Investigating Pitch Accent Recognition in Non-native Speech

Gina-Anne Levow
Computer Science Department
University of Chicago
ginalevow@gmail.com

Abstract

Acquisition of prosody, in addition to vocabulary and grammar, is essential for language learners. However, it has received less attention in instruction. To enable automatic identification and feedback on learners' prosodic errors, we investigate automatic pitch accent labeling for non-native speech. We demonstrate that an acoustic-based context model can achieve accuracies over 79% on binary pitch accent recognition when trained on within-group data. Furthermore, we demonstrate that good accuracies are achieved in cross-group training, where native and near-native training data result in no significant loss of accuracy on non-native test speech. These findings illustrate the potential for automatic feedback in computer-assisted prosody learning.

1 Introduction

Acquisition of prosody, in addition to vocabulary and grammar, is essential for language learners. However, intonation has been less-emphasized both in classroom and computer-assisted language instruction (Chun, 1998). Outside of tone languages, it can be difficult to characterize the factors that lead to non-native prosody in learner speech, and it is difficult for instructors to find time for the one-on-one interaction that is required to provide feedback and instruction in prosody.

To address these problems and enable automatic feedback to learners in a computer-assisted language learning setting, we investigate automatic prosodic labelling of non-native speech. While many prior systems (Teixeia et al., 2000; Tepperman and Narayanan, 2008) aim to assign a score to the learner speech, we hope to provide more focused feedback by automatically identifying prosodic units, such as pitch accents in English or tone in Mandarin, to enable direct comparison with gold-standard native utterances.

There has been substantial progress in automatic pitch accent recognition for native speech, achieving accuracies above 80% for acoustic-feature based recognition in multi-speaker corpora (Sridhar et al., 2007; Levow, 2008). However, there has been little study of pitch accent recognition in non-native speech. Given the challenges posed for automatic speech recognition of non-native speech, we ask whether recognition of intonational categories is practical for non-native speech. To lay the foundations for computer-assisted intonation tutoring, we ask whether competitive accuracies can be achieved on non-native speech. We further investigate whether good recognition accuracy can be achieved using relatively available labeled native or near-native speech, or whether it will be necessary to collect larger amounts of training or adaptation data matched for speaker, language background, or language proficiency.

We employ a pitch accent recognition approach that exploits local and coarticulatory context to achieve competitive pitch accent recognition accuracy on native speech. Using a corpus of prosodically labelled native and non-native speech, we illustrate that similar acoustic contrasts hold for pitch accents in both native and non-native speech. These contrasts yield competitive accuracies on binary pitch accent recognition using within-group training data. Furthermore, there is no significant drop in accuracy when models trained on native or near-native speech are employed for classification of non-native speech.

The remainder of the paper is organized as follows. We present the LeaP Corpus used for our experiments in Section 2. We next describe the feature sets employed for classification (Section 3) and contrastive acoustic analysis for these features in native and non-native speech (Section 4). We

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 269–272,
Suntec, Singapore, 4 August 2009.

ID	Description
c1	non-native, before prosody training
c2	non-native, after first prosody training
c3	non-native, after second prosody training
e1	non-native, before going abroad
e2	non-native, after going abroad
sl	'super-learner', near-native
na	native

Table 1: Speaker groups, with ID and description in the LeaP Corpus

then describe the classifier setting and experimental results in Section 5 as well as discussion. Finally, we present some conclusions and plans for future work.

2 LeaP Corpus and the Dataset

We employ data from the LeaP Corpus (Milde and Gut, 2002), collected at the University of Bielefeld as part of the "Learning Prosody in a Foreign Language" project. Details of the corpus (Milde and Gut, 2002), inter-rater reliability measures (Gut and Bayerl, 2004), and other research findings (Gut, 2009) have been reported.

Here we focus on the read English segment of the corpus that has been labelled with modified EToBI tags[1] , to enable better comparison with prior results of prosodic labelling accuracy and also to better model a typical language laboratory setting where students read or repeat. This yields a total of 37 recordings of just over 300 syllables each, from 26 speakers, as in Table 1.[2] This set allows the evaluation of prosodic labelling across a range of native and non-native proficiency levels. The modified version of EToBI employed by the LeaP annotators allows transcription of 14 categories of pitch accent and 14 categories of boundary tone. However, in our experiments, we will focus only on pitch accent recognition and will collapse the inventory to the relatively standard, and more reliably annotated, four-way (high, downstepped high, low, and unaccented) and binary (accented, unaccented) label sets.

