INTERSPEECH 2011

Modeling Broad Context for Tone Recognition with Conditional Random Fields

Siwei Wang

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

Gina-Anne Levow

Department of Linguistics
University of Washington, Seattle, WA USA
levow@uw.edu

Abstract

We propose a tone recognition approach that employs linear-chain Conditional Random Fields (CRF) to model tone variation due to intonation effects. We implement three linear-chain CRFs which aim at modeling intonation effects at phrase-sentence- and story-level boundaries, where we show that standard recognition techniques degrade and common normalization approaches do not improve. We show that all linear-chain CRFs outperform the baseline unigram model, and the biggest improvement is found in recognizing 3rd tones, (4%) in overall accuracy. In particular, Phrase Bigram CRFs show a drastic 39% improvement in recognizing 3rd tones located at initial boundaries. This improvement shows that the position specific modeling of initial tones in bigram CRFs captures the intonation effects better than the baseline unigram model.

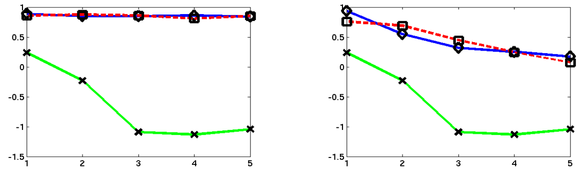
Index Terms: prosody, tone recognition, broad context, conditional random fields.

1. Introduction

Tone languages employ pitch patterns to distinguish syllables which are otherwise ambiguous. In Mandarin Chinese, there are four canonical tones and one neutral tone: 1st: high, 2nd: rising, 3rd: low, and 4th: falling. However, several contextual factors in continuous speech make it challenging to achieve successful tone recognition. First, speaker differences, especially gender differences, make it necessary to compensate for individual variation. Second, coarticulation between adjacent tones can compromise the realization of underlying tone targets. Finally, broad context intonational conditions like phrase, sentence and topic boundaries can also affect pitch; pitch variation has been successfully employed to perform sentence and story segmentation.

Established normalization techniques can compensate for much of the effect of speaker differences. Many machine learning applications modeling local contextual information and coarticulation have been shown to improve tone recognition in continuous speech [1, 2, 3, 4, 5, 6]. However, although some approaches [3, 6, 7] have been proposed to compensate for broader intonational effects, such as declination, these effects still pose significant challenges for tone recognition.

Pitch variation due to intonational effects can have a dramatic effect on tone realization at prosodic unit boundaries and can result in confusion by tone recognition algorithms. Figure 1 depicts two example 3rd tones located at sentence initial boundaries. It is clear that not only are these exemplar tones highly confusable with 1st and 4th tones respectively, but their pitch contours are drastically different from that of the sentence medial 3rd tone also shown in the figure.



Tone (3,1) confusion

Tone (3,4) confusion

Figure 1: Pitch contour confusion of tones at initial positions. Dash lines: phrase medial 1st tone (left) and 4th tone(right). Solid lines with diamonds: Sentence initial 3rd tones (left) and phrase initial 3rd tones(right), Solid lines with cross: 3rd tones at medial positions (no intonation effects)

In this paper, we focus on the challenges of modeling broad context effects on tone recognition. We demonstrate the impact of prosodic boundary effects on tone recognition. We assess the effectiveness of several proposed contextual normalization approaches. We then investigate an alternative approach: encoding the intonational structure in a sequential learning framework, namely, linear chain Conditional Random Fields (CRFs). We also aim to answer the following questions in our sequential tone recognition experiments:

- Individual tone sensitivity to intonation boundaries: Do all four tones react to intonational boundaries in the same way? If not, which tone changes the most?
- Encoding from different sequential graphical models: How do different linear chain CRFs model tone variation at intonation boundaries?

The paper is organized as follows. In section 2, we introduce our dataset, feature representation, and approach for minimizing the effect of local coarticulation. In section 3.2, we briefly overview the sequential graphical models we employed in our experiments. We present our results in 4. In section 5, we will conduct a detailed discussion on how different sequential graphical models, especially the position-specific training incorporated in sequential learning, improves the overall tone recognition accuracy, and especially that for 3rd tones.

2. Data Preparation

We evaluated our models using a subset of the Voice of America Mandarin Chinese broadcast news corpus distributed as part of the Topic Detection and Tracking Task (TDT2) by the Linguistic Data Consortium. In this corpus, the audio was force

28 – 31 August 2011, Florence, Italy

aligned to the corresponding automatically word-segmented anchor scripts using the University of Colorados Sonic Speech Recognizer [8]. This alignment employed a large pinyin pronunciation lexicon and manually constructed mapping from pinyin to APRABET with sandhi rules applied. Severe errors due to mistranscription were manually corrected as well as tone errors due to speaker variations.

