Improved Tone Modeling for Mandarin Broadcast News Speech Recognition

Xin Lei[1], *Manhung Siu*[2], *Mei-Yuh Hwang*[1], *Mari Ostendorf*[1] *and Tan Lee*[3]

[1]Univ. of Washington, Dept. of Electrical Engineering, Seattle, WA 98195 USA
[2]Hong Kong Univ. of Science and Technology, EEE Dept., Clear Water Bay, HK
[3]Chinese Univ. of Hong Kong, Dept. of Electronic Engineering, Shatin, New Territories, HK
{leixin,mhwang,mo}@ee.washington.edu, eemsiu@ust.hk, tanlee@ee.cuhk.edu.hk

Abstract

Tone has a crucial role in Mandarin speech in distinguishing ambiguous words. Most state-of-the-art Mandarin automatic speech recognition systems adopt embedded tone modeling, where tonal acoustic units are used and F_0 features are appended to the spectral feature vector. In this paper, we combine the embedded aproach (using improved F_0 smoothing) with explicit tone modeling in rescoring the output lattices. Oracle experiments indicate 32% relative improvement can be achieved by rescoring with perfect tone information. Recognition experiments on Mandarin broadcast news show that, even with an accuracy of only 70%, the explicit tone classifier offers complementary knowledge and improves performance significantly. Through the combination of tone modeling techniques, the character error rate on the CTV test set can be improved from 13.0% to 11.5%.

Index Terms: speech recognition, Mandarin, tone modeling.

1. Introduction

Quite different from English and other Western languages, Mandarin Chinese is a tone language which benefits from modeling of lexical tones to distinguish ambiguous words. Many studies have been conducted on how to incorporate tone information in continuous Chinese speech recognition. There are two major approaches: embedded tone modeling and explicit tone modeling [1]. In embedded tone modeling, tonal acoustic units are used and F_0 features are appended to the spectral feature vector (MFCC/PLP) at each frame [2, 3, 4, 5]. In contrast, with explicit tone modeling, tones are independently recognized in parallel to phonetic recognition and then combined in a post-processing stage [1] or integrated back in a global search process [6].

Although embedded tone modeling is very successful in most state-of-the-art Mandarin automatic speech recognition (ASR) systems, it does not exploit the supra-segmental nature of tones. First, a tone spans much longer than a phone and is synchronous with the syllable instead of the phone. Second, a tone depends on the shape of the pitch contour of the syllable. The frame-level F_0 and its derivatives may not be enough to capture this shape. Third, tones are very variable in length and the fixed delta window can not capture the shape well. In this paper, we first describe improved smoothing and normalization of pitch features for embedded modeling in our baseline Mandarin broadcast news (BN) system. Then we propose to use explicit tone models in lattice rescoring to complement the embedded tone modeling approach. As shown in the experiments, although the explicit tone classifier has much lower tone accuracy than the recognizer, it can still improve the recognition performance as a complementary knowledge source.

The rest of the paper is organized as follows: In Section 2, we describe the baseline Mandarin BN system. In Section 3, the improved smoothing and normalization algorithm for pitch features is presented, with results using the embedded approach. In Section 4, explicit tone modeling and recognition results are discussed. Finally, we summarize the key points and propose future work in Section 5.

2. Baseline Mandarin BN system

Training and Test Data: The acoustic models of our baseline Mandarin BN system were trained on 28 hours of Hub-4 data released by the Linguistic Data Consortium (LDC) with accurate transcriptions. The language model was trained using 121M words from three sources: Hub4, TDT[2,3,4], Gigaword(Xinhua) 2000-2004. The test set is the RT-04 evaluation set, which includes a total of 1 hour of data from CTV, RFA and NTDTV broadcast in April 2004 (`eval04`).

Features and Models: Standard 39-dim MFCC features with vocal tract length normalization are generated with the front-end of the SRI DECIPHER speech recognizer. The fundamental frequency F_0 is extracted with ESPS's get_f0 and then passed to a lognormal tied mixture model [7] to alleviate pitch halving and doubling problems. Then a smoothing algorithm similar to [2] is applied and derivatives are computed. The 3-dim F_0 features are appended to the spectral features, resulting in a feature vector of 42-dim.

