PERCEPTION OF INTONATION IN CANTONESE: NATIVE LISTENERS VERSUS AN EXEMPLAR-BASED MODEL*

Una Y. Chow and Stephen J. Winters
University of Calgary

1. Introduction

Speech perception researchers (Peterson and Barney 1952, among others) have known for many years that the phonemic categories of speech have highly variable acoustic/phonetic manifestations. Variation in speech depends on many factors, including the size and shape of the vocal tract (Fant 1960, 1972), the phonetic environment in which speech sounds are produced, and the speaker's sociolect. For example, Peterson and Barney (1952) measured the formant frequencies of 10 vowels in hVd words (i.e., *heed*, *hid*, *head*, *had*, *hod*, *hawed*, *hood*, *who'd*, *hud*, and *heard*), produced by 76 English speakers. They found considerable variation among different groups of speakers primarily due to their vocal tracts: on average, children produced the highest formant frequencies, women produced the second highest formant frequencies, and men produced the lowest formant frequencies. Such variation in speech poses a serious challenge to the question of how human beings can successfully identify consistent, abstract phonemic and lexical categories across a wide spectrum of different speakers, styles, and contexts.

One theoretical approach that has had success in dealing with this variability in real-world tests is Exemplar Theory, which takes the variability in speech to be a resource for perception, rather than a problem. In this paper, we investigated whether a computational model, based on Exemplar Theory principles, could account for the human perception of intonation.

1.1 Exemplar Theory of speech perception

Originating from psychological models of categorization, exemplar theories of perception claim that a category is represented by all experienced instances of the category (Hintzman 1986; Nosofsky 1986, 1988). Johnson (1997) adapted Nosofsky's (1986, 1988) model to speech perception and proposed that listeners store exemplars of speech that they experience in rich phonetic detail in memory. Since the phonetic details of these exemplars are not *normalized*, or filtered out of their mental representations, listeners can use the inherent variability of exemplars to categorize new tokens based on how similar these tokens are to the exemplars in memory, without the need for speaker normalization (Johnson 1997, 2005).

* We would like to thank the reviewers and audience of WICL 2016 and CLA 2016 for their helpful comments. This research was supported by the Social Sciences and Humanities Research Council of Canada, through a Joseph-Armand Bombardier Canada Graduate Scholarship to Una Chow. The research reported in this paper is part of a larger MA research conducted by Una Chow under the mentorship of her advisor, Stephen Winters.

Actes du congrès annuel de l'Association canadienne de linguistique 2016.
Proceedings of the 2016 annual conference of the Canadian Linguistic Association.
© 2016 Una Y. Chow and Stephen J. Winters

To test this hypothesis, Johnson (1997) simulated vowel perception using an exemplar-based model (Nosofsky 1986, 1988). The test tokens consisted of 10 different (h)Vd words, read by 14 male and 25 female native English speakers five times each. Each of these word tokens was presented to the model for categorization, while the rest of the tokens served as experienced exemplars in memory. The best-fitting model correctly categorized these word tokens 80% of the time—a success rate that is comparable to human listener performance on synthesized vowels (Lehiste and Meltzer 1973, Ryalls and Lieberman 1982, Johnson 1997). The model's confusion matrix also significantly correlated with the human listeners' confusion matrix in Peterson and Barney's (1952) vowel identification task, which used a similar list of hVd words. The results of Johnson's (1997) study demonstrated that an exemplar-based model could in principle account for certain aspects of human vowel perception.

1.2 Exemplar-based perception of intonation

Exemplar Theory has been applied less often to the perception of intonation in speech, even though there is great variation in the prosodic elements of speech as well. For instance, Flynn (2003) analyzed the six Cantonese tones produced by five native speakers of Hong Kong Cantonese, and found variation in the pitch height and slope of individual tones due to coarticulation with adjacent tones. Carryover and anticipatory effects altered the onset and offset of the target tones, respectively. Also, Warren (2005) compared the onsets of high-terminal rises in statements and questions produced by two groups of native New Zealand English speakers: 1) six male and six female teenagers between 16 and 19 years old, and 2) six male and six female adults between 30 and 45 years old. Same-sex dyads from each group produced a variety of sentences that were controlled in the study. They also freely produced a variety of sentences while performing a map task (Brown et al. 1984). Warren (2005) found that the teenage group produced more high terminal rises that started at the nuclear syllable, whereas the mid-age group produced more rises that started at a post-nuclear syllable.

Assuming that human speech perception draws on the rich phonetic details of speech, an exemplar-based model should be able to account for intonation perception as well. To date, only a few studies have investigated the classification of prosodic elements by an exemplar-based model. For example, Walsh et al. (2013) demonstrated that an exemplar-based model (Nosofsky 1988, Johnson 1997) could successfully categorize pitch accents (L*H and H*L) extracted from five hours of read speech of German radio broadcast news. However, as far as we know, currently there is no study that has investigated the classification of whole-sentence intonation contours by an exemplar-based model for a tone language.

