Production of English Prominence by Native Mandarin Chinese Speakers

Andrew Rosenberg[1], *Julia Hirschberg*[2]

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

Abstract

Native-like production of intonational prominence is important for spoken language competency. Non-native speakers may have trouble producing prosodic variation in a second language (L2) and thus, problems in being understood. By identifying common sources of production error, we will be able to aid in the instruction of L2 speakers. In this paper we present results of a production study designed to test the ability of Mandarin L1 speakers to produce prominence in English. Our results show that there are some consistent differences between the L1 and L2 speakers in the use of pitch to indicate prominence, as well as in the accenting of phrase-initial tokens. We also find that we can automatically detect prominence on Mandarin L1 English with 87.23% and an f-measure of 0.866 if we train a classifier with annotated Mandarin L1 English data. Models trained on native English speech can detect prominence in Mandarin L1 English with an accuracy of 74.77% and f-measure of 0.824.

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

1. Introduction

While prosodic variation is a key method of conveying meaning in English, it is rarely taught in Second Language (L2) instruction (cf. [1]). In languages such as English, failing to produce appropriate prosodic variation can lead to unintended semantic or pragmatic interpretations. In this work, we identify similarities and differences in the accenting behavior of native Standard American English (SAE) speakers and native Mandarin Chinese (MC) speakers. We present results of a production study designed to test the ability of native speakers of Mandarin to produce intonational prominence (*pitch accent*[1]). There have been few attempts to include instruction in native-like prosodic variation in online language tutoring systems. (cf. [3, 4]). Our ultimate goal is to create tutoring systems which can train students learning English in prosodic variation — particularly accenting behavior — targeting those aspects of prosodic variation that are most difficult for the language group being tutored.

In our production study, MC L1 speakers were asked to read news stories written in English. We then compare the intonation the subjects used to that of native SAE speakers reading the same material. In Section 2 we describe the details of the production study. The analyses of prominent tokens in non-native speech is presented in Section 3. We next describe the results of experiments detecting prominence in non-native English automatically in Section 4. In Section 5 we summarize our results and discuss future work.

2. Materials

We selected two news stories drawn from the Boston University Radio News Corpus (BURNC) [5] for subjects to read for comparison with the original BURNC SAE native speakers. The BURNC corpus is is a corpus of professionally read radio news data. The *labnews* portion of the corpus includes laboratory recordings from six speakers reading four stories each. We selected two of these *labnews* stories for the Mandarin speakers to read. These were: **p** — computerized parole officers — and **r** — the Safe Roads Act. Story **j** — Massachusetts Supreme Court Justice contains a high rate of proper names. These names caused difficulty for pretest subjects leading to the decision to omit this story from the production study. We also decided not to use labnews story **t**, as its subject was teen pregnancy, sex and contraception. We were concerned that this topic might make some subjects uncomfortable, modifying their intonation in unexpected ways. Each subject was also asked to read an introductory transcribed broadcast news paragraph concerning NASA and a delayed spacecraft launch, chosen to acclimatize subjects to the task and to the recording environment.

Our subjects were 4 native Mandarin Chinese speakers, two male and two female. No subjects reported any hearing problems. At the time of recording, the four speakers of the annotated material were between 25 and 30 years old, with 6 to 19 years of experience with English; they had spent from 7 months to 3 years living in the United States.

Subjects were asked to read all materials in a sound-proof booth in the Columbia Speech Lab and were recorded using a Tascam HD-P2 solid state recorder at 16 bit with a 44.1kHz sampling rate and a Crown CM-311 headset microphone. All but the introductory paragraph was orthographically transcribed. Disflunecies were annotated and non-standard pronunciation and lexical stress placement was also noted. Prosodic annotation of this material was performed using the ToBI standard [2] by one of the authors. The annotated material comprises 37.6 minutes of speech. Throughout this paper we describe the BURNC material spoken by native speakers of Standard American English as SAE material. The collected material produced by native speakers of Mandarin Chinese will be referred to as the MC corpus.

3. Analysis

We conduct a number of analyses of the MC productions for comparison with the SAE data. We examine the rate of accenting used by native vs. non-native speakers. We also assess the two groups' similarity in other dimensions of accenting behavior: 1) positional (the location of accented words in a phrase); 2) syntactic distribution (the distribution of accents by POS); and 3) acoustic properties of L1 vs. L2 speakers' accents.

