yes/no questions. This final rise has the same tonal direction as T2. In [8], native speakers in a perception experiment misperceived many of the final tones of T3, T4, T5 and T6 in questions as T2. The interaction between tone and intonation thus provides an interesting test case for our study.

Previous research found that Cantonese listeners perceive statements and questions mainly based on the F0 cues of the utterance-final syllable [9]. Therefore, in this study, we examined the F0 values of the final syllables of statements and echo questions in Cantonese in order to determine whether an exemplar-based model without F0 normalization could classify these two types of sentences. We also tested the performance of the model with stimuli produced by different speakers from different genders in order to gauge the effects that talker variability had on the performance of the model. Based on the results of previous research [5, 7], we hypothesized that 1) the model could accurately classify stimuli from multiple speakers without normalizing the acoustic representations of those stimuli, and 2) less variability in the test stimuli would lead to better performance of the model.

2. METHOD

We recorded native Hong Kong Cantonese speakers producing two repetitions each of 200 pairs of statements and echo questions. Then we extracted eleven F0 measurements at successive time points of the intonation contour of the sentence-final syllable. We tested an exemplar-based model with these measurements to see how accurately it could classify the recorded sentences by sentence type according to the classification algorithm presented in [7].

2.1. Participants

Five male and five female native speakers of Cantonese from the University of Calgary, aged 21 to 26, participated in the production study. All speakers were born and raised in Hong Kong, except for one speaker who was born in Canada but lived in Hong Kong from age 2 to 16. They reported no history of speech or hearing disorders.

2.2. Stimuli

The stimuli for the production study comprised five blocks of four discourses: blocks A, B, C, D and E with target sentences that were 5, 7, 9, 11 and 13 syllables long, respectively. Each discourse included a target pair of Cantonese sentences: a statement and a question that were lexically and syntactically identical (e.g. *Wong1 Ji6 gaau3 lik6 si2. Wong1 Ji6 gaau3 lik6 si2?* 'Wong Ji teaches history'). A filler *wh*-question or a *maa*-question preceded the pair, thus introducing a statement reading for the first target utterance (e.g. *Wong1 Ji6 gaau3 mat1 je5? Wong1 Ji6 gaau3 lik6 si2.* 'What does Wong Ji teach? Wong Ji teaches history.'). The second target utterance, an echo question, followed the statement (e.g. *Wong1 Ji6 gaau3 lik6 si2?* 'Wong Ji teaches history?'). After the echo question, a filler affirmative statement terminated the dialogue (e.g. *Hai6, Wong1 Ji6 gaau3 lik6 si2.* 'Yes, Wong Ji teaches history.').

Within each block, all target pairs ended with the same syllable (i.e. *si*, *ji*, *maa*, *fu* or *fan*) but with a different Cantonese tone for each of the four discourses. We constructed the production stimuli in this fashion in order to create both intonation and tone variation for the final syllable.

2.3. Procedures

The stimuli were presented to the speakers in Chinese characters on an iMac computer in a sound-attenuated booth at the University of Calgary. The speakers were instructed to read the dialogue in their natural voice and at their normal speed. Each speaker read the stimuli in a different block order. The readings were recorded with high quality equipment at a sampling rate of 48 kHz.

2.4. Acoustic measurements

We manually labelled the periodic portion of the final syllable of the target pairs of sentences in Praat [1]. A Praat script extracted F0 values at eleven equally spaced time points of the syllable, including the start and end of the periodicity. In the model, these time-normalized points function as auditory properties for calculating perceptual distance between a new token and an exemplar in memory. Excluding sentences with final syllables that Praat failed to analyze due to devoicing or creaky voice, a total of 189 pairs of statements and echo questions from the first reading were used to test the model. In addition, speaker S14's second reading (of all 20 target pairs of sentences) was included to test simulation 3 in section 3.3.

