Classification of Statement and Question Intonations in Mandarin

Fang Liu[1], Dinoj Surendran[2] & Yi Xu[3]

[1]Departments of Linguistics and Statistics, The University of Chicago
[2]Department of Computer Science, The University of Chicago
[3]Department of Phonetics and Linguistics, University College London & Haskins Laboratories
`{liufang; dinoj}@uchicago.edu, yi@phon.ucl.ac.uk`

Abstract

Conflicting reports abound in the literature regarding the critical characteristics of statement and question intonations in Mandarin. In this paper, decision trees with three different sets of feature vectors are implemented to determine the most significant elements in an utterance that signify its sentence type (statement vs. question). For 10-syllable utterances, the highest correct classification rate (85%) is achieved when normalized (to remove the effects of speaker, tone, and focus) final F_0's of the 7th and the last syllables are included in the tree construction. This performance is close to previously reported human performance (89%) for the same testing set. The results confirm the previous finding that the difference between statement and question intonations in Mandarin is manifested by an increasing departure from a common starting point toward the end of the sentence.

1. Introduction

There has been much controversy over the difference between statement and yes/no question intonation in the studies of Chinese prosody. One of the prevailing theories is that the whole pitch level is shifted upward in questions as compared to statements [5,11,17], whereas an opposing view asserts that the essential difference between the two sentence types resides only in the last word or boundary tone [6,9]. A more recent study [7] finds that the pitch contour difference between statement and question varies according to focus conditions: (1) with initial focus, question shows an overall higher pitch contour than statement (Figures 1a-4a), (2) with medial focus, the difference is manifested as a moderate raise in pitch range starting from the focused words in questions (Figures 1b-4b), (3) F_0 contours of statements and questions with final focus are similar to those with no focus (or neutral focus), i.e., showing the greatest difference in the final syllable (Figures 1c-4c and 1d-4d), and (4) across all four focus conditions, the difference in pitch between statement and question increases nonlinearly toward the end of the sentence.

Despite much research on intonations in different languages, speech engineers rarely use any of the proposed intonation models in detecting sentence types or dialog acts (e.g., statement, question, incomplete utterance, backchannel, etc.) in a corpus because it leads to little improvement [13,16]. Most often they employ as many prosodic features as possible in their implementation of decision trees to differentiate one dialog act from another. Disturbingly, removal of one set of features (e.g., F_0) can be compensated for by another, functionally different, set of features (e.g., pause) to achieve roughly the same overall accuracy [12]. Therefore, new approaches need to be explored to both improve the understanding of speech intonations and to apply intonation theories to the practice of speech recognition.
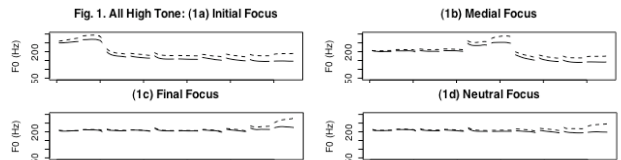

Figure 1: *Time normalized statement (——) vs. question (- - - -) with all High tones under initial, medial, final and neutral focus (ZhāngWēi dānxīn XiāoYīng kāichē fāyūn [ZhangWei worries that XiaoYing will get dizzy while driving]). F_0 contours in each plot were averaged across 40 repetitions by 8 subjects.*
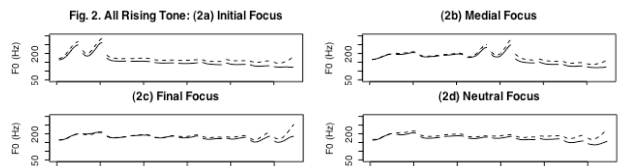