[1] Throughout, we define prosodic events such as pitch accent in the ToBI framework [2].

3.1. Analysis of Accent Rate

We first compare the overall accent rates of native and non-native speakers when reading the same BURNC material. We find that the native speakers produce 51.3% (3218/6277) of words in this corpus with pitch accents, while non-native speakers accent more frequently at 61.9% (2903/4689). A proportion test indicates that this difference is significant with $p < 0.00001$. This finding may be evidence of some degree of *hyperarticulation* on the part of the non-native speakers, perhaps related to slowness and uncertainty in production. The higher accenting rate may also reflect the shorter intonational phrase length consistently observed in non-native speech.

The ToBI intonational standard describes phrase structure — the division of speech into meaningful contiguous units — as a two-tiered hierarchical system. The larger prosodic unit is the *intonational phrase,* in the ToBI framework. These phrases are separated from one another by the greatest degree of perceived disjuncture. This disjuncture is commonly realized by the presence of silence, acoustic reset — the raising of pitch and intensity at the start of a subsequent phrase — and pre-boundary lengthening — segmental durational increases immediately prior to the phrase boundary. Each intonational phrase contains at least one *intermediate phrase* plus a *boundary tone*. Intermediate phrases are distinguished by some of the same features as intonational phrases, though the disjuncture between them is less pronounced and there is less tonal marking. Silence is rarely observed between intermediate phrases and reset and lengthening are less dramatic. In our data, the mean *intermediate phrase* length for native speakers is 3.87 words compared to 2.55 words for non-native speakers. The difference in *intonational phrase* length is still greater: 6.16 words for native speakers, and 3.83 for non-native speakers.

By removing disfluencies from native and non-native speech, we are able to align the material spoken by the ten speakers. This allows us to identify those tokens which are consistently made prominent by native speakers and those which consistently do not bear accent. The analyses in this section is performed on this aligned material with disfluencies removed. On this ’cleaned’ data, the accent rates are not greatly changed: 51.3% on native speech and 63.11% on non-native speech.

We next examine how consistently the two groups of speakers make tokens prominent. We make the simplifying assumption that the native speakers all produce natural, fluent intonational patterns. Therefore we identify three classes of tokens — tokens that *every* SAE speaker accented, tokens that *some but not all* did and tokens *none* did. We also examine whether the non-native speakers are consistent in their use of prominence. We divide the non-native tokens into three similar classes: those that are *always*, *never* and *sometimes* accented by non-native speakers. The contingency matrix and distributions across these three classes by both speaker groups appear in Table 1. It is notable that our non-native speakers are more consistent in their use of accent than are the native speakers. It is possible that non-native speakers use a narrow range of criteria in their accenting decisions — e.g. only accent nouns, or only accent the first word in a phrase — while the native speakers exercise more individual variation in their intonation. There is also a numeric contribution to this difference — it takes only one disagreement to move a token from an *always* or *never* category to the *sometimes* category. Thus by virtue of having six native speakers and four non-native speakers we can expect to have increased disagreement among the native speakers. If we allow the prominence decision to be an independent random process with a prior distribution equal to the accent rate of each of the two groups, we would expect 82.4% disagreement (i.e. *sometimes* accented) in the non-native group and 96.8% in the native group.

		MC			
	Group	Always	Sometimes	Never	Total
SAE	Always	153	38	0	191
	Sometimes	311	304	130	745
	Never	4	74	75	153
	Total	468	416	205	1089

Table 1: *Contingency Matrix and Distribution of tokens* always, sometimes *and* never *accented by SAE and MC speakers.*

Even though our Mandarin speakers agree more than the SAE speakers, as the non-native speakers have different levels of experience speaking American English, we do not expect their agreement with the native speakers to be consistent. Omitting the *optionally* prominent tokens, we find that the four non-native speakers show 87.8%, 87.8%, 85.8% and 84.3% rates of agreement with the native speakers. We next ask if the “errors” — those tokens where a non-native speaker accents a word that is never accented by native speakers, or fail to accent a word that is always accented — are consistent across the non-native speakers. There are only four tokens that are produced with prominence inconsistent with the native speakers by all native speakers. All of these are instances in which none of the native speakers accented the token, but all of the non-native speakers did. These examples are: 1) “Computerized phone calls, **which** do everything from selling magazine subscriptions, . . .”, 2) “First time offenders used to lose their license for thirty days, well now **they** could lose it for as many as ninety.”, 3) “. . . but he’s already **set** its goal.” and 4) “. . . and your blood content levels register .10 or higher, **you** can automatically lose your license for ninety days.”