1.3 Exemplar-based model of intonation perception in Cantonese

Chow and Winters (2015) demonstrated that an exemplar-based model (adapted from Nosofsky 1988, Johnson 1997) could correctly classify 95% of the statements and echo questions produced by 10 native speakers of Cantonese, based on the F0 values of the pitch contour of the final syllable. In this follow-up study, we compared human performance on an identification task with the model's performance in order to address

the question: can exemplar theory account for the human perception of intonation in statements and echo questions in Cantonese?

Cantonese is a tone language with six contrastive tones (Bauer and Benedict 1997, Flynn 2003). Represented using Chao's (1947) five-scale pitch levels—with 1 being the lowest and five being the highest point in a speaker's F0 range—these six tones have the pitch levels of [55], [25], [33], [21], [23], and [22], respectively. Table 1 lists all six tones, each with an example.

Tone	Number	Shape	Pitch level	Example (Jyutping)
T1	1	high-level	55	師 *si1* 'teacher'
T2	2	high-rise	25	史 *si2* 'history'
T3	3	mid-level	33	試 *si3* 'to test'
T4	4	low-fall	21	時 *si4* 'time'
T5	5	low-rise	23	市 *si5* 'city'
T6	6	low-level	22	視 *si6* 'to look at'

Table 1. *Cantonese tones*

The motivation for investigating Cantonese with our exemplar-based model of intonation perception was that its inventory of three different level tones (T1, T3, and T6) and contour tones (T2, T4, and T5) provided a wide variety of tones for testing the model. In addition, since echo questions in Cantonese typically end with a high F0 rise (Gu et al. 2005) or a high boundary tone (Wong et al. 2005)—regardless of the pitch level and direction of the final tone (as shown in Figure 1)—they can be confusable with statements that end with a [25] high-rising tone. Likewise, listeners can misperceive the final tone in echo questions as the [25] high-rising tone. In Ma et al.'s (2006) perception study of the identification of lexical tones in the final syllables of Cantonese statements and echo questions, they found that native listeners misperceived many of the [33], [21], [23], and [22] tones in the final syllable of questions as a [25] tone. Therefore, the interaction between tone and intonation in Cantonese provided a challenging test case for our exemplar-based model.

For an initial evaluation of the model's performance, we tested it on a basic but pragmatically salient intonation classification task: differentiating between statements and echo questions. These sentence-type pairs were identical syntactically and lexically but differed in their prosody and intonation contours. We hypothesized that an exemplar-based model could correctly categorize statements and echo questions in Cantonese at better-than-chance rates, without the need for speaker normalization. That is, we expected the model to be able to consistently identify the abstract category of each sentence type in spite of the speaker and tonal variability in the signal. However, we also hypothesized that our human listeners would perform better than the exemplar-based computational model on the identification task. Although we expected the computational

model to approximate human levels of performance, we recognized that our model is, at present, at an early stage of development. While similar levels of performance between the human listeners and our computational model might provide encouraging evidence that Exemplar Theory can function as an accurate model of human intonation perception, any differences in performance between the model and the human listeners in this task could be instructive as to how to refine the model to develop a more accurate understanding of how human listeners perform this perceptual task.

Figure 1. *F0 contours of statements and echo questions ending in all six tones, produced by a native female speaker. The pitch track setting is 100-450 Hz.*

2. Methods

In order to compare the model's performance with human listeners' performance on the identification of statements and echo questions in Cantonese, we conducted a production study, perception study, and a simulation using an exemplar-based model. The production study provided stimuli for testing both listeners and the exemplar-based model. The goal of the simulation was to determine how well an exemplar-based model

could categorize statements and echo questions in Cantonese, based on their intonation contours, without normalization for differences in the speaker's F0 range. The goal of the perception study was to determine how well the model performed in comparison to human listeners on the same task.

2.1 Production study

2.1.1 Speakers

Five male and five female native speakers from Hong Kong participated in the production study. They were 18-35 years of age (M = 23, SD = 1.49) and were recruited from the University of Calgary. They reported no history of visual, speech, or hearing disorders. They were paid $15 for their participation in a one-hour session.

2.1.2 Stimuli

The stimuli comprised 20 dialogues, arranged in five equal blocks. Since the utterances were intended to test the listeners' and the model's ability to identify sentence types based on intonation alone, each dialogue contained a target pair of sentences: a statement and an echo question, both identical in lexical and syntactic form. A filler question preceded the target pair, and a filler statement followed it to provide pragmatic context. In order to record a variety of sentences for testing human listeners and the exemplar-based model, the target sentences differed in both syllable length and in their final syllable across each block. Blocks A, B, C, D, and E contained target sentences of 5, 7, 9, 11, and 13 syllables long, respectively. These five blocks also contained target sentences that ended in the syllables *si*, *ji*, *maa*, *fu*, and *fen*, respectively. In order to gauge the effect of tone on sentence intonation, the target pairs in each block ended in four different tones. In addition, the target sentences in each block began with the same disyllabic name, but the name differed across each block. Example (1) shows a dialogue between A and B. The target pair in this dialogue began with the name *Wong1 Ji6* and ended with the syllable *si* in T2. The other three target pairs in this block also began with *Wong1 Ji6* and ended with *si*, but the final tone on these syllables was one of T1, T4, or T6.