2.5. Exemplar-based model

The exemplar-based model of intonation perception of statements and echo questions followed [7] and [10], as shown in (1-4). The model compared each newly encountered token (statement or echo question) with every stored exemplar in memory (experienced statement or echo question). The auditory distance between the new token i and an exemplar j was taken to be the Euclidean distance between them, calculated using the attention weight w_m of auditory property m:

$$(1) \quad d_{ij} = [\ \sum w_m (x_{im} - x_{jm})^2\ ]^{1/2}$$

Auditory similarity between token i and exemplar j was determined by applying an exponential decay function with sensitivity c to the auditory distance such that only the nearest neighbors were considered:

$$(2) \quad s_{ij} = \exp(-c d_{ij})$$

The degree to which token i activates exemplar j was determined by the base activation level N_j and optional Gaussian noise e_j:

$$(3) \quad a_{ij} = N_j s_{ij} + e_j$$

Finally, the evidence for category C_1 given token i was the sum of the activations of all exemplars j belonging to category C_1 (e.g. a statement):

$$(4) \quad E_{1,i} = \sum a_{ij},\ j \in C_1$$

2.6. Classification of statements and questions

To simulate human perception of intonation, our exemplar-based model calculated similarity between an unknown sentence and a categorized sentence in memory using the eleven F0 measurements of the final syllable ($F0_1$ to $F0_{11}$). The 189 pairs of recorded sentences from the speakers' first reading were used to test the model. Each pair of statement and echo question served as two unknown tokens that were compared with each of the remaining 376 sentences that served as exemplars in memory. The exemplars in memory were associated with sentence-type, gender and speaker categories. Sensitivity and the weight of each auditory property were set to 1. Noise and bias (base activation level) were ignored.

Three classification simulations were conducted. Simulation 1 classified sentences for all speakers. Simulation 2 classified sentences by gender group. Simulation 3 classified sentences produced by a single speaker. The stimuli were classified in these various conditions in order to determine the effects of variation on correct identification rates. We hypothesized that less variation would yield better performance.

For each simulation, auditory similarity was calculated for eleven different conditions. Condition 1 calculated similarity using $F0_1$ only (i.e. the F0 value at the beginning of the sound token). Each successive condition added the next F0 value to the previous condition. That is, condition 2 applied both $F0_1$ and $F0_2$ to the calculation of similarity; condition 3 applied $F0_1$, $F0_2$ and $F0_3$ to the calculation, and so forth. The purpose was to determine at which accumulated time point the model performed best for each simulation. Due to Cantonese's tonal inventory of three level tones, two rising tones and one falling tone, it is likely that perception of these tones relies on both F0 height and direction. A single time point captures only the F0 height. The incremental approach not only reflects how a listener hears a new token from beginning to end, but also reveals the F0 contour of the tone to the model.

3. RESULTS

3.1. Simulation 1

We ran a logistic regression to investigate the effect of time points on the proportion of stimuli correctly classified by the model. The performance [$G^2(10,4147) = 517.2$; $p < .001$] of the model significantly improves over the initial time point once it includes all of the F0 contour at 40% of the way through the token (condition 5 in Fig. 1). Performance continues to be better than the initial point throughout the rest of the token. Question identification starts to get better much later, in condition 7, suggesting that the meaningful rise in question intonation starts at time point 7. The mean F0 contours of questions ending in T2 to T6 produced by male and female speakers (the five bottom lines in Fig. 2 and 3) confirm this analysis. The F0 contours of questions ending in T1 (the top line in both figures), however, rise immediately from the onset in order to reach a high F0 level.

When 80-100% of the F0 contour was included, the model correctly classified 92-95% of the sentences (96-97% for statements and 88-93% for questions). Performance on statements, on average, is better than on questions. [9] also found a bias toward the perception of statements among Cantonese listeners.