Figure 2: *WángMéi huáiyí LiúNíng huáchuán zháomí (WangMei suspects that LiuNing will get obsessed with canoeing).*
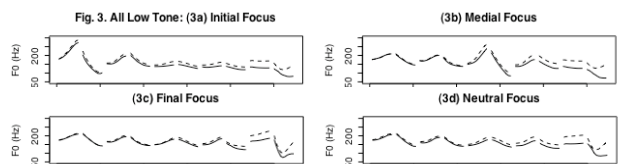

Figure 3: *LǐMǐn fǎngǎn LiǔYǔ diǎnhuǒ qǔnuǎn (LiMin dislikes LiuYu to light a fire to keep warm).*
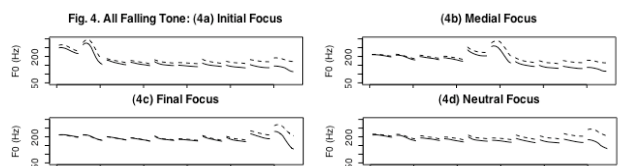

Figure 4: *YèLiàng hàipà ZhàoLì shùijiào zuòmèng (YeLiang is afraid that ZhaoLi will dream while sleeping).*

2. Method

The dataset is a subset from [7], where eight native speakers of Mandarin, 4 males and 4 females, served as subjects. It consists of four basic sentence frames, each having 10

syllables with identical tones: High (Tone 1), Rising (Tone 2), Low (Tone 3) and Falling (Tone 4), as shown in Figures 1-4. The sentences were produced either as a statement or as a yes/no question. There were four possible focus conditions for each sentence (initial, medial, final, and neutral). Each sentence was repeated 5 times by each subject. A total of 1280 sentences (= 8 subjects * 4 tone components * 2 sentence types * 4 focus locations * 5 repetitions) were thus included in the study. F_0 contours of the ten syllables in each sentence were extracted and measured using Praat [1].

To model the difference between statement and question intonations in Mandarin when the target sentences also vary in their tonal composition and focus condition, it is desirable to first extract the most representative information of the syllables in each sentence and then use this information to characterize the entire sentence. We tested the efficacy of three types of features in distinguishing questions from statements, in each case using decision trees as the classification algorithm [2].

3. Analysis

3.1. Coefficients from cubic B-spline regressions

A fixed-knot cubic B-spline regression creates a piecewise cubic polynomial within each knot span that behaves well at the peaks, because each data point affects the global fit [3]. Figure 5 displays four examples of such regression, fitted using the R [8] command $lm(F_0 \sim bs(time, df = 13))$, where $df = 13$ is to have bs place 10 (= 13 - 3 because of cubic spline) knots uniformly along the range of *time* [14].
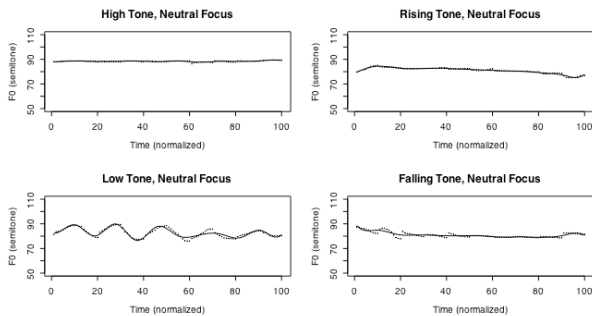

Figure 5: *Cubic B-spline regression of F_0 (in semitone) on normalized time (1-100), where dots represent original data points and lines denote fitted curves.*

Since the B-spline fitting captures the F_0 trend of a sentence reasonably well, it may be possible to use the 14 coefficients (*intercept* and *bs1-bs13*) together with *sex*, *tone*, and *focus* as the input feature vector for each sentence in constructing decision trees. The *intercept* indicates the average pitch of a sentence, and *bs1-bs13* characterize local peaks and valleys, with *bs1-bs5* supporting syllables 1-5, *bs6-bs7* supporting syllables 4-7, and *bs8-bs13* supporting syllables 6-10.