(1) A: 汪 義 教 乜 嘢?
Wong1 Ji6 gaau3 mat1 je5?
'What does Wong Ji teach?'

B: 汪 義 教 歷 史。
Wong1 Ji6 gaau3 lik6 si2.
'Wong Ji teaches history.'

A: 汪 義 教 歷 史?
Wong1 Ji6 gaau3 lik6 si2?
'Wong Ji teaches history?'

B: 係, 汪 義 教 歷 史。
Hai6, Wong1 Ji6 gaau3 lik6 si2.
'Yes, Wong Ji teaches history.'

2.1.3 Reading task

The speakers were recorded individually in a sound-attenuated booth at the University of Calgary. They sat in front of a computer monitor, with a microphone approximately four inches from their mouths. The monitor displayed all 20 dialogues in Chinese characters, one dialogue at a time. The dialogues were randomized by block between speakers. The speakers were asked to read all four sentences in each dialogue naturally, using the same volume and speed as they would when talking with their friends. After reading all 20 dialogues, they read through the entire sequence of dialogues again. Utterances from the first reading were used to create the stimuli for testing the listeners and the model. Utterances from the second reading were intended to test the frequency effects of same-speaker exemplars on the model's classification rates (Chow and Winters 2015) and were excluded from this study.

2.2 Perception study

2.2.1 Listeners

Ten male and ten female native Cantonese listeners participated in the perception study. They were 18-35 years of age (M = 23.10, SD = 3.74) and were recruited from Calgary, Canada. They reported no history of speech or hearing disorders. For their participation in two one-hour sessions, participants received either $30 or 2% course credit.

2.2.2 Stimuli

To create the stimuli for the sentence intonation identification task, we first randomly selected two speakers of each gender from the production study. In total, these speakers had produced 160 utterances in the production task (2 speakers x 2 genders x 5 blocks x 4 dialogues x 2 sentence types). Since we wanted to test how well the model would perform without normalizing for each speaker's voice, we first verified that the F0 ranges of the selected speakers varied from each other considerably. Figure 2 shows the F0 ranges of the four speakers' production of '*Wong1 Ji6*'. These ranges are presented by sentence type, averaged over the target utterances in block A. Since the phrase *Wong1 Ji6* consisted of a high tone and a low tone, we could estimate the speakers' F0 ranges in these utterances from the phrase's maximum F0 (maxF0) and minimum F0 (minF0).

We ran six one-way ANOVAs with the maxF0, minF0, and mean F0 of each sentence type (statement and question) as the dependent measure, and with speaker as the independent factor. All of these ANOVAs revealed significant main effects of speaker [$F > 100$, $p < .001$]. Post-hoc Tukey HSD tests found significant differences in maxF0, minF0, and mean F0 of each sentence type among all four speakers ($p < .01$). These results indicated that the speakers had fairly different F0 ranges—both between themselves and between the two different sentence types.

Since the participants in the identification task were native listeners of Cantonese, they could be tested without training on the task. However, because their responses would be compared with the model's responses—and because the model needed prior experience in order to learn how to perform the task—it was necessary to train both the listeners and the model in the same way prior to testing. Therefore, half of the 80 pairs of sentences produced by the four speakers were used for training and half for testing. For

the 10 listeners of each gender, we randomized the 80 pairs of sentences five times. For five of the listeners in each gender group, the first 40 pairs of sentences in a particular randomization were used as training stimuli and the last 40 pairs were used as test stimuli. The training and test stimuli were then reversed for the other five listeners in the same gender group.

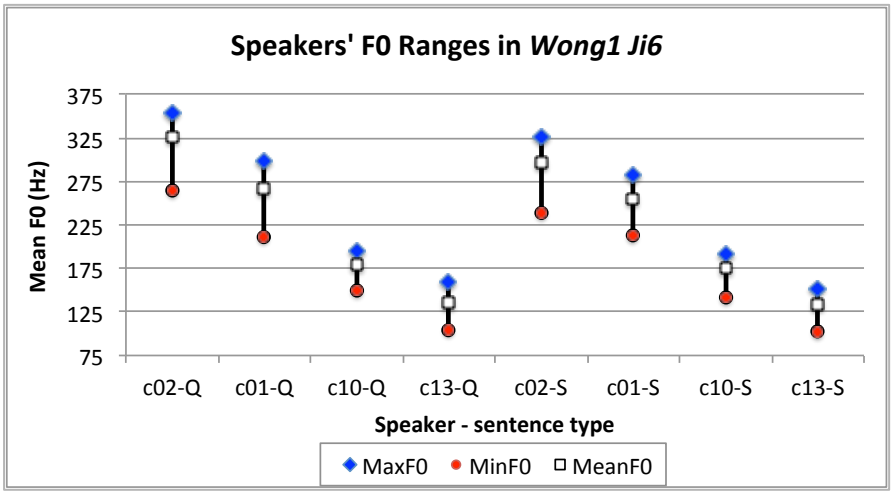

Figure 2. *F0 ranges of the female (c02 and c01) and male (c10 and c13) production of 'Wong1 Ji6', by sentence type (Q = question, S = statement)*

To test which portion of the utterance was perceptually salient (and to bring listener performance down from ceiling levels), we gated the utterances into three forms: 1) the whole utterance, 2) the final syllable, and 3) the non-final portion of the utterance. Figure 3 shows all three stimulus types (*Whole*, *Non-final*, and *Final*), gated from an echo question ending in T2, a [25] high-rising tone.

