Context in Multi-lingual Tone and Pitch Accent Recognition

Gina-Anne Levow

Department of Computer Science
University of Chicago, Chicago, IL, USA
levow@cs.uchicago.edu

10.21437/Interspeech.2005-552

Abstract

Tone and intonation play a crucial role across many languages. However, the use and structure of tone varies widely, ranging from lexical tone which determines word identity to pitch accent signalling information status. In this paper, we employ a uniform representation of acoustic features for recognition of both Mandarin tone and English pitch accent. The representation captures both local tone height and shape as well as contextual coarticulatory and phrasal influences. By exploiting multiclass Support Vector Machines as a discriminative classifier, we achieve competitive rates of tone and pitch accent recognition. We further demonstrate the greater importance of modeling preceding local context, which yields up to 24% reduction in error over modeling the following context.

1. Introduction

Tone and intonation play a crucial role across many languages. However, the use and structure of tone varies widely, ranging from lexical tone which determines word identity to pitch accent signalling information status.

Recent research has demonstrated the importance of contextual and coarticulatory influences on the surface realization of tones.[1, 2] The overall shape of the tone or accent can be substantially modified by the local effects of adjacent tone elements. Furthermore, broad scale phenomena such as topic [3] and phrase structure can affect pitch height, and pitch shape may be variably affected by the presence of boundary tones.

In addition to earlier approaches that employed phrase structure [4], several recent approaches to tone recognition in East Asian languages [5, 6, 7] and to tone generation [8] have incorporated elements of local and broad range contextual influence on tone. Many of these techniques create explicit context-dependent models of the phone, tone, or accent for each context in which they appear, either using the tone sequence for left or right context or using a simplified high-low contrast, as is natural for integration in a Hidden Markov Model speech recognition framework. With StemML[8], templates corresponding to canonical tone models are presumed to be deformed to conform to the current context. Studies of pitch accent have often included features providing contrasts with neighboring words or syllables, though less explicitly in a coarticulatory framework [9]. [10]'s work captures elements of local influence on accent identity, but employs no broader range features.

In this work, we bring together both local and broader contextual influences. Local effects are captured by creating an "extended syllable" representation that incorporates additional features from the local syllabic environment as detailed below. We also apply a phrase-based transformation of the pitch features to compensate for falling slope across a phrase. Using this representation to train Support Vector Machine classifiers, we perform pitch accent recognition in English and tone recognition in Mandarin, achieving competitive classification results, 81.3% and 76.5% respectively. We further demonstrate the utility of local and long range context features, showing that preceding context is crucially more important than following context, though for pitch accent detection the latter does contribute to the best results.

2. Experimental Framework

2.1. Tone and Intonation Modeling

Our model is inspired by the pitch target approximation model of [1]. This approach is grounded in articulatory constraints such as maximum speed of pitch change that predict tonal coarticulation. Each tonal element is viewed as having an underlying target characterized by pitch slope and height. Under coarticulatory constraints, the target may not be achieved immediately, but is gradually approached, with the difference decaying exponentially.

We take the syllable as the domain of tone and pitch accent prediction, consistent with [9]. We employ a purely acoustic model at the syllable level, employing pitch, intensity and duration measures. The acoustic measures are computed using Praat's [11] "To pitch" and "To intensity" functions. Since our datasets include a variety of speakers, we compute a per-speaker log-scale and normalized form for all pitch and intensity values.

For pitch features, we extract pitch values for five evenly spaced points in the voiced region of the syllable[1]. We also calculate pitch maximum and mean. Following [12], we assume that the pitch target can be expected to be closely approached by the middle of the syllable. Thus, we compute a linear fit to pitch slope from the midpoint to the end of the syllable. Finally we obtain maximum and mean intensity and syllable duration.