Figure 1: Percent of statements (- -), questions (...), and statements and questions (—) correctly classified by the model

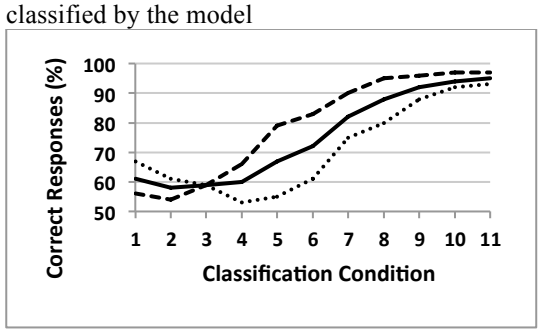


Figure 2: Mean F0 contours of statements (- -) and questions (—) ending in T1 to T6, produced by male speakers

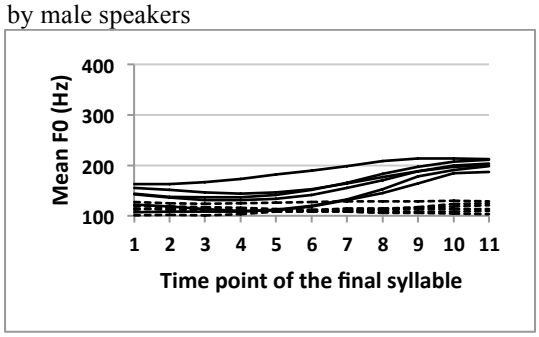
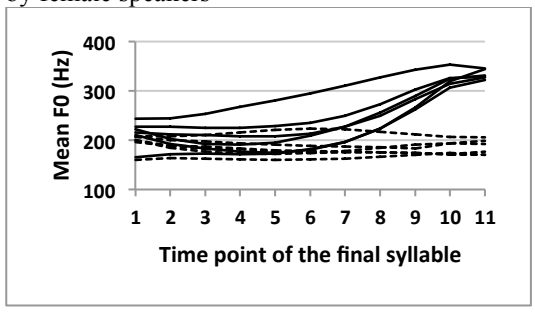


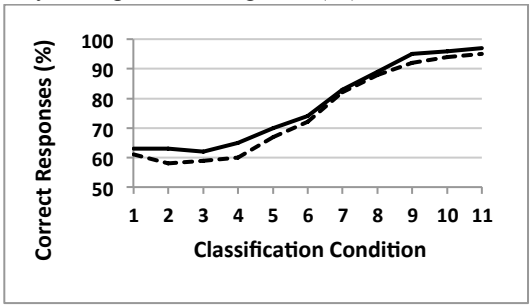
Figure 3: Mean F0 contours of statements (- -) and questions (—) ending in T1 to T6, produced by female speakers



3.2. Simulation 2

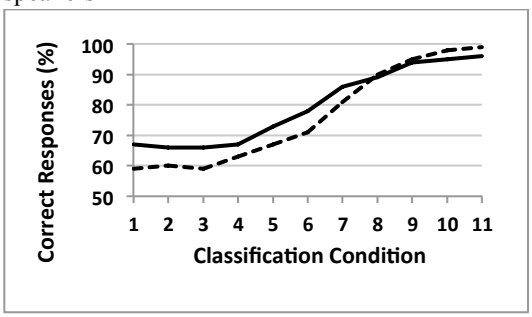
A second logistic regression investigated the effects of different classification sets on the model's performance. In one set, we tested the model on all of the recorded stimuli–produced by both male and female speakers–at one time. In another set, we tested the model on the male- and female-produced stimuli separately, one gender at a time. This analysis [$G^2(1,8314) = 9.20$; $p = .002$] revealed that the model performed better when presented with one gender at a time, rather than two together (Fig. 4). This result confirms what we expected from the variability hypothesis.

Figure 4: Percent of correct responses when the genders were presented separately (—) vs. when they were presented together (- -)



A third logistic regression investigated whether gender had an effect on the model when it was trained on only one of the two sets of speakers. In fact, it did [$G^2(1,4156) = 6.05$; $p = .013$], with the model performing significantly better overall on the male speakers than on the female speakers (Fig. 5). The male speakers in our study produced the sentences with a smaller F0 range for each sentence type than the female speakers (Fig. 2 and 3). As a result, there is less overlap through 70% of the token between statements and questions in the male speakers' utterances. In conditions 9, 10 and 11, the model performed better on female speakers due to greater F0 difference between sentence types at time points 9-11 in the female utterances.