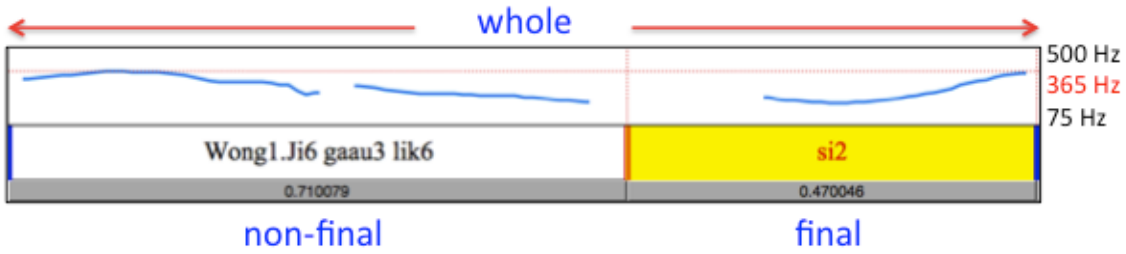

Figure 3. *Stimulus types used in the identification task: Whole, Non-final, and Final.*

2.2.3 Identification task

The identification task required two sessions. In session 1, the listeners were trained on whole sentences only, but they were tested on both whole sentences in one phase and on the non-final and final stimuli together in a subsequent phase, as shown in Table 2. The listeners also worked through two brief practice tasks, prior to testing. The sentences used

in the practice tasks differed from the sentences used in training/testing and were also spoken by a different speaker from the four speakers used in testing. Session 2 tested listeners on two additional stimulus types: the first two syllables and the last two syllables of each utterance. Since session 2 was designed for cross-linguistic comparisons, it was excluded from the Cantonese-only analysis reported here.

Part	Phase	Number of Trials	Stimulus Types
I	Practice	4	Whole
II	Training	80	Whole
III	Testing	80	Whole
IV	Practice	8	Non-final, Final
V	Testing	160	Non-final, Final

Table 2. *Phases and stimulus types used in session 1 of the identification task.*

2.2.4 Procedure

The listeners sat in a quiet room in front of a computer, wearing headphones. The stimuli were presented to the listeners one at a time. In each trial, the listeners indicated whether the stimulus they had just heard was a statement or a question (or whether it was extracted from a statement or a question, for stimulus types Non-final and Final), by pressing one of the following six keys on the keyboard: 1 = definitely a statement, 2 = likely a statement, 3 = maybe a statement, 7 = maybe a question, 8 = likely a question, and 9 = definitely a question. Only the sentence-type selection values ('statement' or 'question') will be addressed in this analysis, because the current model cannot simulate confidence ratings.

During practice and training, the correct sentence type was displayed on the screen as feedback after the listener had provided their response to each trial. During testing, the correct sentence type was not given, but the number of correct responses was displayed after every 10 trials to try to keep the listeners motivated to do the task well.

2.3 Simulation

2.3.1 Exemplar-based model

To simulate an exemplar-based process of categorizing statements and echo questions, the model first *experienced* the training stimuli by storing them as exemplars in memory according to their categories: a statement or question. Then the model processed each new token from the test stimuli by comparing it with all of the stored exemplars in each category. Through these comparisons, the model calculated whether the overall similarity value of the exemplars in the question category was greater than that of the exemplars in the statement category. If so, it classified the new token as a ‘question’ and stored it in memory as a question exemplar. If not, it classified the new token as a ‘statement’ and stored it as a statement exemplar. The stored exemplars were then used in the similarity calculations during the processing of subsequent test stimuli.

Similarity values between the new token and each of the exemplars were calculated using a simplified version of the algorithm from Nosofsky (1988) and Johnson (1997). Then the similarity values between the new token and all of the individual exemplars were summed up to derive the overall similarity value for a category. To calculate the similarity value s_{ij} between a token i and an exemplar j, the auditory distance d_{ij} between them was first calculated based on their auditory properties x_i and x_j, using the formula in (2). Then an exponential function was applied to the auditory distance between them, using the formula in (3). This step ensured that auditorily *close* exemplars would have a greater impact on the overall similarity calculation than auditorily *distant* exemplars.

(2) Auditory distance: $d_{ij} = [\ \Sigma (x_i - x_j)^2\ ]^{1/2}$

(3) Auditory similarity: $s_{ij} = \exp(-d_{ij})$

Since the goal of this study was to determine how well the model could categorize statements and echo questions based on intonation alone, the auditory properties used to calculate the auditory distance were the F0 values of the token's intonation contour, measured at 11 equidistant time-normalized points in each token. The first point began at the first voiced cycle of the token and the last point ended at the last voiced cycle of the token. Linear interpolation was used to obtain the F0 values at any of the time-normalized points that occurred in the voiceless portion(s) of the contour.