3.2. Syntactic analysis

In native SAE, the part-of-speech (POS) of a word has a significant influence on whether or not the word will bear accent or not. In this section we compare the relationship between word class — Noun, Verb, Adverb, Adjective, Cardinal or Function Word — and accenting behavior of native and non-native speakers. By using only the aligned tokens we are able to analyze this relationship using an ANOVA with repeated measures on the within group accent rate of each set of tokens. That is, for each token with each part-of-speech word class we determine the rate at which native and non-native speakers accent the token. We then use a paired t-test to determine if the accent rate within each word class differs between the native and non-native speaker groups.

The accent rate of each POS word class by native and non-native speakers is shown in Table 2. The ANOVA reveals an effect of both language — native or non-native — and POS with $p < 2{*}10^{-16}$; however, there is no significant combined effect, $p = .2912$. This suggests that, while the non-native speakers accent tokens more frequently, this effect does not impact the

	N	Vb.	Adj.	Adv.	Card.	Fn.
SAE	71.02	54.79	73.22	76.89	75.50	24.90
MC	83.48	69.48	79.10	85.86	82.41	35.53

Table 2: *Accent rate of tokens of each POS-based word-class by Native and Non-Native speakers.*

accent rate of different parts of speech with any significant observable difference.

3.3. Phrase Position

In our analysis of accent rates (cf. Section 3.1), we observe that non-native speakers accent with greater frequency than native SAE speakers. We also note that, in the same amount of lexical material, the non-native speakers use more, and therefore shorter, intermediate and intonational phrases. In this section, we compare the relationship between accenting and phrasing.

Using ANOVA tests, we evaluate the effects of native language and phrase position on accenting. We perform this analysis for both intermediate and intonational phrase position. On both of these analyses we find significant effects of native language and phrase position, as predicted, but we also see a significant combined effect. For intermediate phrases, the phrase position effect is significant with $p < 2.2 * 10^{-16}$; the corpus effect is significant with $p = 1.12 * 10^{-19}$; and the combined effect has $p < 2.2 * 10^{-16}$. Examining the intonational phrase position, the position effect is significant with $p < 2.2 * 10^{-16}$; the corpus effect is significant with $p = 8.68 * 10^{-15}$; and the combined effect has a p-value of $7.77 * 10^{-12}$.

We find that both speaker groups have a tendency to accent phrase final[2] words more frequently than phrase-initial or -medial words. However, native SAE speakers are more likely to accent medial tokens than initial words, while native MC speakers accent initial and medial tokens at approximately the same rate. This effect is consistent whether we examine intermediate or intonational phrase position. In Figure 3 we present the accent rates for each corpus based on intermediate and intonational phrase position, respectively. Here we can see the $\sim 20\%$ increase in accent rate of phrase initial tokens. Also, we observe that the accent rate at phrase final tokens is roughly equivalent on intermediate and intonational phrases in native SAE speech, but MC speakers accent intermediate phrase final words 9% more frequently than intonational phrase final words.

Corpus	ip BEGIN	ip MEDIAL	ip FINAL
SAE	33.7	50.2	77.7
MC	**54.7**	53.2	85.4

Corpus	IP BEGIN	IP MEDIAL	IP FINAL
SAE	31.0	48.3	78.8
MC	**52.8**	51.8	76.8

Table 3: *Accent rates (%) of native SAE and native MC speakers based on intermediate phrase (ip) intonational phrase (IP) position.*