For the experiments in this paper, we extracted a dataset consisting of 600 single speaker news stories from the TDT2 corpus. The duration of each news story is around 1 minute. For each story, we identify syllable, word, phrase, and sentence boundaries. Syllable and word boundaries are produced by the alignment above. Phrases are identified as regions delimited by at least 100 ms of silence. Sentence boundaries were marked in the original text transcripts. There are 83,199 syllables and 7,428 phrases in total.

Segments	Phrase	Sentence	Story
Count	7423	2306	600

Table 1: Number of sequences at different levels

Segments	Phrase	Sentence	Story
min	1	1	23
median	9	31	135
max	50	218	259

Table 2: Syllable count of each sequence level

The tone distribution of this dataset is shown in Table 3, we notice that the 3rd tone is the least frequent of the four canonical tones. The 4th tone is the most frequent. We report the recognition performance on these four canonical tones.

We perform landmark-based tone nucleus modeling [9] for every tonal syllable. Landmark-based tone nucleus modeling aims at extracting the region of each tone that is least affected by local coarticulation [9]. Assuming that the best-articulated segmental region should correspond to the best articulated tonal region, we built this tone nucleus modeling technique based on the vowel landmark detection introduced in [10]. This landmark-based tone nucleus modeling has shown improvement in tone recognition [9] and outperformed both pitch contour based nucleus modeling [1] and the supratone Hidden Markov Modeling [4] [11].

For each of these tone nucleus regions, we extract pitch and intensity features using Praat; values are log-scaled, z-score normalized. In additional to local syllable-based features, we include local-context features [5] that are computed as the difference in feature values between the current tone and its previous/following tones. Table 4 lists the complete feature representation.

	1st tone	2nd tone	3rd tone	4th tone	neutral
%	23.07	25.02	13.20	32.79	5.91

Table 3: Tone distribution

3. Modeling Broader Context

There are two major approaches to model the broad context: either we can try to adapt feature values to compensate for intonational effects or we can encode the tone variation into sequential learning frameworks.

3.1. Normalization for Intonational Effects

A variety of approaches have been proposed to compensate for wider window context effects on tone recognition. To assess their utility and establish a baseline, we have implemented the following three techniques for normalization of intonational effects:

- Mean Slope (MS) [3]: Compute a collection average phrase slope and adjust the observed pitch to compensate.
- Moving window by syllable (MW(S))[7]: For every tone, normalize the measured pitch values by the average pitch in a window from two previous to four following syllables.
- Moving window by time (MW(T))[6]: Similar to the previous approach, normalize the measured pitch values by pitch in a window from 0.5s preceding to 1s following the current syllable.

3.2. Using Linear Chain CRF to Model Corresponding Boundary Conditions

CRFs manipulate a class of undirected, conditionally trained graphical models to learn dependencies in both input and output space. First order linear chain CRFs have been employed to perform POS tagging, sentence boundary detection and pitch accent prediction. In this paper, we consider four different CRFs, three of which are linear chain CRFs.

- Unigram only: No bigram connections, equivalent to a maximum entropy classifier
- Phrase bigram+unigram CRF: Bigrams connect all syllables in a phrase, within words or across word boundaries.
- Sentence bigram+unigram CRF: Bigrams connect all syllables inside of a sentence.
- Story bigram+unigram CRF:The sequence contains all syllables in each news story.

GRMM [12] was used for all CRF experiments.

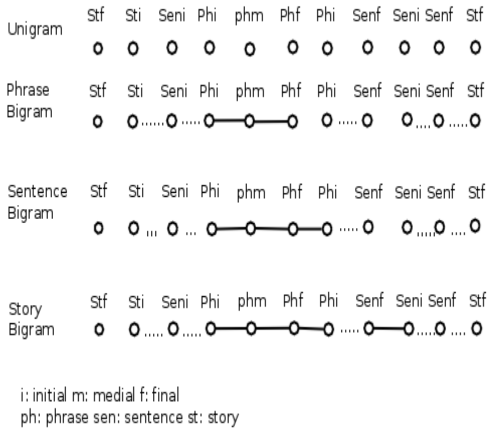

Unigram
Phrase Bigram
Sentence Bigram
Story Bigram
Stf Sti Seni Phi phm Phf Phi Senf Seni Senf Stf
i: initial m: medial f: final
ph: phrase sen: sentence st: story

Figure 2: Four different CRF structures: solid lines indicate bigram connections. A sentence CRF differs from a phrase CRF by connecting a phrase final syllable with its following phrase initial syllable if in the same sentence.