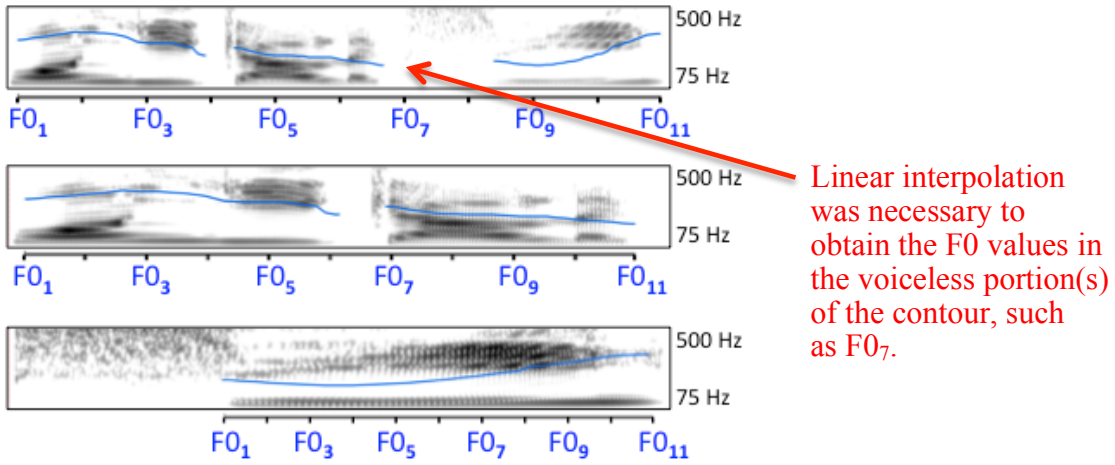

Figure 4. *F0 variables ($F0_1$..$F0_{11}$) corresponding to the F0 values at the 11 equidistant time-normalized points of the stimulus. Shown are examples for stimulus type Whole (top), Non-final (middle), and Final (bottom) for the sentence: Wong1 Ji6 gaau3 lik6 si2.*

Figure 4 displays the F0 variables corresponding to the eleven F0 values for all three stimulus types: Whole, Non-final, and Final. Using the eleven F0 values as auditory properties, the auditory distance was the Euclidean distance between corresponding F0s of the token i and the exemplar j, as in (4).

(4) $d_{ij} = [\ (F0_{1i} - F0_{1j})^2 + (F0_{2i} - F0_{2j})^2 + (F0_{3i} - F0_{3j})^2 + \ldots + (F0_{11i} - F0_{11j})^2\ ]^{1/2}$

2.3.2 Stimuli

The stimuli used for testing the model and for testing the human listeners were identical.

2.3.3 Procedure

In order to compare the model's performance with the listeners' performance on the sentence intonation identification task, we tested the model using the same stimuli in the same orders as we did with the listeners. In 10 separate runs, the model simulated the 10 tests that were assigned separately to the 10 human listeners of each gender.

2.4 Analysis

For the analysis of the listeners' performance versus the model's performance, we first converted the responses into measures of sensitivity (d') and bias (ß), based on signal detection theory (Macmillan and Creelman 2005). Then we ran ANOVAs on both d' and ß, with listener type (human, model) and stimulus type (Whole, Non-final, Final) as independent variables in each ANOVA.

3. Results

3.1 Perceptual sensitivity

We analyzed the combined responses of the human listeners and the model with a two-way ANOVA which treated d' as the dependent measure and both listener type and stimulus type as independent factors. This ANOVA revealed significant main effects of listener type [$F(1, 114) = 9.32$, $p < .001$] and stimulus type [$F(2, 114) = 193.38$, $p < .001$], and a significant interaction between listener type and stimulus type [$F(2, 114) = 14.21$, $p < .001$]. A post-hoc Tukey HSD test indicated that the human listeners performed significantly better on stimulus types Whole and Final than on stimulus type Non-final, and on stimulus type Whole than on stimulus type Final ($p < .001$). In addition, the Tukey HSD test indicated that the model performed significantly better on stimulus types Whole and Final than on stimulus type Non-final ($p < .001$). For the model, however, there was no significant difference between stimulus type Whole and stimulus type Final (at $\alpha = .05$). Comparing human listeners to the model, the Tukey HSD test indicated that the human listeners performed significantly better than the model on stimulus types Whole and Non-final ($p < .001$). There was no significant difference between the two listener groups on stimulus type Final.

Figure 5 shows that both the human listeners and the model performed well above chance on stimulus types Whole and Final, and only slightly above chance on stimulus type Non-final (human listeners: Whole = 4.2, Final = 2.6, and Non-final = .9; model: Whole = 2.9, Final = 3.0, and Non-final = .2). Since both stimulus types Whole and Final contain the final syllable (and stimulus type Non-final does not), the significantly better performances on these stimulus types suggest that the salient cue for distinguishing between statements and echo questions in Cantonese is on the final syllable. In addition, since the model's similarity calculation only used F0 values, the high d' values of 2.9-3.0 on stimulus types Whole and Final suggest that F0 is a primary cue for distinguishing between these two types of sentences. However, the fact that the human listeners

performed significantly better than the model on stimulus types Whole and Non-final suggests that the human listeners might have also used secondary cues in the non-final portion of the utterance besides just F0 (e.g., duration or intensity) to assist them in identifying the correct sentence category for the utterance.