The dataset was first divided into a training and a testing set, with the former containing 960 sentences (480 statements and 480 questions) by 6 subjects (3 males and 3 females), and the latter 320 sentences (160 statements and 160 questions) by 2 subjects (1 male and 1 female). A classification tree was then grown on the training set, as shown in Figure 6.

As can be seen in Figure 6, only *tone*, *focus*, *intercept*, *bs7*, *bs8*, *bs9*, *bs12*, and *bs13* are actually used in the tree construction. Sentences are first split depending on whether

bs13 is greater than or equal to -5.688. If so, they are split according to *tone* being Falling/Low or High/Rising; if not, they are again split according to *intercept* being greater than or equal to 102.3. On the left branch, sentences with Falling/Low tones are classified as questions, of which 184 are indeed questions but 34 are actually statements; for sentences with High/Rising tones, *bs13 >= 1.505* becomes another criterion for further splitting. On the right branch, sentences having *intercept >= 102.3* are grouped into questions with probability 0.75 (=24/32); those having *intercept < 102.3* are further split according to *bs13 >=-8.793* or not.
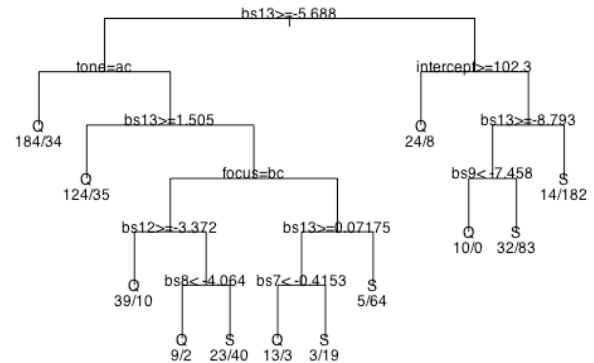

Figure 6: *Classification tree of sentence type (Q: question vs. S: statement) on sex, tone, focus, and 14 coefficients from B-spline regressions for individual sentences.*

The relationship between sentence type and the B-spline coefficients selected by the tree seems compatible with the findings in [7], since the sequential importance of *bs13, bs12, bs9, bs8* and *bs7* agrees with the fact that the difference between statement and question becomes more and more pronounced as the sentence approaches its end. Prediction of sentence type based on this tree for the testing set gives correct classification rate of 82.19% (= 263/320).

3.2. Original final F_0's of the 10 syllables

Due to articulatory constraints, the underlying pitch target of a tone is most fully realized in its final region [15]. Therefore, final F_0's of the syllables may largely represent the global pitch trend of a sentence. Pitch trajectories of individual statements and questions represented by the original final F_0 of each syllable are displayed in Figure 7. The large variability seen in the figure is due to this feature set not being normalized for speaker, tone, or focus.
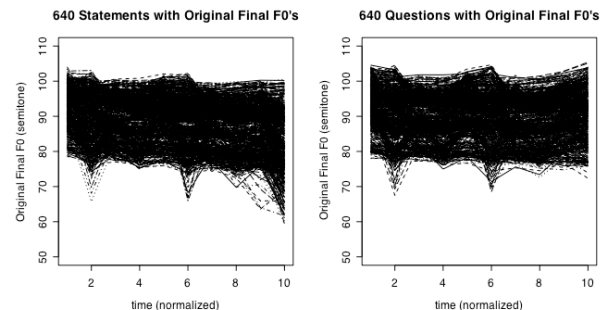

Figure 7: *Pitch trajectories of individual sentences represented by the original final F_0 (in semitone) of each syllable.*
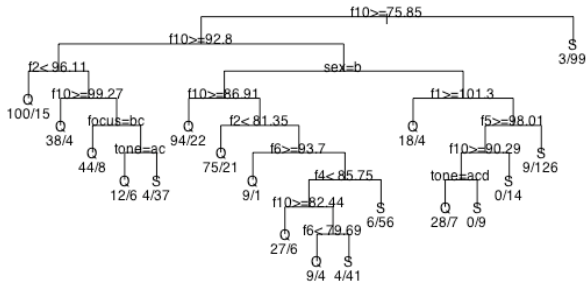
Figure 8: *Classification tree of sentence type (Q: question vs. S: statement) on sex, tone, focus, and original final F_0's (f1 – f10, in semitone).*

