Perception of English Prominence by Native Mandarin Chinese Speakers

Andrew Rosenberg[1], *Julia Hirschberg*[2], *Kim Manis*[3]

[1]Computer Science Department, Queens College CUNY, New York, USA
[2]Computer Science Department, Columbia University, New York, USA
[3]Microsoft, Washington, USA
andrew@cs.qc.cuny.edu, julia@cs.columbia.edu, kimani@microsoft.com

Abstract

Native-like perception of intonational prominence is important for spoken language competency. Non-native speakers may have trouble interpreting prosodic variation in a second language like English, where intonational variation can critically influence utterance semantics. By identifying types of prosody non-native learners find difficult to perceive, we can improve our ability to teach L2 speakers a language. In this paper we present results of a perception study in which Mandarin speakers with knowledge of English were tested on their ability to identify prosodic prominence in English in a variety of contexts. Through this analysis we identify particular contexts which make it difficult for Mandarin speakers to recognize pitch accent in English.

Index Terms: prosody, pitch accent, intonational prominence, perception, non-native speech

1. Introduction

While the importance of acquiring native-like prosody has long been noted as an important but often neglected part of Second Language (L2) Learning (cf. [1]), there has been but little attempt to incorporate training in prosodic variation into online language tutoring systems (cf. [2, 3, 4, 5]). However, in languages such as English, failing to recognize prosodic variation appropriately can have major consequences for semantic interpretation. For example, *John only introduced* **Mary** *to Sue* differs from *John only introduced Mary to* **Sue**, for native speakers of Standard American English (SAE) in terms of how many people who were introduced to Sue and to Mary.

In this paper, we present results of a study designed to test the ability of native speakers of Mandarin to recognize intonational prominence, or (*pitch accent*[1]) in English. In this study, native speakers of Mandarin Chinese were asked to identify all prosodically prominent words in a short utterance of Standard American English via a web-based interface that varied in location of pitch accent, part-of-speech (POS) and length of words (in syllables), and pitch contour over the utterance. Our goal is to design language tutoring systems that are better able to address the particular needs of these L2 speakers. In Sections 2 we describe the experiment. The analysis and results are presented in Section 3. Section 4 contains an error analysis of the subject responses. We conclude and discuss future work in Section 5

[1]We assume the ToBI convention for representing prosodic events in SAE [6] .

Roman women rarely marry
Roman women marry rarely
Rarely roman women marry
Rarely marry roman women
Marry rarely roman women
Marry roman women rarely

Figure 1: *The six orderings of two-syllable words.*

2. The Experiment

2.1. Materials

The stimulus materials were 144 four-word utterances recorded by a native speaker of Standard American English (SAE). Two sets of four words (one consisting of one-syllable words — ALL, MEN, NOW, and RUN — and one of two-syllable words — ROMAN, WOMEN, RARELY, and MARRY) were varied systematically in terms of utterance position. To decouple the influence of position and syntactic function, we varied the word order of these four words with the constraint that the determiner must precede the noun. This resulted in 6 different orderings for each set: 1) Det Noun Adv Verb, 2) Det Noun Verb Adv, 3) Adv Det Noun Verb, 4) Adv Verb Det Noun, 5) Verb Adv Det Noun and 6) Verb Det Noun Adv. Figure 1 shows the two-syllable stimuli.

Note that each utterance was fully sonorant, for ease of prosodic analysis. The utterances were produced with three different intonational contours, each of which contain a single accented word; these were 1) H* L-L% (a standard 'declarative' contour; 2) H* H-H% (a high-rising contour) and 3) L* H-H% (a standard *yes-no* question contour. The position of the accented word was systematically varied across each of the four word positions. Thus we produced utterances of 2 syllable lengths, 6 word orders, 4 prominence positions, and 3 intonational contours, for a total of (2*6*4*3) 144 tokens. The stimuli were pre-tested by two native speakers of SAE with ToBI labeling experience to ensure that the contours and accented words were easily identifiable by L1 speakers.

2.2. Subjects

Our subjects were 12 native speakers of Mandarin Chinese, 6 male and 6 female, with no reported hearing problems. They were recruited from the Columbia University community. Subjects varied in age, gender, experience speaking English, and length of time living in an English-speaking country. They had a mean age of 26.75 years. Their mean reported length of experience speaking English was 16.75 years, with a maximum of 25 years and a minimum of 8 years. Each subject reported being comfortable speaking English, though 5 (3 male, 2 female) had lived in an English-speaking country for less than 6 months.