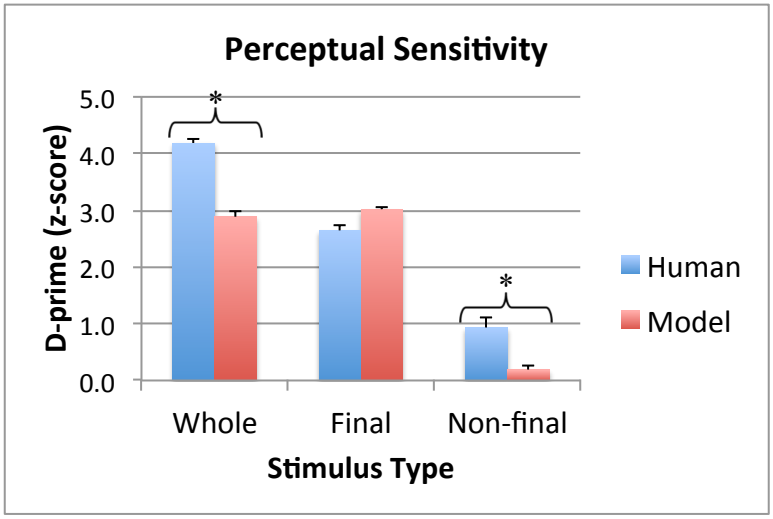

Figure 5. *Interactions between listener type and stimulus type on d' (* = p < .001).*

3.2 Response bias

We then analyzed the combined responses of the human listeners and the model with a two-way ANOVA which treated ß as the dependent measure and both listener type and stimulus type as independent factors. This ANOVA revealed a significant main effect of stimulus type [$F(2, 114) = 7.44$, $p < .001$], and a significant interaction between listener type and stimulus type [$F(2, 114) = 4.84$, $p < .001$]. A post-hoc Tukey HSD test indicated that the human listeners showed significantly more bias towards 'statement' responses on stimulus type Non-final than on stimulus types Whole and Final ($p < .001$), and on stimulus type Whole than on stimulus type Final ($p = .01$). However, the Tukey HSD test revealed no significant difference in bias in the model's responses to any of the different stimulus types. The Tukey HSD test also indicated that the human listeners showed significantly more bias towards 'statement' responses than the model on stimulus type Non-final ($p < .001$). However, the model was more biased towards 'statement' responses than the human listeners on stimulus type Final ($p = .002$). No significant difference was found between the human listeners and the model on stimulus type Whole.

Figure 6 shows that the human listeners were fairly unbiased on stimulus type Whole, but flipped from bias towards 'question' responses on stimulus type Final to bias towards 'statement' responses on stimulus type Non-final (Whole = .03, Final = .25, and Non-final = .81). The greater bias the listeners displayed towards 'statement' responses on stimulus type Non-final suggests that the statement response option may function as a default, or unmarked, category in the perception task, and that the human listeners tend to assume that an utterance is a statement until a possible pitch rise at the end of the utterance may convince them otherwise. However, the greater bias towards 'question' responses on stimulus type Final suggests that the human listeners tend to assume that an

utterance is a question if the stimulus ends with a high rising pitch, whether this rise is the result of a high final boundary tone or a high rising lexical tone. The model, for its part, was fairly unbiased on all three stimulus types, with a modicum of bias towards statement responses for the one stimulus type to which the human listeners also exhibited the greatest bias towards statements (Whole = .06, Final = .07, and Non-final = .17).

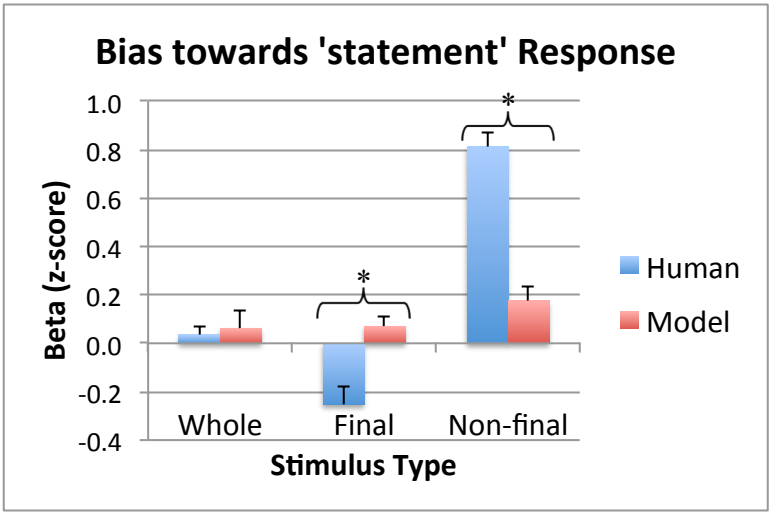

Figure 6. *Interactions between listener type and stimulus type on ß (* = p < .01).*

