EXPLOITING PROSODIC AND LEXICAL FEATURES FOR TONE MODELING IN A CONDITIONAL RANDOM FIELD FRAMEWORK

Hongxiu Wei, Xinhao Wang, Hao Wu, Dingsheng Luo, Xihong Wu*

Speech and Hearing Research Center
State Key Laboratory of Machine Perception, Peking University
Beijing, 100871, China
{weihx,wangxh,wuhao,dsluo,wxh}@cis.pku.edu.cn

ABSTRACT

Tonal cues play an important role in distinguishing ambiguous words in Mandarin speech recognition. This paper explores an innovative tone modeling framework using prosodic and lexical features, as well as syllable context information. A discriminative model, namely a Conditional Random Field (CRF), is adopted, which is sufficiently flexible to handle multiple interacting features and long-range dependencies of observations. After the first pass search of a recognition system, the CRF based tone models are employed to rerank N-best hypotheses according to the tonal scores which can represent the correctness of the tone sequence given each candidate hypothesis and the observed speech signal. Experiments results show that the tonal cues help to achieve 7.8% and 8.6% relative reductions of character error rate on two widely used Mandarin speech recognition tasks, Hub-4 test and 863 test.

Index Terms— Mandarin speech recognition, tone modeling, CRF, reranking

1. INTRODUCTION

In contrast to English, Mandarin is a tonal syllabic language, in which tones have an essential role in lexical disambiguation. Tonal cues have also been proved useful in automatic speech recognition[1, 2, 3, 4, 5, 6, 7].

Research on how to incorporate tonal cues into speech recognition has focused on two main areas: embedded tone modeling, and explicit tone recognition [2]. With the embedded tone modeling, tonal and spectral variations are modeled within a uniform framework. In a single-stream system, the spectral features, such as MFCC (Mel-Frequency Cepstral Coefficient) and PLP (Perceptual Linear Predictive coefficient), are appended by additional F0 (Fundamental frequency) features[1, 3]. While, in a multi-stream system, MFCC and F0 are separated into two streams assigned with different stream weights[4]. On the contrary, in the explicit tone recognition, tone patterns are independently modeled and recognized in parallel to phonetic recognition. In combination with the spectral scores, the tone recognition scores can be directly integrated in the first pass search process[7], or used to rescore the N-best recognition results[5] or word lattices[6] of the phonetic recognizers. Tonal information also provides phonological and lexical constraints, and therefore the N-best output of tone recognition can be utilized to expand a toneless syllable lattice to a tonal syllable lattice [2].

Prosodic features, including F0, energy and duration, are commonly used in the work of tone modeling. However, they are not sufficient to model the complex changes of tones in continuous speech. Studies have proven that with the prosodic, lexical and syntactic features, the performance of prosody labeling can be improved[8, 9]. Thus it may be effective to construct tone models with higher level and longer range features. This paper describes a framework of building explicit tone models that exploit both prosodic and lexical features, as well as syllable context information. Due to the poor performance of parsers on imperfect hypotheses, syntactic features are not used. After the first pass search of a recognition system, the tone models are employed to score the correctness of the tone sequence for each N-best hypothesis. The tonal scores combined with other scores are then used to rerank N-best hypotheses. Consequently, experimental results show the efficiency of our framework in incorporating tonal cues to improve the recognition performance.

For modeling tones, this paper adopts a discriminative model - Conditional Random Field (CRF)[10, 11], which has been successfully implemented in information extraction[11], shallow parsing[12], phonetic recognition[13] and many other fields. Compared with the conventional tone modeling methods, the advantages of CRFs lie in two major aspects[10, 11]: Firstly, they relax the very strict independence assumptions on the observations made by generative models like Hidden Markov Models (HMMs), hence can robustly handle the overlapping and long-range dependent features of observations. Secondly, unlike classifiers, such as Support Vector Machine (SVM), neural networks, and decision trees, which make the

*Hongxiu Wei graduated from Peking University in Jul, 2007 and is now with the Research & Development Center, TOSHIBA (China).

best decision on each local point of observations, CRFs are optimized globally to make an optimal decision on the whole sequence.