3.4. Acoustic analysis

It is often suggested that differences between native and non-native intonation might be due to artifacts from the native language. In Mandarin Chinese, focus-bearing words are typically produced with an expanded pitch range [6]. While pitch range is used to indicate prominence, durational cues are also used [7]; moreover, pitch cues need not be present for the perception of emphasis [8]. To see whether native and non-native speakers produce accents differently in *English*, we compare their acoustic correlates of prominence. We employ t-tests to compare acoustic features extracted from prominent and non-prominent words. We extract acoustic features based on pitch, intensity and duration of each word. The acoustic features we examine are intended to capture either excursion in one of these three acoustic domains, or to represent a quality of the shape of the pitch or intensity contour. The excursion features are aggregations — minimum, maximum and mean — of speaker-normalized (using z-score normalization) pitch, intensity or word duration. We also perform context normalization of these values using a context window of two previous and two following words. To capture the shape of the pitch and intensity contours, we identify 1) the maximum, mean and standard deviation of slope, 2) the relative location of the maximum value, 3) tilt [9] and skew [10] parameters, 4) the standard deviation of the values of each contour, and 5) the slope of the contour leading into the local maxima, and trailing from the local maxima.

We find that many of the acoustic features used by native SAE speakers to indicate the prominence of a word are also used by native MC speakers. We observe increased duration on prominent words by both groups as well as increased mean and maximum intensity; these effects are observed when the value is evaluated in isolation and when normalized by the surrounding context. We find that the context-normalized pitch aggregations (max and mean) are increased on prominent words for both speaker groups, though these aggregations taken without context normalization only show this effect in non-native speech. That is, native speakers do not demonstrate significantly different maximum and mean speaker-normalized pitch on accented vs. non-accented words; only when the surrounding context is included in the analysis does pitch reveal a significant difference between prominent and non-prominent tokens.

Shape features also show similar effects in both speaker groups. We observe increased maximum, mean and minimum slope of pitch and intensity in prominent words by both native and non-native speakers. Also, we find that tilt and skew parameters are greater in prominent tokens. For tilt parameters, this indicates an earlier peak, and greater rise than fall for both pitch and intensity. The increased skew that we observe indicates a pitch peak that is earlier relative to the intensity peak in prominent productions than in non-prominent ones. This relationship is also observed by measuring the distance between f0 and intensity peaks. The positive correlation between this distance and prominence is observable in both speaker groups (SAE: p=$1.6 * 10^{-83}$ MC: p=$4.73 * 10^{-23}$) with later intensity peaks correlating with prominence.

In addition to these broad consistencies in duration, pitch, and the shape of pitch and intensity contours, however, we find some acoustic differences. In particular, native speakers use intensity differently from non-native speakers. Native SAE speakers produce a significantly later intensity peak during prominent tokens, as calculated by tilt parameters fitted to the energy contour (-0.119 prominent, -0.042 non-prominent; p=$3.94 * 10^{-5}$). Additionally, the intensity slope over prominent words significantly differs in native SAE speech (p=$1.41 * 10^{-7}$): Prominent tokens have a slope of -0.038 dB/sec, while this value is -1.51 dB/sec in non-prominent tokens. This feature does not show any difference in the Mandarin speaker's productions. This use of energy dynamics represents a set of signals that native speakers employ to indicate prominence that non-native speakers do not. There are also features that show a significant difference in the non-native material, that show no difference in native speech. We observe that the maximum (1.09 v. 0.32) and mean (.12 v. -.28) speaker normalized pitch are significantly different only in non-native speech (p=$3.81 * 10^{-65}$, p=$2.21 * 10^{-67}$). In native

[2] We consider tokens in single word phrases to be phrase-final, although they are both phrase-initial and -final.

SAE, the maximum and mean normalized pitch do not show any significant difference (p=0.913, p=0.261) across prominent and non-prominent words — these pitch aggregations are only significant when normalized by their surrounding context.

4. Automatic Detection of Prominence in Non-native speech

In Section 3.4, we observed that many of the acoustic correlates of native SAE productions of prominence are similar to native MC productions. These similarities raise the possibility of using models trained on native speech to predict prosodic events such as accent in non-native speech. There has been a significant amount of work on the automatic detection of accent in native speech (cf. [11, 10, 12, 13]). In this section we evaluate the ability of these approaches to detect prominence in native Mandarin speakers' production of English.

To this end, we use the corrected energy-based classifiers approach described in [14, 10]. We believe this technique has the best pitch accent detection rates on SAE speech at 84.95% ± 0.787 accuracy under speaker independent evaluation on BURNC material. This ensemble technique generates accent detection predictions based on band-pass filtered energy features, then corrects these predictions based on pitch and duration information before combining the ensemble of predictions into a single hypothesis.