To capture local contextual influences and cues, we incorporate two sets of features. The first set of features ("difference features") correspond to differences between the current syllable and its preceding and following syllables. They include difference between pitch maxima, pitch means, pitch at the midpoint of the syllable, pitch

[1] We restrict our experiments to those syllables with at least 50 milliseconds of voicing.

slopes, intensity maxima, and intensity means. The second set of features, which we will refer to as "extended syllable" features, are simply the last pitch values from the end of the preceding syllable and the first from the beginning of the following syllable, as well as the pitch maxima and means of these adjacent syllables.

2.2. Data Sets

We consider two corpora: one in English for pitch accent recognition and one in Mandarin for tone recognition. We introduce each briefly below; in each case, approximately one fourth of the samples were held out for testing.

2.2.1. English Corpus

We employ a subset of the Boston Radio News Corpus [13], read by female speaker F2B, comprising 40 minutes of news material. The corpus includes pitch accent, phrase and boundary tone annotation in the ToBI framework [14] aligned with manual transcription and syllabification of the materials. Following earlier research [9, 10], we collapse the ToBI pitch accent labels to four classes: unaccented, high, low, and downstepped high for experimentation.

2.2.2. Mandarin Corpus

We extracted a subset of the Voice of America Mandarin broadcast news corpus distributed as part of the Topic Detection and Tracking [15] task. The audio material includes news stories read by several anchors, and recorded during the first half of 1998. Text versions of anchors' scripts were also provided with the corpus, but were aligned only at the story level. To obtain tones and syllable boundaries, we performed a forced alignment based on the scripts using the language porting framework provided by the University of Colorado's Sonic speech recognizer [16]. We created a pronunciation lexicon, based on the pinyin entries in a Chinese-English lexicon, with a noisy mapping from Mandarin to the base English phoneme set. Finally, we hand-verified the alignment to correct errors in the original transcripts or severe misalignments. We perform five-way classification of the four canonical tones and the neutral tone; labels assume that tone sandhi transformation has been applied.

2.3. Classifier

For all 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 $f(x) = sign(w^T x - b)$. Furthermore, SVMs have been generalized from binary classification to multiclass classification as well as semi-supervised frameworks. The corresponding weights can also provide insight into the contribution of different features to the classification process. We employ two publicly available implementations of SVMs, SVMlight [17] and LIBSVM [18].

2.4. Contrasts

We consider two sources of contextual influence: local coarticulatory constraints and broader range phrasal effects. Both coarticulation and declination have been shown to operate across a wide range of languages [19], and physical constraints, such as lung volume and articulator speed, have been suggested as a common cause. [20, 21]

2.4.1. Local Context

To assess local influences on tone recognition and tone modeling, we perform ablation experiments contrasting the impact of the two feature sets directed to local context, difference features and "expanded syllables" features as described above. We further contrast the effects of modeling preceding versus following context.

2.4.2. Phrase Effects

To evaluate the impact of long range phrasal and discourse features, we compute a modified representation of all scalar pitch features. Since we are using only acoustic features, we estimate phrases as story or silence delimited segments.[2] We computed a linear fit to the overall phrase contour excluding the final syllables, and, we hope, sharp boundary tone effects. For both Mandarin and English, we find, as expected, that phrases have an overall falling slope. We then used the median slope, calculated per syllable, across the corpus as phrase-based falling contour compensation. As [5] did, we find the reestimation based on individual phrase slope overfit to the specific pitch accent or tone configuration, and reduced accuracy. In the phrase based feature representation, each pitch value is thus replaced with an estimate of the pitch value without phrasal effects, by adding back the estimated pitch drop to pitch values later in the phrase.

3. Results

We trained multiclass Support Vector Machines with linear kernels on syllable-based feature vector representations, testing on a held out subset. One classifier performed the 4-way classification on English pitch accent, the other the 5-way classification of Mandarin tone. Using the full feature vector described above, we obtain classification accuracies competitive with those cited in the literature: 81.3% for pitch accent [9, 22, 10] and 76% for Mandarin newswire speech [5, 6]. Baselines for most frequent class assignment are 63.4% for pitch accent and 34% for tone recognition with the current training/test splits. A confusion matrix for Mandarin tone appears in Table 1.