Finally the features are mean and variance normalized per speaker. We have used a pronunciation dictionary that includes consonants and tonal vowels, with a total of 72 phones. There are only 4 tones in the phone set, with tone 5 mapped to tone 3. The acoustic models are maximum-likelihood-trained, within-word triphone models. Decision-tree state clustering was applied to cluster the states into 2000 clusters, with 32 mixture components per state. The language models are word-level bigram models.

Decoding Structure: The decoding lexicon consists of 49K multi-character words. The test data `eval04` was automatically segmented into 565 utterances. The length of each utterance is between 5 to 10 seconds. Speaker clustering is applied to cluster the segments into acoustically similar clusters. After first pass decoding, the top hypothesis is used for 3-class MLLR adaptation. The adapted results are evaluated in terms of character error rate (CER).

3. Smoothing and Normalization of F_0

Since pitch is present only in voiced segments, the F_0 needs to be interpolated in unvoiced regions to avoid variance problems in recognition. In our baseline Mandarin BN system mentioned in Section 2, the IBM-style smoothing algorithm [2] has been applied. Recently we have changed the F_0 smoothing and normalization algorithm as follows:

1. Interpolate the F_0 contour with piecewise cubic Hermite interpolating polynomial (PCHIP) [8].
2. Take the log of F_0.
3. Moving window normalization (MWN).
4. 5-point moving average (MA) smoothing.

Compared with the general spline interpolation, the PCHIP spline interpolation has no overshoots and less oscillation. In practice, we find the recognition performance is better than with general spline or linear interpolation. The MWN subtracts the moving average of a long-span window, with a window size of 1-2 seconds. The purpose of MWN is to normalize out phrase-level intonation effects, as described in [3, 1]. The moving average smoothing reduces the noise in F_0 features. Figure 1 shows the original raw F_0 and the final feature, together with the tonal syllable alignments. As we can see from the upper plot, the F_0 level of the second tone 1 (in "gen1") is much lower than the first tone 1 (in "zhong1") due to the F_0 declination over the utterance. The processed F_0 features alleviate this problem through the MWN step.

The speech recognition performance for different F_0 features is evaluated on `eval04`, as shown in Table 1. The RFA and NTDTV shows are broadcasted outside mainland China and have more mismatch with the training data. Therefore, the performance on these two sources are significantly worse than the CTV data. Overall, the new

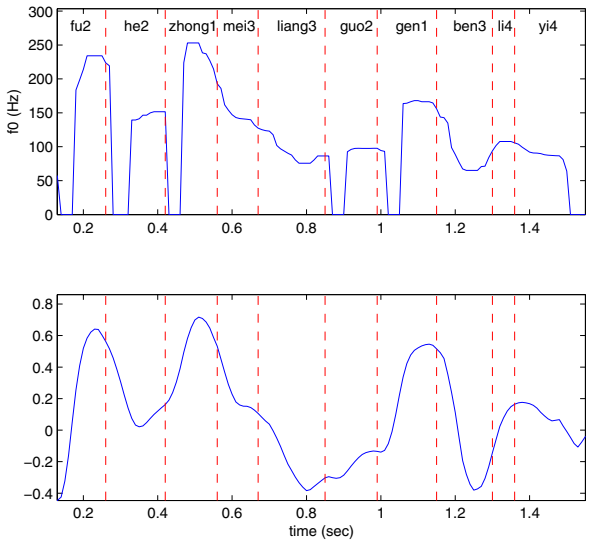

Figure 1: *Raw F_0 contour and the processed F_0 features. The vertical dashed lines show the forced aligned tonal syllable boundaries.*

smoothing and normalization algorithm improves the baseline IBM-style smoothing by 0.8% absolute, with 1.0% absolute improvement on CTV. This improved smoothing is also useful in explicit tone modeling.

Table 1: *CER on* `eval04` *using different F_0 processing.*

Feature	CTV	RFA	NTDTV	Overall
MFCC only	14.0	38.5	21.5	24.1
+ Baseline IBM style F_0	13.0	35.4	19.8	22.2
+ Spline F_0	12.9	35.0	19.7	22.0
+ Spline+MWN+MA F_0	12.0	35.2	18.8	21.4

4. Explicit Tone Modeling

By embedded tone modeling, we have achieved significant improvement in recognition performance. The embedded modeling uses frame-level F_0 as tone features. However, the most important acoustic cue of lexical tones is the segment-level F_0 contour. Therefore, we want to explore whether we can further improve the ASR performance by explicit segment-level tone modeling from rescoring the ASR output lattices of the embedded tone modeling. We choose to rescore lattices instead of n-best lists because a lattice is a much richer representation of the search space.