2.3. Procedure

Subjects participated in the study in the Columbia Speech Lab. The test was administered through a web interface, with subjects listening to stimuli through headphones and indicating their decisions on transcripts presented with the aural stimulus. For each stimulus, subjects clicked an icon to play the utterance for the first time. The text of the utterance was visible in grey during the presentation of the stimulus. After the first aural presentation, the text turned black and subjects were asked to "underline the word or words that are prominent by clicking on them". Selected words were highlighted in yellow and underlined; they could be deselected by clicking a second time. (While only one word in the stimuli was deemed by L1 speakers to be accented, subjects were allowed to selected as many words as they wished.) Subjects could replay the stimulus as many times as they liked. When the subject was satisfied with their answer, he or she clicked a button labeled "Next" to proceed to the next token. Before the actual experiment, subjects were given a short training session using the same procedures and could ask any questions during this period. We record subject responses, timing information and the number of replays.

3. Subject Score Analysis

Since subjects were allowed to mark any number of words in an utterance as intonationally prominent, the response to each utterance consisted of four binary decisions – one for each word. We calculate a subject's score for each token as the proportion of correct decisions, leading to a score between 0 and 1 for each subject response. Pitch and intensity contours for each token are calculated using Praat [7]. Correlations with subject scores are calculated using linear regression for continuous variables and t-tests for categorical variables, using R [8].

3.1. Demographic Analysis

We first examine the effects of gender, age, and experience speaking English on subject performance. We find that age and experience both *negatively* correlate with subject scores. Interestingly, older subjects and those reporting more experience speaking English are *less* successful in detecting prominent words, with $p = 1.354 * 10^{-5}$ and $p = 1.640 * 10^{-8}$, respectively. It is possible that increased exposure to the language leads subjects to rely on other information streams such as semantic and pragmatic context in their perception of prominence. Those speakers with less experience, and perhaps less entrenched intuitions about which word or words *ought* to be prominent may be more successful in making judgments purely based on the acoustics of the utterance. We also find a relationship between gender and scores that approaches significance ($p = 0.0624$). Male subjects had a mean score of 0.847, while female subjects had a mean score of 0.826. Both male and female subjects had a nearly identical mean age (M: 26.83, F:26.67) and mean amount of experience speaking English (M:16.83 years, F: 16.67). However, the female subjects had, on average, spent twice as much time in the United States as the males at the time of the study – five years compared to two and a half years. This difference may account for difference in performance similarly to the negative effect of greater reported experience speaking English.

3.2. Analysis of Syntactic and Positional Qualities

In terms of the contexts in which prominent words appeared, subjects are significantly more successful in identifying prominence in utterances that contained two syllable words (0.873) than those containing one syllable words (0.799). We hypothesize that the presence of three types of syllables in these stimuli (lexically stressed and accented, lexically stressed and deaccented, and not bearing lexical stress) provides more contrast for the successful identification of accent.

The POS of the accented word shows a significant influence on subject scores, with $p = 7.404 * 10^{-4}$. Subjects are better able to identify accented adverbs (mean score = 0.868) or determiners (0.842) than verbs (0.833) or nouns (0.802). It is surprising that perception of prominent nouns is more difficult for non-native speakers than any other POS. Perhaps the accenting of other parts-of-speech is less expected, and therefore it is more noticeable when these words are made intonationally prominent. The position of the accented word in the utterance also shows a strong interaction with subject scores with $p < 2.2 * 10^{-16}$. Subjects are much more successful at recognizing accented words at the end of an utterance than at the beginning. In first position, the mean score for accent detection is 0.772; in position two, 0.815; in three, 0.858; and when the final word is prominent subjects have a mean detection score of 0.900. This monotonic increase in accuracy may be due to a tendency for final content words to bear accent in SAE (cf. [9]) – subjects may have come to expect that a phrase final word will be accented, or this may be due an interaction with the phrase ending intonation. We address this possibility in greater detail in Section 4. While POS and utterance position both interact with subject's perceptions of prominence, word order shows no significant interaction ($p = 0.439$). Recall that the word order of the stimuli (cf. Figuredetermines the syntactic structure of the utterance. However, type of intonational contour also significantly correlates with subject scores with $p \leq 2.2 * 10^{-16}$. In particular, prominence in declarative contours is easier for L2 subjects to correctly identify, with a mean score of 0.961. Prominence in the H* H-H% (high rising) contour is the most difficult to identify, with a mean score of 0.740, while L* H-H% (question) contours has a mean detection rate of 0.808. Curiously, the *type* of pitch accent (H*) is the same for the easiest and most challenging contours to judge. We also note a learning effect; over the duration of the study, subject scores significantly improve ($p = 0.0364$). This suggests that being asked to pay attention to prominent words can lead to an improved prominence detection, and thus that training in this area can be useful.