3.1. Contributions of Local Context

In order to understand the relative contributions of the two types of local context features as well as relative contributions of preceding and following context, we conducted a set of contrastive experiments training and testing on subsets of the features. We grouped contextual

[2] Clearly a more sophisticated prosodic approach to phrase boundary detection could be applied, but it is beyond the scope of the current paper.

Reco	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5
Tone					
Tone 1	84.3%	6.7%	0%	9%	0%
Tone 2	8.9%	78.6%	3.6%	5.3%	0%
Tone 3	5%	10%	70%	5%	10%
Tone 4	13.2%	7.4%	7.4%	70.6%	1.5%
Tone 5	0%	27.3%	27.3%	0%	45.4%

Table 1: Confusion Matrix for Mandarin Tone Recognition

features by type – difference ("Diff") or extended syllable ("Extend") – and by position, relative to the preceding, left context syllable ("L") or following, right context syllable ("R"). We compared different combinations of the available contextual features as follows:

1. All contextual features ("Full context")
2. Extended syllable features for left context, right context, and both ("Extend L", "Extend R", "Extend LR")
3. Difference features for left context, right context, and both ("Diff L", "Diff R", "Diff LR")
4. All preceding context features ("Both L")
5. All following context features ("Both R")
6. No context features ("No context")

The results of these comparisons appear in Table 2.

The results indicate clearly that representation of the local context contributes to accuracy of classification in both languages and tone types. One observes that classification with any context features outperforms the "no context" condition shown in the last row. Furthermore, preceding context is substantially more important than following context. All conditions in which preceding context information is added outperforms comparable conditions without that left context information. For example, the "Diff LR" and "Extend LR" conditions outperform the "Diff R" and "Extend R" conditions, respectively, for both languages. In addition, classification with left context features alone outperforms classification with right context features alone, as shown in the differences in effectiveness for "Diff L" versus "Diff R", "Extend L" versus "Extend R", and "Both L" versus "Both R." In fact, at least for this collection, including following context degrades performance for Mandarin, while it makes a small positive contribution in the pitch accent case. This contrast in left and right context is consistent with phonetic findings [20] that coarticulatory effects are primarily carryover, rather than anticipatory. It is also advantageous both in terms of language perception and in terms of on-line computational recognition. The majority of the information required for the listener, human or machine, to appropriately interpret the pitch contour is available at the time the tone is produced.

The relative contribution of the difference features and the extended syllable features is less clear. Both contribute and performance is at least as good with both as with either alone.

Context	Mandarin Tone	English Pitch Accent
Full context	74.5%	81.3%
Extend LR	74%	80.7%
Extend Left	74%	79.9%
Extend Right	70.5%	76.7%
Diffs LR	75.5%	80.7%
Diffs Left	76.5%	79.5%
Diffs Right	69%	77.3%
Both L	76.5%	79.7%
Both R	71.5%	77.6%
No context	68.5%	75.9%

Table 2: Effect on classification accuracy of different subsets of local context features. Contrasts are between "difference features" and "extended syllable" features and between preceding (left) and following (right) contexts.

	Mandarin Tone	English Pitch Accent
Phrase	75.5%	81.3%
No Phrase	72%	79.9

Table 3: Effect on classification accuracy of pitch features with (Phrase) and without (No Phrase) compensation for common falling phrasal contour.

3.2. Contributions of Phrase Compensation

The overall results reported above incorporate phrase-based feature transformations. Table 3 presents the results for comparable feature sets with and without phrase-based modification. We can observe that in both cases, compensation for phrasal contour improves classification. Although the effect is not large, it is encouraging in that the phrase segmentation employed here was very simple, and more nuanced approach with finer grained phrase boundary and possibly phrase accent detection would likely yield greater benefit.