Because only the CTV portion of the test data is from the same source as the training Hub4 data, the CTV set is used for evaluation of explicit tone modeling. The other two sources, RFA and NTDTV are more conversational (thus

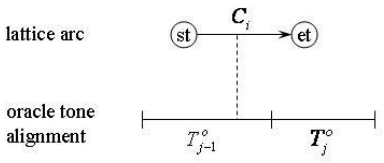

Figure 2: *Aligning a lattice arc i to oracle tone alignments.*

have more tonal variation) and not well matched to the training data, so we expect less benefit from the explicit tone model trained on Hub4. In the following part of this section, we first evaluate the oracle upperbound for explicit tone modeling, then describe the tone classifier, and finally outline the approach for integrating tone scores into lattice rescoring with associated experiments.

4.1. Upperbound evaluation

An error analysis was performed on the CTV test set. Table 2 shows the recognition accuracy of tones, base syllables, tonal syllables and characters, computed from the same decoding run. We find the character errors with correct syllable but wrong tone account for only 0.6% absolute (BS vs. TS). This might lead to the conclusion that by using perfect tone information, we can at most achieve 0.6% improvement. However, different tone decisions might change the phonetic decision since the acoustic units are context-dependent tonal phones.

Table 2: *Accuracy of tones, base syllables (BS), tonal syllables (TS), and characters (Char) on the CTV test set.*

	Tone	BS	TS	Char
Acc. (%)	90.7	89.6	89.0	88.0

To more objectively evaluate the upperbound for tone modeling, we incorporate the perfect tone information in lattice search. Forced alignment is performed against the references to get the oracle tone alignments. For each character in the lattice, we get the oracle tone label according to the center time of the character. As shown in Figure 2, character C_i is aligned to oracle tone T^o_{j-1}. If the tone T_i of C_i is different from the oracle tone T^o_{j-1}, the corresponding arc is pruned in the lattice via applying a large penalty score. Then we re-decode the lattice with the Viterbi algorithm.

The re-decoded top best hypothesis achieves 8.2% CER compared to the baseline 12.0%. This indicates the upperbound for improvement is 3.8% absolute (or 32% relative) if we have a perfect tone recognizer.

4.2. Tone classification

Figure 3 shows the averaged F_0 contours of the four tones from one show in Mandarin BN speech. Similar to the findings in [6], the co-articulation of tones is significant, especially from the left context. Therefore, explicit tone classification is a very challenging problem. Various pattern recognition methods have been tried for Chinese tone recognition, such as decision trees, neural networks, Gaussian mixture models and support vector machines [9, 10, 11, 12]. In this study, we use a single-hidden-layer neural network to train tone classifiers due to the fast training and straightforward integration. The `quicknet` package from ICSI is used in the implementation.

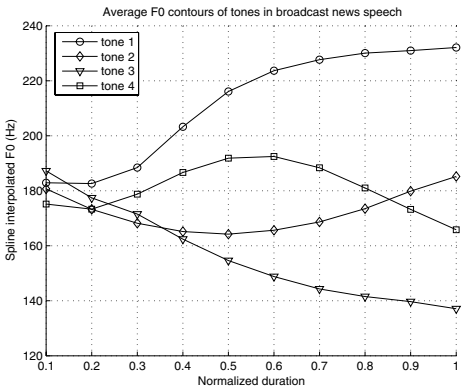

Figure 3: *F_0 contour of four tones.*

The Mandarin syllable consists of two parts: initial (consonant part) and final (vowel part). The features we tried for tone classification are the pitch contour and duration of either the final part or the whole syllable. The pitch contour is processed as described for embedded acoustic modeling, finally sampled to a fixed number of points.

A context-independent 4-tone classifier is trained on all tones longer than 15 frames, since it is almost impossible to distinguish the very short tones due to co-articulation effects. The tone classification results are shown in Table 3. We find that the tone classification accuracy improves significantly by adding the F_0 features in the initial part of the syllable. This is probably because these features contain some information about the co-articulation of the tones. Other features such as the polynomial regression coefficients have also been tried, but no significant improvement was achieved.

We also train context-dependent tone models by classifying the left context into 6 categories: tone 1-4, silence, and noise. Features from the previous syllable are also concatenated onto the feature vector. By using the context-dependent tone models, the 4-tone classification accuracy can be further improved by 1.8% absolute as shown in Table 3.