Figure 5: Percent of correct responses when the model was tested on male (—) vs. female (- -) speakers



3.3. Simulation 3

Finally, we selected one speaker (S14) at random to determine if a particular reading or the number of readings would affect the model's performance on the classification task. In one classification set, we tested the model separately on S14's first and second readings of each token. In another classification set, we tested the model on both of S14's readings of each utterance.

There were trends in the direction of classification being better for reading two than reading one, and also for classification being better when the model had two different readings "in memory" than one. However, neither of these trends turned out to be statistically significant.

This simulation also tested whether training a model on a single speaker would yield better performance than training with multiple speakers. In conditions 9-11 of reading one, the single-talker model correctly classified 100% of the sentences (compared to 92-95% when tested on all speakers in simulation 1), likely due to the lack of cross-speaker variability.

4. CONCLUSION

This study demonstrated that an exemplar-based model could accurately classify statements and questions in Cantonese based on F0 measurements of the final syllable alone, without normalization for speaker or gender. When 80-100% of the F0 contour of the final syllable was included in the testing, the model correctly classified 92-95% of the sentences. This near-ceiling performance provides evidence that the phrase-final high boundary tone is a primary cue for distinguishing between statements and echo questions in Cantonese. It also suggests that H% can be stored in exemplar fashion in memory.

As expected, the model performed better when presented with less variable stimuli. A logistic regression revealed that the model performed significantly better in classifying the sentences when presented with one gender at a time than with both genders together. As the performance patterns of our exemplar-based model have also been found with human listeners [5, 7], the results of this study suggest that exemplar theory provides a promising model for the human perception of intonation.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] Boersma, P., Weenink, D. 2013. *Praat: doing phonetics by computer* (version 5.3.51) http://www.praat.org

[2] Calhoun, S., Schweitzer, A. 2012. Can intonation contours be lexicalised? Implications for discourse meanings. In: Elordieta, G., Prieto, P. (eds), *Prosody and Meaning (Trends in Linguistics), Volume 25*. Munchen: Walter de Gruyter, 271-327.

[3] Chao, Y.-R. 1947. *Cantonese primer*. Cambridge: Cambridge University Press.

[4] Fant, G. 1972. Vocal tract wall effects, losses, and resonance bandwidths. *Speech Transmission Laboratory Quarterly progress and status report* 2(3), 28-52.

[5] Goldinger, S. D., Pisoni, D. B., Logan, J. S. 1991. On the nature of talker variability effects in recall of spoken word lists. *J. Experimental Psychology: Learning, Memory and Cognition* 17(1), 152-162.

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

[7] 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.

[8] Ma, J. K.-Y., Ciocca, V., Whitehill, T. L. 2006. Effect of intonation on Cantonese lexical tones. *J. Acoust. Soc. Am.* 120, 3978-3987.

[9] Ma, J. K.-Y., Ciocca, V., Whitehill, T. L. 2011. The perception of intonation questions and statements in Cantonese. *J. Acoust. Soc. Am.* 129, 1012-1023.

[10] Nosofsky, R. M. 1988. Similarity, frequency and category representation. *J. Experimental Psychology: Learning, Memory and Cognition* 14, 54-65.

[11] Peterson, G., Barney, H. 1952. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24, 175-184.

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

[13] 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*. New York: Oxford University Press, 271-300.

[14] Xu, Y. 2007. Speech as articulatory encoding of communicative functions. *Proc. 16th ICPhS* Saarbrücken, 25-30.

[15] Xu, Y., Xu, C. X. 2005. Phonetic realization of foucs in English declarative intonation. *J. Phon.* 33(2), 159-197.