[1]While the full corpus includes speakers from a range of languages, the EToBI labels were applied primarily to data from German speakers.

[2]Length of recordings varies due to differences in syllabification and cliticization, as well as disfluencies and reading errors.

3 Acoustic-Prosodic Features

Recent research has highlighted the importance of context for both tone and intonation. The role of context can be seen in the characterization of pitch accents such as down-stepped high and in phenomena such as downdrift across a phrase. Further, local coarticulation with neighboring tones has been shown to have a significant impact on the realization of prosodic elements, due to articulatory constraints (Xu and Sun, 2002). The use of prosodic and coarticulatory context has improved the effectiveness of tone and pitch accent recognition in a range of languages (Mandarin (Wang and Seneff, 2000), English (Sun, 2002)) and learning frameworks (decision trees (Sun, 2002), HMMs (Wang and Seneff, 2000), and CRFs (Levow, 2008)).

Thus, in this work, we employ a rich contextual feature set, based on that in (Levow, 2008). We build on the pitch target approximation model, taking the syllable as the domain of tone prediction with a pitch height and contour target approached exponentially over the course of the syllable, consistent with (Sun, 2002). We employ an acoustic model at the syllable level, employing pitch, intensity and duration measures. The acoustic measures are computed using Praat's (Boersma, 2001) "To pitch" and "To intensity." We log-scaled and speaker-normalized all pitch and intensity values.

We compute two sets of features: one set describing features local to the syllable and one set capturing contextual information.

3.1 Local features

We extract features to represent the pitch height and pitch contour of the syllable. For pitch features, we extract the following information: (a) pitch values for five evenly spaced points in the voiced region of the syllable, (b) pitch maximum, mean, minimum, and range, and (c) pitch slope, from midpoint to end of syllable. We also obtain the following non-pitch features: (a) intensity maximum and mean and (b) syllable duration.

3.2 Context Modeling

To capture local contextual influences and cues, we employ two sets of features. The first set of features includes differences between pitch maxima, pitch means, pitches at the midpoint of the syllables, pitch slopes, intensity maxima, and intensity means, between the current and preceding or fol-

lowing syllable. The second set of features adds the last pitch values from the end of the preceding syllable and the first from the beginning of the following syllable. These features capture both the relative differences in pitch associated with pitch accent as well as phenomena such as pitch peak delay in which the actual pitch target may not be reached until the following syllable.

4 Acoustic Analysis of Native and Non-native Tone

To assess the potential effectiveness of tone recognition for non-native speech, we analyze and compare native and non-native speech with respect to features used for classification that have shown utility in prior work. Pitch accents are characterized not only by their absolute pitch height, but also by contrast with neighboring syllables. Thus, we compare the values for pitch and delta pitch, the difference between the current and preceding syllable, both with log-scaled measures for high-accented and unaccented syllables. We contrast these values within speaker group (native: na; non-native: e1, c1). We also compare the delta pitch measures between speaker groups (na versus e1 or c1).

Not only do we find significant differences for delta pitch between accented and unaccented syllables for native speakers as we expect, but we find that non-native speakers also exhibit significant differences for this measure (t-test, two-tailed,$p < 0.001$). Accented syllables are reliably higher in pitch than immediately preceding syllables, while unaccented syllables show no contrast. Importantly, we further observe a significant difference in delta pitch for high accented syllables between native and non-native speech. Native speakers employ a markedly larger change in pitch to indicate accent than do non-native speakers, a fine-grained view consistent with findings that non-native speakers employ a relatively compressed pitch range (Gut, 2009).

For one non-native group (e1), we find that although these speakers produce reliable contrasts in delta pitch between *neighboring* syllables, the overall pitch height of high accented syllables is not significantly different from that of unaccented syllables. For native speakers and the 'c1' non-native group, though, overall pitch height does differ significantly between accented and unaccented syllables. This finding suggests that while all speakers in this data set understand the locally contrastive role of pitch accent, some non-native speaker groups do not have as reliable global control of pitch.

The presence of these reliable contrasts between accented and unaccented syllables in both native and non-native speech suggests that automatic pitch accent recognition in learner speech could be successful.

5 Pitch Accent Recognition Experiments

We assess the effectiveness of pitch accent recognition on the LeaP Corpus speech. We hope to understand whether pitch accent can be accurately recognized in non-native speech and whether accuracy rates would be competitive with those on native speech. In addition, we aim to compare the impact of different sources of training data. We assess whether non-native prosody can be recognized using native or near-native training speech or whether it will be necessary to use matched training data from non-natives of similar skill level or language background.