4. Discussion and Conclusion

We have employed a uniform acoustic feature representation for syllables for both pitch accent detection in English and tone recognition in Mandarin. Multiclass classification results using an SVM framework with a linear kernel of 81.3% and 76.5% represent greater than 50% reduction in error from a majority class classification baseline. The representation models both local context to capture coarticulatory effects and phrase slope to capture longer range prosodic phenomena. We find that successful recognition of both tone and pitch accent rely crucially on modeling of the preceding context, while modeling following context yields at best small gains and a worst an introduction of noise which degrades classification accuracy. The observations are consistent with current linguistic theory which claims strong persistence or carryover effects in tonal coarticulation and only very weak anticipatory ones and predicts effects due the influence of the broader phrasal context due to declination, downstep, and final lowering.

In future work, we plan to extend the contextual

model to capture a wider range of influences including topic and turn initiation or finality, focus, and other lexical, syntactic and semantic constraints. We are currently preparing experiments on conversational speech in both Mandarin and English, as well as application of the current tone recognition framework to both Cantonese and Bantu tone languages, to more fully assess the range of contextual effects and of different tone and intonational structures.

5. Acknowledgements

We would like to thank Yi Xu for useful discussion. This work was supported by NSF Grant 0414919.

6. References

[1] Y. Xu, "Contextual tonal variations in Mandarin," Journal of Phonetics, vol. 25, pp. 62–83, 1997.

[2] X.-N. Shen, "Tonal co-articulation in Mandarin," Journal of Phonetics, vol. 18, pp. 281–295, 1990.

[3] G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," Computational Linguistics, vol. 27, no. 1, pp. 31–57, 2001.

[4] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in The Production of Speech. Springer-Verlag, 1983, pp. 39–55.

[5] C. Wang and S. Seneff, "Improved tone recognition by normalizing for coarticulation and intonation effects," in Proceedings of 6th International Conference on Spoken Language Processing, 2000.

[6] J. L. Zhou, Y. Tian, Y. Shi, C. Huang, and E. Chang, "Tone articulation modeling for Mandarin spontaneous speech recognition," in Proceedings of ICASSP 2004, 2004.

[7] N. Thubthong and B. Kijsirikul, "Support vector machines for Thai phoneme recognition," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 9, no. 6, pp. 803–813, 2001.

[8] C. Shih and G. P. Kochanski, "Chinese tone modeling with stem-ml," in Proceedings of the International Conference on Spoken Language Processing, Volume 2, 2000, pp. 67–70.

[9] M. Ostendorf and K. Ross, "A multi-level model for recognition of intonation labels," in Computing Prosody, Y. Sagisaka, N. Campbell, and N. Higuchi, Eds., 1997, pp. 291–308.

[10] X. Sun, "Pitch accent prediction using ensemble machine learning," in Proceedings of ICSLP-2002, 2002.

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

[12] X. Sun, "The determination, analysis, and synthesis of fundamental frequency," Ph.D. dissertation, Northwestern University, 2002.

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

[14] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labelling English prosody," in Proceedings of ICSLP, 1992, pp. 867–870.

[15] C. Wayne, "Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation," in Language Resources and Evaluation Conference (LREC) 2000, 2000, pp. 1487–1494.

[16] B. Pellom, W. Ward, J. Hansen, K. Hacioglu, J. Zhang, X. Yu, and S. Pradhan, "University of Colorado dialog systems for travel and navigation," 2001.

[17] T. Joachims, Making large-Scale SVM Learning Practical. MIT Press, 1999.

[18] C-C.Cheng and C.-J. Lin, "LIBSVM:a library for support vector machines," 2001, software available at: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[19] C. Gussenhoven, The Phonology of Tone and Intonation. Cambridge University Press, 2004.

[20] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," Journal of the Acoustical Society of America, vol. 111, 2002.

[21] Y. Xu, "Fundamental frequency peak delay in Mandarin," Phonetica, vol. 58, pp. 26–52, 2001.

[22] M. Hasegawa-Johnson, J. Cole, C. S. abd Ken Chen, A. Cohen, S. Chavarria, H. Kim, T. Yoon, S. Borys, and J.-Y. Choi, "Speech recognition models of the interdependence among syntax, prosody, and segmental acoustics," in HLT/NAACL-2004, 2004.