As can be seen from the above classification tree fitted based on the training set, all predictors except *f3*, *f7*, *f8* and *f9* are included in the tree. Sentences are first split depending on whether *f10* (final F_0 of the last syllable) is greater than or equal to 75.85. If not, they are classified as statements (with 3 of them misclassified); if so, they are again split according to *f10* being greater than or equal to 92.8. For sentences with *f10* less than 92.8, either *f10 >= 86.91* or *f1 >= 101.3* determines further splitting of the tree depending on whether they were produced by male or female speakers. For sentences with *f10* greater than or equal to 92.8, *f2 < 96.11* becomes another criterion for differentiating questions from statements.

Prediction of sentence type based on this tree for the testing set gives correct classification rate of 80.00% (= 256/320), which is slightly worse than the tree in 3.1. for which *sex*, *tone*, *focus*, and B-spline coefficients served as predictors. However, the drawback of the above two classification trees is the existence of collinearity between the factors (*sex*, *tone*, and *focus*) and the numeric predictors (*intercept*, *bs1 – bs13*, and *f1 – f10*). For example, the pitch contour of a sentence is modulated severely by the effects of *focus*: the pitch range of the focused words is expanded, that of the post-focus words compressed and lowered, and that of the pre-focus words largely unaffected. Likewise, F_0 of a syllable varies differently depending on the tone. Therefore, it is not entirely clear which mechanism (*sentence type* vs. *tone/focus*) should be considered as having taken effect when interpreting the partition rules based on the values of *bs1 – bs13* and *f1 – f10*.

3.3. Normalized final F_0's of the 10 syllables

From Figure 1d, we can see that F_0 contours of the sentences with all High tones and neutral focus reflect most directly the nonlinear increase in the difference between statement and question along the time axis. In a sense, the difference pattern there is largely free of tone and focus effects. Then, it may help to remove these potentially confounding effects by transforming the F_0 contours under other conditions toward those under High-tone and neutral-focus condition through the following normalization method.

Suppose μ_{stf} and σ_{stf} are the mean and standard deviation of final F_0's of the syllables by speaker s (1, 2, …, 8), with tone t (1: High, 2: Rising, 3: Low, 4: Falling), and under focus condition f (1: pre-focus, 2: post-focus, 3: initial/medial focus, 4: final focus), final F_0 (x_{stf}) of each syllable under such speaker/tone/focus condition (*stf*) is standardized to z_{stf} in equation (1). Note that all syllables in neutral focus sentences

are treated as pre-focus, and syllables under final focus are treated as different from those under initial/medial focus [7].

$$z_{stf} = (x_{stf} - \mu_{stf}) / \sigma_{stf} \qquad (1)$$

Then, to remove the effects of *speaker*, *tone*, and *focus*, z_{stf} with $s \neq 1$, $t \neq 1$, and $f \neq 1$ is normalized to become y_{stf}, where

$$y_{stf} = z_{stf} \bullet \sigma_{111} + \mu_{111} = ((x_{stf} - \mu_{stf}) / \sigma_{stf}) \bullet \sigma_{111} + \mu_{111} \quad (2)$$