3.3. Analysis of Acoustic Qualities

We also examine acoustic correlates of subjects' prominence detection, to see which parameters subjects appear to be attending to in making their judgments. We compute average and maximum intensity and pitch for the accented word in isolation and for the whole utterance. Not surprisingly, higher average and maximum pitch values within the accented word lead to higher subject scores, with $p < 1.348 * 10^{-6}$ and $p < 2.2 * 10^{-16}$, respectively. Average and maximum pitch over the whole phrase are negatively correlated with subjects scores, with $p < 2.2 * 10^{-16}$ for both, indicating that higher pitch within the accented word and lower pitch for the rest of the sentence leads to higher contrast and, thus, easier prominence detection. Linear regression of intensity values with subject scores shows that a higher average intensity, in both the accented word and the whole phrase, leads to fewer correctly

accented words, with $p < 5.279 * 10^{-10}$ and $p < 2.2 * 10^{-16}$, respectively. Maximum intensity of the accented word does *not* significant correlate with subject scores ($p = 0.5561$), while maximum intensity of the whole phrase does ($p = 0.03687$). These findings indicate that, the louder the phrase is overall, the more difficult it becomes to identify accented words. We also examine the ratios of the average and maximum values of pitch and intensity drawn from the accented word to those calculated over the entire phrase. For pitch, both the average and maximum ratios show a statistically significant correlation with subject scores ($p < 2.2 * 10^{-16}$ for both). The ratios for intensity are also significantly correlated with the subjects' accuracy with $p = 9.197 * 10^{-05}$ for ratio of averages and $p = 0.00168$ for ratio of maxima. The duration of the accented word and the ratio of the accented word length to the whole phrase length also show a positive correlation with subject accuracy, with $p < 2.2 * 10^{-16}$ and $p = 3.862 * 10^{-15}$, respectively. The longer the accented word, the easier it is to identify as accented. All of these findings are plausible: increased pitch, intensity and duration relative to the surrounding utterance are known correlates of prominence in Standard American English. However, it is interesting to note that L2 speakers are indeed attending to the same acoustic correlates of accent that L1 speakers use to produce their native prosody.

4. Misses vs. False Alarms

In Section 3, we calculate a subject score for each stimuli based on the ratio of correct prominence decisions to the total number of words in the utterance – four. In this section, we treat *misses* and *false alarm* (FAs) as distinct errors. We discuss the contexts which are most difficult for non-native speakers to perceive prominence as native speakers do by describing tokens that led subjects to generate misses vs. FAs. Misses occur when a subject does not mark the correct word as accented. FAs occur when a subject marks a deaccented word as accented. For each utterance, we calculate miss and FA rates and identify syntactic, structural and acoustic correlates of this value.

We first note, not surprisingly, that misses and FAs are strongly correlated. Using a linear regression, we find a correlation coefficient, $r = 0.900$, with an associated p-value, $p < 2*10^{-16}$. 81.54% of all subject responses marked only one word as accented. On these tokens, any error would be manifested as *both* a miss and a FA. However, the FA rate is notably higher than the miss rate. Across all tokens, 35.8% contain a FA error, while 21.6% contain a miss error. Moreover, 97.1% of all miss errors are associated with a FA. Only 11 of 1728 responses contain no annotation of prominence; thus, each other miss error must have an associated FA. On the other hand, 58.6% of false alarms are associated with a miss.