Significant lexical correlates of prominence have been noted (cf. Section 3.2, [10, 15], *inter alia*). However, in our material here, lexical content is the same across speakers. In order to perform a speaker-independent evaluation of this task, we thus cannot use lexical information, or risk identical content occurring in both the testing and training data, inflating the results. In order to avoid this pitfall, we opt to evaluate the automatic detection of prominence using only acoustic information.

We first evaluate the use of the corrected energy-based classifier technique on the MC data. We perform this evaluation using leave-one-speaker-out cross validation — we train four models, each time omitting the material spoken by one speaker, these models are used to generate accent predictions for the material from the omitted speaker. This technique is able to detect prominence in MC material with $87.23\% \pm 0.79$ accuracy. This corresponds to an accent detection f-measure of 0.899 ± 0.00672 with a slightly higher recall (0.921) than precision (0.878). By way of comparison, under speaker independent evaluation accents in the BURNC material are detected with 85.4% ±0.78 accuracy (f-measure=0.866) using classifiers trained on BURNC. In general, the rate of human agreement in detecting prominence is somewhere between 81% [16] and 91% [17]. The accuracy of this approach therefore approaches the maximum rate of human agreement on this task.

While these results indicate that the automatic detection of prominence in non-native speech can be accomplished with high performance, they rely on the use of non-native training data. To evaluate the dependence on this consistent training data, we evaluate the performance of models trained on BURNC material in the detection of prominence on the MC material. In this evaluation the accuracy is 74.77% with an f-measure of 0.824. The BURNC model significantly overpredicts accented tokens; while the recall of the accent class is 0.951, the precision is only 0.726. Approximately 81% of tokens are hypothesized to be accented. This suggests that, while the same technique can be applied — using the same features, the model parameters required to generate high performance on SAE and MC material are significantly different.

5. Conclusions and Future Work

In this paper we have described a production experiment designed to identify Mandarin speakers' realization of intonational prominence in English using speech read in the laboratory. Comparing these to native productions from the BURNC corpus, we find that native speakers accent fewer words than non-native speakers, perhaps because the latter hyperarticulate or at least produce shorter phrases. When we compare tokens that are always, sometimes or never accented across the two speaker groups, we find considerable agreement between the two groups, with non-native speakers exhibiting more within-group consistency. We also find that native and non-native speakers differ in their propensity to accent phrase-initial vs. phrase medial words, altho both tend to accent phrase-final words at a similar rate. We also have identified a number of acoustic correlates of accent realization that both groups share — in duration, pitch and pitch and intensity contour shape, with differences however in native speakers' use of energy dynamics to indicate prominent which non-native speakers do not share. We have also explored the automatic identification of prominence in non-native speech utilizing models train on material from non-native vs. native speech, with encouraging results.

6. References

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

[2] 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.

[3] 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.

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

[5] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Boston University, Tech. Rep. ECS-95-001, March 1995.

[6] F. Liu, "Parallel encoding of focus and interrogative meaning in mandarin intonation," *Phonetica*, vol. 62, pp. 70–87, 2005.

[7] C. Shih, *Intonation: analysis, modelling and technology*. Springer, 2000, ch. A Declination Model of Mandarin Chinese.

[8] X. S. Shen, "Relative duration as a perceptual cue to stress in mandarin," *Language and Speech*, vol. 36, pp. 415–433, 1993.

[9] P. Taylor, "The tilt intonation model," in *ICSLP*, 1998.

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

[11] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in *HLT-NAACL*, 2009.

[12] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling," in *ICSLP*, 2006.

[13] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processig*, vol. 2, no. 4, October 1994.

[14] A. Rosenberg and J. Hirschberg, "On the correlation between energy and pitch accent in read english speech," in *Interspeech*, 2006.

[15] V. R. Sridhar, A. Nenkova, S. Narayanan, and D. Jurafsky, "Detecting prominence in conversational speech: Pitch accent, givenness and focus," in *Speech Prosody*, 2008.

[16] J. Pitrelli, M. Beckman, and J. Hirschberg, “Evaluation of prosodic transcription labeling reliability in the tobi framework.” in *ICSLP*, 1994.

[17] A. Syrdal and J. McGory, “Inter-transcriber reliability of tobi prosodic labeling,” in *ICSLP*, 2000.