4. Discussion

In general, the results of this study confirmed our first hypothesis: the exemplar-based computational model was able to consistently distinguish between statements and echo questions in Cantonese. The model was unsurprisingly much better at performing this task when it was presented with the final syllable in an utterance, where the most salient prosodic cue to the statement/question distinction resides. We also learned that human listeners can, under some conditions, perform the identification task better than the computational model. Our human listeners correctly identified the sentence types of the test utterances more often than the model when presented with the whole sentence and non-final stimuli. Interestingly, however, there was no difference in performance between the two listener types for just the final syllable-only stimuli. This pattern of results suggests that the human listeners are able to extract some information out of the pre-final portions of the stimuli that the model is currently not sensitive to. The primary cue to the statement/question distinction in Cantonese is therefore in the final syllable of utterances, but this is not the only cue that Cantonese speakers rely on to make this distinction.

The results of this study are fairly consistent with the results of Ma et al.'s (2011) acoustic analysis of statement and question intonations in Cantonese. For the identification of questions, Ma et al. (2011) found that the most salient cue resides in the final syllable and that the difference in F0 height between the start and end of the voiced portion of the final syllable also provides a significant cue. However, due to differences in their speakers' F0 ranges, Ma et al. (2011) normalized the F0 measurements of the speakers prior to conducting their acoustic analysis. In this study, we did not normalize the F0 values to account for intra- and inter-speaker variation prior to categorizing the

tokens. In line with Exemplar Theory, the model was presented with the entire F0 contour without any explicit direction as to which cues or parts of the stimulus it ought to focus on in order to determine the utterance's stimulus type. Rather, specific exemplars in memory were associated with one of the two sentence type categories, and from this information alone, the model was able to make use of the relevant F0 cues and their relative position in the contour to perform the identification task. The F0 contour in the final syllable emerged as a particularly salient cue to the statement/question distinction in this simulation, without ever being represented as such in the exemplar model. In addition, the model was able to identify the salient statement and question intonation cues from the final syllable even though the final syllable appeared in different phonetic contexts: since we did not use a carrier phrase, the preceding syllable and tone varied across all target pairs.

Differences in performance between the model and human listeners also emerged in the analysis of response bias. Human listeners were significantly more biased towards 'statement' responses for the non-final stimuli than the model, and more biased towards 'question' responses for the final stimuli than the model. While this pattern may—as we mentioned above—indicate a default, or unmarked, response option for the human listeners, it is telling that these biases emerge primarily in the human responses to the partial sentence stimuli, which they are unlikely to have heard before in non-laboratory settings. Ma et al. (2011) found the same bias towards statements in their Cantonese listeners' performance on a similar statement and question identification task. The model, for its part, remained largely unbiased throughout the study, while exhibiting a small bias towards 'statement' responses in the non-final condition. Given that no bias was built into the model, this result is unsurprising; however, it is clear from the results of the human listening experiment that we will need to devise a way to incorporate bias in future, advanced versions of the model in order to better reflect how human listeners perform this intonation identification task.

It is important to note that we presented more information to the human listeners—in the form of unfiltered, natural speech stimuli—than we presented to the model, which only processed F0 contours extracted from those same speech stimuli. The ability to use other prosodic information, such as syllable duration, may partially account for the human listeners' better performance in some of the conditions of the experiment. In addition, intonation provides cues to not only sentence types, but also other information, such as focus (Liu and Xu 2006), emotion (Mozziconacci and Hermes 1999), and social identity (Bolton and Kwok 1990). Exemplar Theory would hold that this information could potentially serve as a perceptual resource to listeners performing the identification task in this study as well. Human listeners also, of course, have considerably more experience with the perception and classification of intonation in natural speech than our model had, which would also provide them with an advantage in performing the experimental task.

4.1 Future directions

In this study, we used a simple identification task to demonstrate, conceptually, that an exemplar-based computational model could categorize speech stimuli based only on intonation (i.e., F0 contours). There are, of course, more complicated aspects of intonation perception (e.g., pitch range expansion and phrase-final lengthening) that might prove to be more challenging to an exemplar-based model. It would be worthwhile

to test the model's ability to categorize more complicated prosodic structures in future research. Since intonation systems differ across languages, it would also be revealing to evaluate how the model performs with different languages. For instance, signaling echo questions with a high boundary tone is a language-specific property. Some African languages, instead, use falling intonation to signal questions (Rialland 2009). It would be helpful to apply different or additional auditory and semantic properties to the similarity calculation, in order to learn more about the salience of different cues across languages.