4.1. False Alarms

In this section we examine the contexts that give rise to false alarms. We find that significantly more errors occur on utterances constructed with one syllable words (FA rate: 0.424) than on two syllable words (0.293), $p = 0.00810$. We also find that contour type has an effect: declarative tokens (H* L-L%) show the lowest rate of FAs, with 0.116, while high-rise (H* H-H%) utterances had a rate of 0.528 and L*H-H% contours had an error rate of 0.431. This effect of contour type is significant with $p = 7.54 * 10^{-14}$. The distributions of accent annotations by word position and contour type are shown in Table 1. It is remarkable that more than half of the high-rise utterances were incorrectly marked as accented by our Mandarin subjects. Examining the rising tokens – both L* H-H% and H* H-H% – we find a marked increase in the number of phrase-final words which are marked as prominent. This is an effect we did not observe in the pretest native SAE speakers. For the high rise utterances, 54.0% of subject responses indicated that the final word was prominent. Recall that in each stimulus utterance only one of four words (25%) is actually prominent. If we look at the distribution of accent annotations on rising contours compared to declarative utterances, we can see a clear interaction between contour type and prominence perception. It appears that phrase-final tones appear to lead to confusion in prominence detection more than the type of pitch accent: The rate of FAs on rising contours is much higher than for falling contours while there are no clear differences between contours containing H* accents in contrast to those containing L* accents.

Contour	1	2	3	4
H* L-L%	28.3%	28.3%	25.9%	26.2%
H* H-H%	23.3%	27.1%	27.1%	**54.0%**
L* H-H%	26.2%	27.3%	25.3%	**47.7%**

Table 1: *Distribution of accent annotations by contour type and word position*

We also find that the position of the accented word in the utterance has a effect on the rate of FAs. Significantly more words are incorrectly marked as accented when the first word is actually prominent. Across all contours, when the accented word comes first, 51.9% of responses include a FA. This rate drops to 35.4% and 37.7% when the second or third word is accented, and only 18.3% when the utterance final word is prominent. This difference is significant with $p = 1.88*10^{-5}$. Recall that the false alarm rate is lower for declarative utterances than for utterances produced with rising intonation (L* H-H% or H* H-H%). Despite this *overall* lower rate of false alarms, the influence of accent position on FA rate is still observed. In declarative utterances, the FA rate when the first word is accented is 19.4%, 10.4% when the second word is accented, 6.3% on the third word and 10.4% when the last word is prominent.

Contour type and accent position show an independent effect on the FA rate, and we also observe a *combined* effect of these two features. This combined effect is significant with $p = 0.00404$ under a multi-variate ANOVA. The FA rate varies less significantly based on accent position in declarative contours. For high-rising contours, the FA rate when only the utterance final word is prominent is 25.7%, but leaps to 70.14% when only the first word is accented.

Thus, Mandarin speakers appear to be more likely to incorrectly identify words as accented if they occur on one-syllable words, on utterances with rising contours, especially for utterance-final words, and on utterances whose first word is accented.

4.2. Acoustic Correlates of Misses and False Alarms

The acoustic correlates of misses and false alarms are, for the most part, identical to the correlates of subject score which we identified in Section 3. There is, however, one acoustic quality on which the correlation with miss rate and FA rate differs.

When we examine the correlation between the standard deviation of the intensity of the prominent word and the subject score on the token, we find a significant positive correlation ($p = 4.8 * 10^{-5}$). This is paired with a significant negative correlation with the miss rate ($p = 0.00411$). However, this

acoustic feature does not correlate significantly with FA rate ($p = 0.133$). These findings together suggest that a narrowly varying intensity on a prominent word may make it less likely to be recognized as accented by a non-native speaker, while not making it significantly more likely that *another* token will be perceived as accented. All of the other examined acoustic qualities correlate identically with misses and FAs – the difficulty in recognizing one token as accented is matched by the likelihood of recognizing an unaccented word as prominent. However, FAs occur with some regularity without corresponding misses. Even though a word may not be missed outright by a particular subject, the same features that make a word more likely to be missed make it more likely that a subject will generate a FA. This feature, the standard deviation of intensity, was the only which we observed **not** to follow this relationship.

4.3. Two Categories of False Alarms

We observed in Section 4.1 that miss and FA rates are tightly correlated. While nearly all misses are associated with FAs, the converse does not hold. A FA may occur when a subject marks an incorrect word as prominent *instead* of the correct word. This FA error will thus coincide with a miss error. An FA may also occur when a subject marks the correct word as prominent, and *also* marks another word in the utterance is accented. In this section we compare these two errors. We compare the 363 responses in which FAs occur in isolation – *additional* annotation – with those 256 cases where the FA is coincident with a miss error – *replacement* annotation. In the previous analyses, we collapsed subject responses into an aggregate score, either "subject score" or an error rate. Here we examine each rating separately, examining only FAs, and identifying differences between *replacement* and *additional* FAs.