Feature Type	Description	Feature IDs
Pitch	5 uniform points across word Maximum, minimum, mean Differences in max, min, mean	p_0,p_0.25,p_0.5,p_0.75,p_1 pmax, pmin, pmean diff_pmax, diff_pmin, diff_pmean
	Difference b/t boundaries	diff_pitch_endbeg
	Pitch slope Difference b/t slopes	pslope diff_slope_endbeg
Intensity	Maximum, minimum, mean Difference in maxima	imax, imin, imean diff_imax

Table 4: Prosodic features for classification and analysis, first introduced in [5]

4. Contrasting Results

4.1. Baselines and effects of intonational boundaries

Using a one-fifth subset of the data, we conducted an exploratory experiment to assess the impact of intonational boundary effects on tone recognition and to determine the effectiveness of the proposed compensation techniques. Using our standard feature representation, we employed a Support Vector Machine classifier with an RBF kernel to perform tone recognition.

Test set	MS	MW (S)	MW (T)	Z-score
Phrase initial	54.23%	47.55%	41.15%	52.79%
Phrase final	42.34%	26.98%	40.85%	48.07%
Sentence initial	49.01%	43.18%	39.18%	47.55%
Sentence final	43%	28.58%	44.3%	44.65%
Story initial	56.85%	48.13%	35.7%	49.79%
Story final	37.13%	29.78%	50%	48.91%
All position	63.47%	61.05%	64.67%	66.19%

Table 5: Tone recognition results at intonational boundaries using normalization techniques: Mean slope, Moving Window (Syll), Moving Window (Time) and Z-score.

Table 5 compares these three feature normalization techniques with the standard z-score log normalization which does not explicitly include intonational compensation. We show that tones at intonational unit boundaries are much more poorly recognized than those in medial positions. Furthermore, none of the above feature normalization techniques outperforms z-score normalization overall in recognizing tones located at intonation boundaries or in all positions. Those which improve in one position often find these gains offset by poorer performance in others.

4.2. Sequence Modeling Results

We compute the overall accuracy of every CRF by averaging the accuracy from 5 fold cross validation. Accuracy is reported in Table 6. All CRFs incorporating bigram-based dependencies improve over the unigram by 1%. Breaking down accuracy by tone, we observe improvements for all tones under all CRF bigram models, except for 1st (high) tone with phrase CRFs. We further note that the accuracy on 3rd tones improves the most, with gains of up to 4% absolute.

	Unigram	Phrase	Sentence	Story
Total Accuracy	**63.81%**	**64.80%**	**65.08%**	**64.99%**
1st tone Accuracy	62.9%	61.5%	63.8%	64.3%
2nd tone Accuracy	71.0%	72.8%	72.08%	71.6%
3rd tone Accuracy	**38.9%**	**43.2%**	**42.5%**	**42.5%**
4th tone Accuracy	69.02%	69.8%	69.8%	69.6%

Table 6: Overall Accuracy using different CRFs

5. Discussion

5.1. Unigram vs. Phrase CRF on Recognizing Tones at Initial Boundaries

Since our goal in sequential modeling was to improve the relatively poor recognition observed at intonational boundaries, we compare the accuracy of unigram and phrase-based bigram CRFs in recognizing tones located at all three initial intonational boundaries (phrase, sentence and story). We notice the most drastic improvement in recognizing 3rd tones at phrase initial positions, shown in Table 7. While the unigram model almost entirely fails to recognize 3rd tones located at initial boundaries, the phrase-based bigram CRF successfully recognizes 46.7% of these initial 3rd tones, which yields a 39% improvement over unigram.

	1st tone	2nd tone	3rd tone	4th tone
Unigram	56.3%	73.0%	**7.5%**	67.7%
Phrase	50.8%	69.0%	**46.7%**	68.5%

Table 7: Accuracy on tones at all three (phrase, sentence and story) initial boundaries using unigram and phrase bigram CRFs

To understand the improvement in the phrase initial 3rd tone recognition, we compared the feature values of 3rd tones recognized by the unigram model and the phrase bigram Model. Features which differ significantly between these two subsets are shown in Table 8. It is obvious that the 3rd tones recognized by the phrase bigram model have much higher mid point pitch

Significantly Differing Features		
	Unigram	Phrase based CRF
p_0.5	-0.8353	-0.2974**
pslope	-0.0414	-0.0280**
imean	0.0734	0.5220**

Table 8: Contrasting features between tones recognized by unigram and phrase based bigram CRF. Features are marked with ** indicated significant differences with $p <= 0.001$

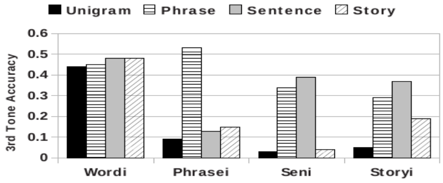


Figure 3: 3rd tone accuracy obtained from four different CRFs and four different initial conditions: word initial(wordi), phrase initial(phrasei), sentence initial(seni) and story initial(storyi)

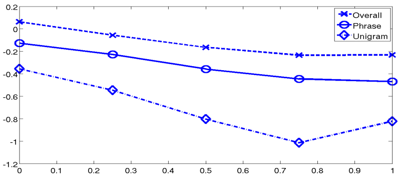


Figure 4: Pitch contours for phrase initial 3rd tones recognized by Phrase Bigram and Unigram models

and much higher mean amplitude compared to those 3rd tones recognized by the unigram model. Considering the accuracy difference in Table 7, we can conclude that the phrase based CRF can encode the pitch level and amplitude variations from phrase initial effects better than the unigram model.