Or, equivalently,

$$z_{stf} = (y_{stf} - \mu_{111}) / \sigma_{111} \qquad (3)$$

That is, for each speaker $s \neq 1$, assuming that the distribution of his/her syllables' final F_0's for any fixed tone (t) and focus (f) condition is Gaussian, then normalization can be viewed as mapping all syllables' final F_0's to another Gaussian distribution (in this case $s = 1$, $t = 1$, and $f = 1$, i.e., speaker 1's High-tone pre-focus syllables).
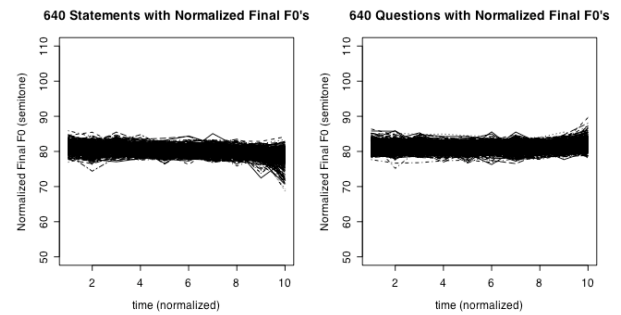

Figure 9: *Pitch trajectories of individual sentences represented by the normalized final F_0 (in semitone) of each syllable.*

The effects of speaker, tone, and focus on F_0 trajectory are largely removed in Figure 9, where statements can be seen to have a gradually falling contour, and questions a rising contour.
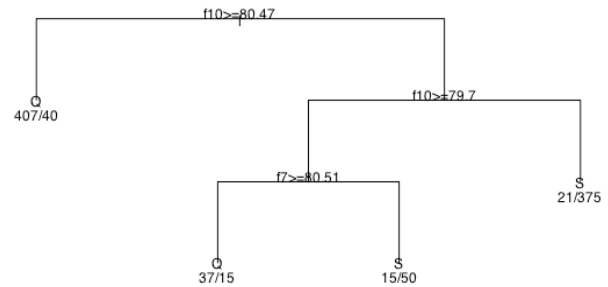

Figure 10: *Classification tree of sentence type (Q: question vs. S: statement) on normalized final F_0's (f1 – f10, in semitone).*

As shown in the above classification tree based on normalized final F_0's in the training set, among the 10 syllables in each sentence, only *f10* and *f7* are employed in the tree construction. The split on *f10 >= 80.47* partitions the 960 observations into groups of 447 and 513 individuals, with probability of question equal to 0.9105 (=407/447) and 0.1423 (=73/513), respectively. This second group is then partitioned

into groups of 117 and 396 individuals, depending on whether $f10$ is greater than 79.7 (inclusive) or not. The former group is subdivided into groups of 52 and 65 individuals, depending on whether or not $f7$ is greater than or equal to 80.51, with probabilities of question equal to 0.7115 (=37/52) and 0.2308 (=15/65), respectively. The latter group is classified as statement with probability 0.9470 (=375/396).

Prediction of sentence type based on this tree for the testing set gives correct classification rate of 85.31% (= 273/320), which is much better than that obtained in the second tree model (80.00% = 256/320), where *sex*, *tone*, *focus*, and original $f1 - f10$ are predictors.

4. Discussion and Conclusion

As a screening method for selecting predictors, classification trees were used to extract the most useful information in an utterance that characterizes its sentence type. Three different sets of feature vectors were input into the tree and the corresponding results are summarized as follows.

Table 1: *Summary of the three classification trees.*

Decision trees	Variables used	Correct classification rate in testing set
Coefficients from B-spline regressions	focus, tone, intercept, bs7, bs8, bs9, bs12, bs13	82.19% (=263/320)
Original final F_0's	focus, tone, sex, f1, f2, f4, f5, f6, f10	80.00% (=256/320)
Normalized final F_0's	f7, f10	85.31% (=273/320)

The coefficients from B-spline regressions are reasonably good in classifying statement and question in Mandarin, as demonstrated by the 82.19% correct classification rate. However, the fact that they performed only slightly better than non-normalized original final F_0's of syllables shows that direct representation of detailed surface F_0 has limited benefit for sentence type classification. In contrast, the much larger improvement brought about by the normalization method (85.31%) demonstrates the importance of directly taking into account the effects of tone and focus. This performance is only slightly worse than the human performance (89.12%) reported in [7] for the speech of the same two subjects in the testing set here.