4.3. Integration of tone scores

The above explicit tone modeling gives around 25% tone error rate. However, from the error analysis in Table 2 we

Table 3: *Four-tone classification accuracy results on CTV data. CI denotes context-independent models. CD denotes context-dependent models.*

Feature	Dim	#of nodes	Acc.
CI: final f_0 + dur	4	35	70.6%
CI: syllable f_0 + dur	7	40	74.4%
CD: syllable f_0 + dur	14	100	76.2%

find the tone error rate of the recognizer is 9.3% on CTV data. In this sense the recognizer is actually a much better tone classifier since it utilizes more complex acoustic and language model information. We propose to use the explicit tone classifier as a complementary knowledge source in lattice rescoring. In this initial work, we only use tone scores from the simple context-independent tone models in the lattice.

For each lattice arc i, which has tone T_i associated with character C_i, the tone score is computed as:

$$\psi_i = w d_i \log p(T_i|f_i) \quad (1)$$

where w is the weight for the tone score, d_i is the number of frames in T_i, and $p(T_i|f_i)$ is the posterior probability of T_i given the tone features f_i. For short tones, a uniform score is used instead of the posterior probability.

A tone weight of smaller than 0.5 gives improved performance. The best CER is 11.5%, achieved with $w = 0.35$. Compared with the embedded modeling, this 0.5% absolute improvement is statistically significant at the level $p = 0.039$ according to the matched pair sentence segment test. It shows that the inferior explicit tone classifier provides complementary information for recognition and improves the system performance significantly. However, there is still a lot of room to improve compared with the oracle bound.

5. Conclusions and Future Work

In this paper, we have described the recent progress on tone modeling in Mandarin BN speech recognition. In embedded tone modeling, better smoothing of F_0 features led to 0.8% improvement on `eval04` and 1.0% on CTV shows. By using the scores from explicit segment-level tone models in lattice decoding, another 0.5% improvement was achieved on CTV data.

Future work includes integrating context-dependent tone modeling in lattice decoding. Tone model adaptation will also be explored. We will also investigate different techniques for combining the embedded and explicit tone models.

6. Acknowledgments

The authors would like to thank BBN for sharing the pronunciation dictionary. Thanks also to Xiao Li in SSLI lab and Tim Ng from BBN for providing useful discussions. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

7. References

[1] T. Lee et al., "Using tone information in Cantonese continuous speech recognition," *ACM Trans. Asian Language Info. Process.*, vol. 1, pp. 83–102, 2002.

[2] C.J. Chen et al., "New methods in continuous Mandarin speech recognition," in *Proc. Eur. Conf. Speech Communication Technology*, 1997, vol. 3, pp. 1543–1546.

[3] H.C. Huang and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition," in *Proc. ICASSP*, 2000, vol. 3, pp. 1523–1526.

[4] E. Chang et al., "Large vocabulary Mandarin speech recognition with different approaches in modeling tones," in *Proc. ICSLP*, 2000, vol. 2, pp. 983–986.

[5] M. Hwang et al., "Porting DECIPHER from English to Mandarin," in *Proc. DARPA 2004 Rich Transcriptions Workshop*, 2004.

[6] C. Wang, *Prosodic Modeling for Improved Speech Recognition and Understanding*, Ph.D. thesis, Massachusetts Institute of Technology, 2001.

[7] M.K. Sonmez et al., "A lognormal model of pitch for prosody-based speaker recognition," in *Proc. Eur. Conf. Speech Communication Technology*, 1997, vol. 3, pp. 1391–1394.

[8] F.N. Fritsch and R.E. Carlson, "Montone Piecewise Cubic Interpolation," *SIAM J. Numerical Analysis*, vol. 17, pp. 238–246, 1980.

[9] P.F. Wong and M. Siu, "Decision tree based tone modeling for chinese speech recognition," in *Proc. ICASSP*, 2004, vol. 1, pp. 905–908.

[10] S.H. Chen and Y.R. Wang, "Tone recognition of continuous Mandarin speech based on neural networks," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 146–150, 1995.

[11] Y. Qian, *Use of Tone Information in Cantonese LVCSR Based on Generalized Posterior Probability Decoding*, Ph.D. thesis, The Chinese University of Hong Kong, 2005.

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