5.2. 3rd Tone Accuracy at All Four Initial Boundary Types

Since 3rd tone is the least frequent tone in the dataset, it is the most poorly modeled. The drastic improvement achieved by phrase based bigram CRF motivates us to look into the 3rd tone accuracy achieved by different CRFs at different initial boundary positions. In figure 3, we show the accuracy we obtained by four CRFs on four distinctive subsets corresponding to four initial conditions: word initial, phrase initial, sentence initial and story initial. The unigram model shows particularly low accuracy on 3rd tone recognition, confuses most of the 3rd tones at sequence initial positions with other tones.

We also found that, for phrase initial and sentence initial 3rd tones, the best performance is achieved by the corresponding sequence-specific CRF. These sequence-specific CRFs model initial tones with a Unigram structure while all other sequence CRFs model them as Bigram+Unigram. This observation shows that this position-specific modeling captures intonational boundary effects, yielding significant improvements for otherwise highly confusable 3rd tones in these positions.

5.3. Pitch-level Confusion of Phrase Initial 3rd Tones

In Figure 4, we compare the average pitch contours of all phrase initial 3rd tones recognized by the unigram model, phrase bigram CRF, and in the overall dataset. Based on the Tukey posthoc test, we found that 3rd tones recognized by the phrase bigram model have significantly higher pitch compared to those recognized by the unigram model. However, the pitch levels of 3rd tones recognized by the phrase bigram model are still lower than the overall dataset of phrase initial 3rd tones. We aim to improve position-specific training in further work.

6. Conclusion

We employed three different bigram CRFs to model broad context effects at boundaries of phrases, sentences and stories. When we compared these sequential CRFs with the baseline zero-order unigram model, we observed that all bigram CRFs improved overall accuracy. The greatest improvement was in recognizing 3rd tones, where all bigram CRFs outperform the unigram model by at least 4%. Further investigation indicated that this sequence modeling approach improves recognition for phrase, sentence, and story initial tones third tones by as much as 39% absolute.

In future work, a natural extension is to employ an alternative graphical model topology that allows position specific training for tones at both initial and final positions with distinct structures. We are also interested in addressing the challenges of general prosodic modeling, by investigating adapting this broad context modeling to other languages and other accent modeling problems with minor changes.

7. References

[1] J. Zhang and K. Hirose, "Tone nucleus modeling for Chinese lexical tone recognition," *Speech Communication*, vol. 42, pp. 447–466, 2004.

[2] E. Chang, J. Zhou, S. Di, C. Huang, and K.-F. Lee, "Large vocabulary Mandarin speech recognition with different approaches in modeling tones," *Proceedings of ICSLP 2000*, pp. 983–986, 2000.

[3] C. Wang and S. Seneff, "Improved tone recognition by normalizing for coarticulation and intonation effects," *Proceedings of ICSLP*, pp. 1–6, 2000.

[4] Y. Qian, T. Lee, and F. K. Soong, "Tone recognition in continuous Cantonese speech using supratone models," *The Journal of the Acoustical Society of America*, vol. 121-5, pp. 2936–45, 2007.

[5] G.-A. Levow, "Context in multi-lingual tone and pitch accent prediction," *Proceedings of Interspeech 2005*, 2005.

[6] G. Peng and W. S.-Y. Wang, "Tone recognition of continuous cantonese speech based on support vector machines," *Journal of Speech Communication*, vol. 45, pp. 49–62, 2005.

[7] T. Lee, W. Lau, Y. Wong, and P. Ching, "Using tone information in Cantonese continuous speech recognition," *ACM Transactions on Asian Language Information Processing*, vol. 1, no. 1, pp. 83–102, 2002.

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

[9] S. Wang and G.-A. Levow, "Mandarin chinese tone nucleus detection with landmarks," *Proceedings of Interspeech 2008*, 2008.

[10] A. Jansen and P. Niyogi, "A probabilistic speech recognition framework based on the temporal dynamics of distinctive feature landmark detectors," The Deparment of Computer Science, Univeristy of Chicago, Tech. Rep. TR-2007-07, 2007.

[11] S. Wang, "Improving tone recognition with nucleus modeling and sequential learning," Ph.D. dissertation, The University of Chicago, 2010.

[12] C. Sutton, "Grmm: Graphical models in mallet." http://mallet.cs.umass.edu/grmm/. 2006.