In all the classifications, the parameters corresponding to the sentence-final F_0 are found to be the dominant factor for determining sentence type. Nevertheless, F_0 before the final syllable are also consistently found to be relevant. This agrees with the previous finding that, across all tone and focus conditions, statement and question intonation are characterized by a nonlinear fall and a nonlinear rise toward the end of the sentence.

The decision trees in the present study are grown on a dataset in which statements and questions are elicited under laboratory conditions with tone and focus systematically controlled. The performances therefore are not equivalent to those on natural speech databases. Importantly, in natural speech, many questions do not have rising intonation while many statements do [4,10], which present an issue that is beyond the scope of the present study. In those cases,

however, it is also an open question whether human listeners can identify the questions and statements that are taken out of context. What the current results show is that for those cases that human listeners can make the identification, syllable final F_0 with normalization can achieve similar performance.

5. References

[1] Boersma, P.; Weenink, D., 2005. Praat: doing phonetics by computer (Version 4.3.19) [Computer program]. Retrieved from http://www.praat.org/.

[2] Clark, L. A.; Pregibon, D., 1992. Tree-based models. In *Statistical Models in S*, Chambers J. M.; Hastie, T. J. (eds.). Wadsworth & Brooks, 377-419.

[3] Hastie, T.; Tibshirani, R.; Friedman, J., 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[4] Hirschberg, J., 2000. A corpus-based approach to the study of speaking style. In *Prosody: theory and experiment. Studies presented to Gösta Bruce*. M. Horne. Kluwer Academic Publishers**,** 335-350.

[5] Ho, A., 1977. Intonation variations in a Mandarin sentence for three expressions: interrogative, exclamatory, and declarative. *Phonetica* 34, 446-456.

[6] Lin, M., 2004. On production and perception of boundary tone in Chinese intonation. In *Proceeding of. International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing, 125-129.

[7] Liu, F.; Xu, Y., 2005. Parallel transmission of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62: 70-87.

[8] R Development Core Team, 2005. R: A language and environment for statistical computing. ISBN 3-900051-07-0, URL http://www.R-project.org.

[9] Rumjancev, M. K., 1972. *Ton i intonacija v sovremennom kitajskom jazyke* (*Tone and Intonation in Modern Chinese*). Izdatel'stvo Moskovskogo Universiteta, Moscow.

[10] Shattuck-Hufnagel, S.; Turk, A. E., 1996. A Prosody Tutorial for Investigators of Auditory Sentence Processing. *J. Psycholinguistic Research* 25(2), 193-247.

[11] Shen, X., 1989. *The Prosody of Mandarin Chinese*. University of California Press, Berkeley.

[12] Shriberg, E.; Bates, R.; Stolcke, A.; Taylor, P.; Jurafsky, D.; Ries, K., Coccaro, N.; Martin, R.; Meteer, M.; Van Ess-Dykema, C., 1998. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech* 41(3-4), 439-487.

[13] Taylor, P.; King, S.; Isard, S.; Wright, H., 1998. Intonation and dialogue context as constraints for speech recognition. *Language and Speech* 41(3-4), 493-512.

[14] Venables, W. N.; Ripley, B. D., 2002. *Modern Applied Statistics with S*. Springer.

[15] Xu, Y.; Wang, Q. E., 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* 33, 319-337.

[16] Yuan, J.; Jurafsky, D., 2005. Detection of Questions in Chinese Conversational Speech. To appear, *Proceedings of. IEEE Automatic Speech Recognition and Understanding Workshop*, Cancun, Mexico.

[17] Yuan, J.; Shih, C.; Kochanski, G.P., 2002. Comparison of Declarative and Interrogative Intonation in Chinese. *Speech Prosody 2002*, Aix-en-Provence, France, 711-714.