Thus we perform experiments on matched training and test data, training and testing within groups of speakers. We also evaluate cross-group training and testing, training on one group of speakers (native and near-native) and testing on another (non-native). We contrast all these results with assignment of the dominant 'unaccented' label to all instances (common class).

5.1 Support Vector Machine Classifier

For all supervised experiments reported in this paper, we employ a Support Vector machine (SVM) with a linear kernel. Support Vector Machines provide a fast, easily trainable classification framework that has proven effective in a wide range of application tasks. For example, in the binary classification case, given a set of training examples presented as feature vectors of length D, the linear SVM algorithm learns a vector of weights of length D which is a linear combination of a subset of the input vectors and performs classification based on the function $f(x) = sign(w^T x - b)$. We employ the publicly available implementation of SVMs, LIBSVM (C-C.Cheng and Lin, 2001).

5.2 Results

We see that, for within group training, on the binary pitch accent recognition task, accuracies

	c1	c2	c3	e1	e2	*sl*	**na**
Within-group Accuracy	79.1	80.9	80.6	81	82.5	*82.4*	**81.2**
Cross-group Accuracy (**na**)	77.2	79	81.4	80.3	82.5	*83.2*	
Cross-group Accuracy (*sl*)	77.3	79.9	82	80.5	82.9		**81.6**
Common Class	56.9	59.6	56.2	70.2	64	*65.5*	**63.6**

Table 2: Pitch accent recognition, within-group, cross-group with native and near-native training, and most common class baseline: Non-native (plain), 'Super-learner' (underline *sl*), Native (bold **na**)

range from approximately 79% to 82.5%. These levels are consistent with syllable-, acoustic-feature-based prosodic recognition reported in the literature (Levow, 2008). A summary of these results appears in Table 2. In the cross-group training and testing condition, we observe some variations in accuracy, for some training sets. However, crucially none of the differences between native-based or near-native training and within-group training reach significance for the binary pitch accent recognition task.

6 Conclusion

We have demonstrated the effectiveness of pitch accent recognition on both native and non-native data from the LeaP corpus, based on significant differences between accented and unaccented syllables in both native and non-native speech. Although these differences are significantly larger in native speech, recognition remains robust to training with native speech and testing on non-native speech, without significant drops in accuracy. This result argues that binary pitch accent recognition using native training data may be sufficiently accurate that to avoid collection and labeling of large amounts of training data matched by speaker or fluency-level to support prosodic annotation and feedback. In future work, we plan to incorporate prosodic recognition and synthesized feedback to support computer-assisted prosody learning.

Acknowledgments

We thank the creators of the LeaP Corpus as well as C-C. Cheng and C-J. Lin for LibSVM. This work was supported by NSF IIS: 0414919.

References

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

C-C.Cheng and C-J. Lin. 2001. LIBSVM:a library for support vector machines. Software available at: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Dorothy M. Chun. 1998. Signal analysis software for teaching discourse intonation. *Language Learning & Technology*, 2(1):61–77.

U. Gut and P. S. Bayerl. 2004. Measuring the reliability of manual annotations of speech corpora. In *Proceedings of Speech Prosody 2004*.

U. Gut. 2009. *Non-native speech. A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Peter Lang, Frankfurt.

G.-A. Levow. 2008. Automatic prosodic labeling with conditional random fields and rich acoustic features. In *Proceedings of the IJCNLP 2008*.

J.-T. Milde and U. Gut. 2002. A prosodic corpus of non-native speech. In *Proceedings of the Speech Prosody 2002 Conference*, pages 503–506.

V. Rangarajan Sridhar, S. Bangalore, and S. Narayanan. 2007. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. In *Proceedings of HLT NAACL 2007*, pages 1–8.

Xuejing Sun. 2002. Pitch accent prediction using ensemble machine learning. In *Proceedings of ICSLP-2002*.

C. Teixeia, H. Franco, E. Shriberg, K. Precoda, and K. Somnez. 2000. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In *Proceedings of ICSLP 2000*.

J. Tepperman and S. Narayanan. 2008. Better non-native intonation scores through prosodic theory. In *Proceedings of Interspeech 2008*.

C. Wang and S. Seneff. 2000. Improved tone recognition by normalizing for coarticulation and intonation effects. In *Proceedings of 6th International Conference on Spoken Language Processing*.

Yi Xu and X. Sun. 2002. Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America*, 111.