5. Conclusion

This study investigated how the human perception of intonation of statements and echo questions in Cantonese compares to the classification of the same statements and questions by an exemplar-based model. The model successfully classified all three stimulus types presented to it at better-than-chance rates, and closely matched the performance of the human listeners on the final syllable-only stimuli. The human listeners, however, performed significantly better on both whole sentence and non-final stimuli. The human listeners also displayed significantly more bias than the model for the final and non-final stimuli, suggesting that both frequency of experience and higher-level syntactic and pragmatic information play important roles in the human perception of intonation. While we have yet to incorporate such information into a working, exemplar-based model of intonation perception, the results of this study suggest that Exemplar Theory provides a promising approach to the study of intonation perception, because 1) the intonation patterns of sentence types are stored in rich detail in memory, 2) an exemplar-based process of categorization can account for intonation perception, and 3) no F0 normalization of speakers' voices is required prior to categorization in an intonation perception task.

References

Bauer, Robert S., and Paul K. Benedict. 1997. *Modern Cantonese phonology*. Trends in Linguistics Studies and Monographs 102. Berlin: Mouton de Gruyter.

Bolton, Kingsley, and Helen Kwok. 1990. The dynamics of the Hong Kong accent: Social identity and sociolinguistic description. *Journal of Asian Pacific Communication* 1(1): 147–172.

Brown, Gillian, Anne Anderson, Richard Shillcock, and George Yule. 1984. *Teaching talk*. Cambridge: Cambridge University Press.

Chao, Yuen-Ren. 1947. *Cantonese primer*. Cambridge: Cambridge University Press.

Chow, Una Y., and Stephen J. Winters. 2015. Exemplar-based classification of statements and questions in Cantonese. In *Proceedings of the 18th International Congress of Phonetic Sciences*, ed. The Scottish Consortium for ICPhS 2015, paper number 0987.1–5. Glasgow: The University of Glasgow.

Fant, Gunnar. 1960. *Acoustic theory of speech production*. The Hague: Mouton.

Fant, Gunnar. 1972. Vocal tract wall effects, losses, and resonance bandwidths. *Speech Transmission Laboratory Quarterly Progress and Status Report* 2(3): 28–52.

Flynn, Choi-Yeung-Chang. 2003. *Intonation in Cantonese*. LINCOM Studies in Asian Linguistics 49. Muenchen: LINCOM GmbH.

Gu, Wentao, Keikichi Hirose, and Hiroya Fujisaki. 2005. Analysis of the effects of word emphasis and echo questions on F0 contours of Cantonese utterances. In *Proceedings of the 2005 Interspeech*, 1825–1828. Lisbon, Portugal: ISCA Archive.

Hintzman, Douglas L. 1986. "Schema abstraction" in a multi-trace memory model. *Psychological Review* 93: 411–428.

Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In *Talker variability in speech processing*, ed. Keith Johnson and John W. Mullennix, 145–165. San Diego: Academic Press.

Johnson, Keith. 2005. Speaker normalization in speech perception. In *The handbook of speech perception*, ed. David B. Pisoni and Robert E. Remez, 363–389. Oxford: Blackwell.
Lehiste, Ilse, and David Meltzer. 1973. Vowel and speaker identification in natural and synthetic speech. *Language and Speech* 16(4): 356–364.
Liu, Fang, and Yi Xu. 2006. Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62(2–4): 70–87.
Ma, Joan K.-Y., Valter Ciocca, and Tara L. Whitehill. 2006. Effect of intonation on Cantonese lexical tones. *Journal of Acoustical Society of America* 120(6): 3978–3987.
Ma, Joan K.-Y., Valter Ciocca, and Tara L. Whitehill. 2011. The perception of intonation questions and statements in Cantonese. *Journal of Acoustical Society of America* 129(2): 1012–1023.
Macmillan, Neil A., and C. Douglas Creelman. 2005. *Detection theory: A user's guide*. 2nd ed. New Jersey: Lawrence Erlbaum Associates, Inc.
Mozziconacci, Sylvie J. L., and Dik J. Hermes. 1999. Role of intonation patterns in conveying emotion in speech. In *Proceedings of the 14th International Congress of Phonetic Sciences*, ed. John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville, and Ashlee C. Bailey, 2001–2004. San Francisco, CA: The University of California.
Nosofsky, Robert M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115(1): 39–57.
Nosofsky, Robert M. 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14: 700–708.
Peterson, Gordon E., and Harold L. Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24(2): 175–184.
Rialland, Annie. 2009. The African lax question prosody: Its realisation and geographical distribution. *Lingua* 119(6): 928–949.
Ryalls, John H., and Philip Lieberman. 1982. Fundamental frequency and vowel perception. *Journal of the Acoustical Society of America* 72(5): 1631–1634.
Walsh, Michael, Katrin Schweitzer, and Nadja Schauffer. 2013. Exemplar-based pitch accent categorisation using the Generalized Context Model. In *Proceedings of the 2013 Interspeech*, 258–262. Lyon, France: ISCA Archive.
Warren, Paul. 2005. Patterns of late rising in New Zealand English: Intonational variation or intonational change? *Language Variation and Change* 17: 209–230.
Wong, Wai Yi P., Marjorie K. M. Chan, and Mary E. Beckman. 2005. An autosegmental-metrical analysis and prosodic annotation conventions for Cantonese. In *Prosodic typology: The phonology of intonation and phrasing*, ed. Sun-Ah Jun, 271–300. New York: Oxford University Press.