This paper is organized as follows: In Section 2, the choice and the derivation of tonal features are introduced. Then, the tone modeling method using CRFs is presented and the approach of reranking the N-best hypotheses with tonal cues is presented in Section 3. Afterwards, the database and the baseline system used for evaluation are described, and the experimental results are discussed in Section 4. Finally, the conclusions and future work are drawn in Section 5.

2. FEATURE SELECTION AND EXTRACTION

2.1. Analysis of Tonal Variations

Syllable based prosodic features, including F0 contour features, average log-energy and duration, are used in this study. Additionally, other possible factors contributing to tone pattern variations are taken into consideration.

One of the main factors affecting to tone pattern variations is co-articulation. Due to articulatory constraints, the exact acoustic realizations of tones are determined not only by the properties themselves, but also by their contexts. Therefore it is natural to combine the features of the processing syllable with those of its neighboring syllables. In addition, the neighboring tone types should also be considered as necessary features.

Besides co-articulation, tone pattern variations of a syllable are also influenced by the location of the syllable within a word. A syllable in the boundary of a word suffers different co-articulation effect from those inside the word. Therefore the location of the syllable within a word is helpful to tone modeling for continuous speech.

2.2. Feature Extraction

Considering the above factors contributing to Mandarin tone variations, a number of prosodic and lexical features were chosen and are listed in Table1.

In training, forced-alignment is performed against the references to get the syllable segmentations information. For test utterances, syllable segmentations are obtained after the first pass recognition.

To extract syllable based F0 contour features, the F0 values are firstly extracted using the cepstrum-based algorithm, and then the Hermite interpolation is applied to the unvoiced regions to get a continuous F0 contour. For each syllable, a two-order polynomial function is employed to fit the continuous F0 contour, and the fitting parameters are viewed as F0 contour features for the syllable. It has been reported that only certain portions of the F0 contours satisfy the canon definitions of tones due to co-articulation effects, and the other regions just represent the co-articulatory transitions from previous to following tone[14]. Therefore, for simplicity, the syllable is segmented into three equal parts, where the first third is discarded as the transition, and the last two thirds is used to get the fitting features. The average energy and its deviation of the finals are also included as features, as well as the duration of the syllable.

Table 1. Features used for tone modeling.

1	Duration of the syllable being processed	1 feature
2	Fitting parameters of the F0 contour of the syllable being processed	3 features
3	Log-energy mean and its deviation of the final of the syllable being processed	2 features
4	The same three kinds of features (i.e., duration, log-energy mean and deviation, fitting parameters of F0 contour) extracted from the preceding and following syllables	12 features
5	Location within a word of the syllable being processed	1 feature
6	Tone types of the preceding and following syllables	2 features

3. TONE MODELING AND N-BEST RERANKING

3.1. Tone Modeling Based on CRFs

This study uses CRFs to build the tone models, which are sufficiently flexible to use the various features mentioned in section 2. Here it should be noted that the features of F0 contour, duration and log-energy are quantized according to their priori probability distributions in advance.

A CRF [10, 11] takes the form of undirected graphical model that directly defines a conditional probability distribution over label sequence Y, corresponding to the tone sequences in this task, given an observation sequence X, corresponding to the speech signal and the candidate hypothesis. And it is specified by a particular chosen set of feature functions $f_i(Y_{t-1}, Y_t, X, t)$, parameterized by λ_i with respect to the importance of each feature function. Then the conditional probability takes the form as

$$p(Y|X,\Lambda) = \frac{\exp(\sum_t \sum_i \lambda_i f_i(Y_{t-1}, Y_t, X, t))}{Z(X,\Lambda)} \qquad (1)$$

where each feature $f_i(Y_{t-1}, Y_t, X, t)$ is either a state feature $s(Y_t, X, t)$ or a transition feature $t(Y_{t-1}, Y_t, X, t)$; $Z(X,\Lambda)$ is a normalization factor, obtained by accumulating all possible label sequences of the observation sequence, and independent with the label sequence Y.