We find that there are significantly more replacement FAs than expected on one-syllable utterances, but more additional FAs on two-syllable utterances ($p=1.62 * 10^{-6}$). This supports our previous hypothesis that it is more difficult for non-native speakers to detect prominence on monosyllabic tokens or in monosyllabic contexts than in utterances where some words contain non-lexically stressed syllables. We again see an influence of contour type and accent position on perception errors. We find fewer than expected replacements on utterances produced with a declarative contour (H* L-L%) and more than expected replacement on H* H-H% contours ($p= 1.261 * 10^{-7}$). This suggests that in declarative contours subjects are better able to perceive the correct prominent word, though they may also be prone to perceiving erroneous additional words as being prominent. On the other hand, phrase-ending intonation – phrase accent and boundary tone – may lead non-native listeners to miss the prominence of the accented word. We find that there are also more additional FAs than expected when the first, second or third word is accented, and more replacements than expected when the final word is accented ($p = 0.0002957$). This suggests that subjects are more likely to perceive an accent incorrectly on the final word. When the final word is accent-bearing, false-alarms are more likely to be replacement errors.

We also examine whether there are observable acoustic qualities that make an utterance more likely to have an additional FA than a replacement FA. Broadly speaking, those tokens which give rise to additional FAs tend to have greater excursions on the accented word. When the accented word has more modest excursions or dynamics, the error tends to be a replacement FA – perceive another word as prominent *rather* than the correct word. We find a positive correlation between addition FAs and the maximum pitch of the accented word ($p = 0.0003472$), the standard deviations of pitch ($p = 3.23 * 10^{-6}$) and intensity ($p = 3.98 * 10^{-06}$), and the ratio of the maximum pitch in the accented word to the maximum pitch in the utterance ($p = 3.04 * 10^{-9}$). We also find that increased duration correlates with replacement FAs whether measured in seconds ($p = 1.33 * 10^{-9}$) or as a ratio with the average word length in the utterance ($p = 5.90 * 10^{-7}$).

5. Conclusions and Future Work

In this paper we have discussed results of a perception study in which native speakers of Mandarin were asked to identify intonationally prominent, or, accented, words in English. We found negative effects of age and experience speaking English on subjects' scores and a positive effect of gender on score that approaches significance, although female subjects did have on average twice as much time living in an English speaking country than their male counterparts. With respect to features of the accented words identified, we found effects of word length in syllables, part-of-speech, word position in utterance, and type of intonational contour on subject performance. We also noted that subjects became more accurate at identifying prominence over the course of the session. We also found influences of pitch and intensity (particularly when accented words were in high contrast to the matrix utterance), as well as duration, on words subjects judged to be accented; these findings suggest that non-native speakers are using similar criteria to native speakers in making their judgments. We also discuss cases in which accented words were *not* recognized as such vs. cases in which deaccented words were deemed to be accented, finding that word length, position, and contour had significant effects, while pitch accent type did not. In sum, we have targeted particularly contexts which make it difficult for Mandarin speakers to recognize pitch accent in English, which should prove useful for future training of Mandarin learners of English.

6. References

[1] A. Wennerstrom, *The Music of Everyday Speech: Prosody and Discourse Analysis*. Oxford University Press, 2001.

[2] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Islands of failure: Employing word accent information for pronunciation quality assessment of english l2 learners," in *SLaTE*, 2009.

[3] R. Hincks and J. Edlund, "Using speech technology to promote increased pitch variation in oral presentations," in *SLaTE*, 2009.

[4] N. Cylwik, A. Wagner, and G. Demenko, "The euronounce corpus of non-native polish for asr-based pronunciation tutoring system," in *SLaTE*, 2009.

[5] M. Duong and J. Mostow, "Detecting prosody improvement in oral rereading," in *SLaTE*, 2009.

[6] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *Proc. of the 1992 International Conference on Spoken Language Processing*, vol. 2, 1992, pp. 12–16.

[7] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9-10, pp. 341–345, 2001.

[8] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[9] A. Rosenberg, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.