Given a syllable sequence S corresponding to a hypothesis W, for the ith syllable s_i associated with tone t_i, the probability of label $T_i = t_i$: $p(t_i|S)$ can be computed through

forward-backward algorithm. This probability is a kind of measurement for the tone tag $T_i = t_i$, therefore it can be regarded as the tonal score of s_i. Then the total tone score of the hypothesis W is computed as:

$$\log\left(P\left(T|W,S\right)\right) = \omega \sum_{i=1}^{N} \log\left(P\left(t_i|S\right)\right) \qquad (2)$$

where ω is the weight for the tone score, T is the tone labeling sequence, and N is the number of syllables in S.

3.2. N-best hypotheses reranking

The tone models are incorporated after the first pass recognition, and N-best hypotheses are reranked according to the total path scores adjusted with the tonal scores as follows

$$\hat{W} = \arg\max_{W} \left(\begin{array}{l} \log\left(P\left(O|W\right)\right) + \log\left(P\left(W\right)\right) \\ +\log\left(P\left(T|W,S\right)\right) \end{array} \right) \qquad (3)$$

where $P\left(O|W\right)$ and $P\left(W\right)$ respectively represent the acoustic score and the language score. By the reranking paradigm, a new best sentence hypothesis is given.

4. EXPERIMENTS

4.1. Experimental Setup

The acoustic models (AM) of the baseline system consist of context-dependent Initial-Final models. The left-to-right model topology is used to represent each unit. The number of states in each model is set to 2 or 3 for initials, and 4 or 5 for tonal finals, according to the corresponding phonetic structures. Each state is trained to have 32 Gaussian mixtures. The used 39-dimension feature vector consists of 12 MFCC coefficients, energy, and their first-order and second-order deltas. In our experiments, the acoustic models and the tone models are trained with a speech database which contains about 360 hours speech of over 750 male speakers. This training data is picked up from three widely used continuous Mandarin speech corpora: the 863-I, 863-II and Intel corpora. The brief information about these three speech corpora is listed in Table 2. The language model is a word-based trigram built on a vocabulary of 60,000 words. The training text consists of about 1.5 billion characters from Chinese newspapers and Internet.

Table 2. The information of speech training data

Corpus	Speakers	Amount of Speech (hours)
863-I (male)	83	56.67
863-II(male)	120	78.08
Intel (male)	556	227.30
total	759	362.05

The baseline system employs a one-pass search by a forward Viterbi algorithm to generate a word lattice, which is then converted into N-best lists.

To build the CRF tone models, the Yet Another CRF++ 0.45 toolkit[1] is employed. This open source implementation of CRFs uses a quasi- Newton LBFGS algorithm to perform the gradient minimization.

4.2. Use of Tone Models

In our experiments, the clean male speech data from two corpora, including the test set of the 2005 continuous Mandarin speech recognition evaluation held by the national "863" program of China(863 Test Set) and the test set of the 1997 Hub-4 Mandarin broadcast news evaluation (Hub-4 Test Set), are adopted as the evaluation data. For the 863 corpus, the speech data above 15 dB SNR is viewed as clean data. For the Hub-4 corpus, the partitioning for clean data is done with the acoustic segmentation software CMUseg_0.5[2]. The brief information about these two evaluation corpora is listed in Table 3.

Table 3. The information of speech test data

Corpus	863 Test	863 Dev	Hub-4 test	Hub-4 Dev
sentences	209	174	192	544

The weights of the tone models for each of the two test sets are empirically derived from the development sets described above. The recognition results are shown in Table 4.

Table 4. Speech recognition results

Test Sets	System	Err.	Sub.	Del.	Ins.
Hub-4	Baseline	16.76	14.85	1.01	0.91
	Tonal incorporation	15.45	13.69	1.21	0.55
863	Baseline	22.10	19.69	1.36	1.05
	Tonal incorporation	20.20	17.85	1.56	0.79

For the Hub-4 test set, the experimental results show a character error rate (CER) of 16.76% with no special tone processing in the AM of the baseline system. When the tonal cues are incorporated into the post-processing stage, a 7.8% relative reduction is achieved. As for the 863 test set, compared with the 22.10% CER in the baseline system, the relative reduction is 8.6%. What is important is that the substitution error rate is reduced from 14.85% to 13.69% for the Hub-4 data, and from 19.69% to 17.85% for the 863 data, which is

[1] Yet Another CRF++ 0.45: CRF toolkit for segmenting/labeling sequential data written by Taku Kudo in C++, http://chasen.org/∼taku/software/CRF++.

[2] CMUseg_0.5:Acoustic segmentation software downloaded from http://www.nist.gov/speech/tools/CMUseg_05targz.htm.

consistent with the expectation that tonal cues can effectively help lexical disambiguation in Mandarin speech recognition.

5. CONCLUSIONS AND FUTURE WORK

This paper addresses an investigation into the utility of tonal cues in Mandarin automatic speech recognition. Implementation of the proposed framework produces 7.8% and 8.6% relative reduction in character error rate over the Hub-4 and 863 test data, respectively. The results show large potential for using CRFs for this topic, and prove the efficiency of our method. The combination of prosodic features and lexical features are shown to be appropriate, too.

To handle prosodic features for CRFs, quantization is applied in this paper. Obtaining more discriminative features is a key area of interest for feature work. We would also like to investigate additional types of features for tone modeling, especially linguistic features, such as Part-of-Speech (POS) and syntactic structures of the hypotheses with errors. Finally, since the proposed approach would be promising for incorporating tonal information into automatic speech recognition, we are also interested in seeing if other information sources, such as duration and pauses, may be effective within the same framework.

6. ACKNOWLEDGMENTS

The work was supported in part by the National Natural Science Foundation of China (60435010; 60535030; 60605016), the National High Technology Research and Development Program of China (2006AA01Z196; 2006AA010103), the National Key Basic Research Program of China (2004CB318005), the New-Century Training Program Foundation for the Talents by the Ministry of Education of China, and a Joint Project with Microsoft Research Asia.

7. REFERENCES

[1] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New methods in continuous Mandarin speech recognition," in *Proc. Eurospeech*, 1997, vol. 3, pp. 1543–1546.

[2] T. Lee, W. Lau, Y.W. Wong, and P.C. Ching, "Using tone information in Cantonese continuous speech recognition," *ACM Transactions on Asian Language Information Processing*, vol. 1, pp. 83–102, March 2002.

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

[4] Y. Sun, D. Willett, R. Brueckner, R. Gruhn, and D. Bühler, "Experiments on Chinese speech recognition with tonal models and pitch estimation using the Mandarin speech data," in *Proc. ICSLP*, 2006, pp. 1245–1248.

[5] G. Peng and William S.Y. Wang, "An innovative prosody modeling method for Chinese speech recognition," *International Journal of Speech Technology*, vol. 7, pp. 129–140, April 2004.

[6] X. Lei, M. Siu, M.Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for Mandarin broadcast news speech recognition," in *Proc. ICSLP*, 2006, pp. 1237–1240.

[7] Y. Cao, Y. Deng, H. Zhang, T. Huang, and B. Xu, "Decision-tree based Mandarin tone model and its application to speech recognition," in *Proc. ICASSP*, 2000, vol. 3, pp. 1759–1762.

[8] M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S.S. Kim, A. Cohen, T. Zhang, J.Y. Choi, H. Kim, T. Yoon, and S. Chavarria, "Simultaneous recognition of words and prosody in the Boston university radio speech corpus," *Speech Communication*, vol. 46, pp. 418–439, 2005.

[9] V. Rangarajan, S. Narayanan, and S. Bangalore, "Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework," in *Proc. NAACL-HLT*, 2007, pp. 1–8.

[10] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001, pp. 282–289.

[11] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. MIT Press, 2006.

[12] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proc. HLT-NAACL*, 2003, pp. 213–220.

[13] J. Morris and E. Fosler-Lussier, "Combining phonetic attributes using conditional random fields," in *Proc. ICSLP*, 2006, pp. 597–600.

[14] Y. Xu and Q. Emily Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Communication*, vol. 33, pp. 319–337, March 2001.