MAXIMUM ENTROPY BASED TONE MODELING FOR MANDARIN SPEECH RECOGNITION

Xinhao Wang, Yansuo Yu, Xihong Wu, Huisheng Chi

Speech and Hearing Research Center
Key Laboratory of Machine Perception (Ministry of Education)
School of Electronics Engineering and Computer Science
Peking University, Beijing, 100871, China
{wangxh,yuys,wxh}@cis.pku.edu.cn, chi@pku.edu.cn

ABSTRACT
To explore the potential of prosody for Mandarin speech recognition, this paper addresses the tone modeling problem and its integration issue. This study adopts the maximum entropy approach to capture both acoustic and lexical characteristics of tones due to its flexibility in handling multiple interacting features. Moreover, considering the phoneme factor, besides a tone model, a phoneme dependent model is also constructed. With regard to the model integration, the presented models are integrated into the recognizer under the one-pass decoding framework, where they are used to prune the active word-final states during beam search. Experimental results on the HUB-4 evaluation material reveal the effectiveness of the presented models. They significantly improve the performance of speech recognition with 7.6% and 11.1% relative reduction of character error rate.

Index Terms— Mandarin speech recognition, Tone modeling, Maximum Entropy, One-pass decoding

1. INTRODUCTION
In contrast to English and other Western languages, Mandarin is a tonal monosyllabic language, in which each Chinese character is pronounced in a single syllable with an associated tone pattern, from four different lexical tones and a neutral tone. As tonal cues play an essential role in lexical disambiguation, much recent research has focused on tone modeling in the study of prosody investigation in large vocabulary Mandarin speech recognition.

One typical way is to incorporate the pitch features based on the framework of Hidden Markov Model (HMM) (e.g.[1, 2]). In a single-stream system, the additional F0 features are appended to the short-time spectral features, such as MFCC and PLP. While in a multi-stream system, spectral features and F0 can be separated into two streams, assigned with different stream weights. However, the frame-level F0 features, extracted from a normal time window as in spectral features, are not sufficient to capture the shape of the pitch contour for a syllable. Hence, the supra-segmental characteristics of pitch based on syllables can not be well described.

On the contrary, there is another way to model the tonal information, where the tone patterns are modeled independently and recognized in parallel with phonetic recognition (e.g. [3, 4]). When these tone models are integrated into the speech recognizer, they usually derive the acoustic features from the force-aligned syllable boundaries after the first-pass recognition, and are commonly used to rescore the word lattice or N-best hypothesis in a second-pass decoding.

Unlike the tone pattern in an isolated syllable, due to co-articulation, the tone pattern may vary according to its different context in continuous speech utterances. Studies have proven that with the acoustic, lexical and syntactic features, the performance of prosody labeling can be improved (e.g. [5]). Thus it may be effective to construct tone models with not only the acoustic features (F0, duration and energy), but also higher level and longer range linguistic information. Considering that the maximum entropy model can robustly handle the overlapping and long-range dependent features of observations, this study proposes maximum entropy based tone modeling method to exploit both the acoustic and lexical features.

Furthermore, it can be observed that the pitch contour of tones may be different when attaching them to different phonemes in continuous speech. As in [6], the phoneme dependent model outperformed the phoneme independent model in the task of tone recognition. In this contribution, two different kinds of tone models are built, namely tone-model and tonal-syllable-model, which consider the tone patterns or tonal syllables as category labels respectively. Finally, these two models are integrated into speech recognizer under the one-pass framework, in which the tonal cues are employed to score the tonal correctness of each candidate word, and then the tonal score is combined with acoustic and language scores to prune the current active space. That is, the tone models directly participate in the generation of word lattice and N-best hypotheses. At last, the presented models and the adopted integration manner are evaluated to be effective by a series of experiments on the widely used HUB-4 Mandarin speech recognition task.

The remainder of this paper is structured as follows. In Section 2, the maximum entropy based tone models and selected features are introduced in detail. Afterwards, under the one-pass recognition framework, the tone models are incorporated into a Mandarin speech recognition system in Section 3. In Section 4, a series of experiments are carried out to evaluate the proposed models. Finally, the conclusion is drawn in Section 5.

2. TONE MODELING
In Mandarin, the syllables are generally considered as the basic prosodic units. Since the tone pattern variations of syllables are greatly influenced by their linguistic context, especially the word it belongs to, the ideal case is to build a tone model for each distinct word. However, this is unfeasible due to data sparseness and intractable parameter estimation. By means of deriving lexical features from words, a word-related tone model can be achieved.

As a discriminative model, the maximum entropy (MaxEnt)
model [7] has been successfully employed in various labeling tasks of natural language processing, and it also achieves state-of-the-art results in the task of prosody labeling [5]. Compared with the generative models, such as HMM, the MaxEnt model directly optimizes the conditional distribution, and relaxes the very strict independence assumptions on the observation, hence it can robustly handle the overlapping and long-range dependent features of observations. Accordingly, the maximum entropy approach is an ideal choice for our tone modeling.

2.1. MaxEnt based Tone Models

Given a word with $K$ syllables, $w = s_1 s_2 \cdots s_K$, and corresponding observation sequence $x$, the conditional probability for each syllable $s_k$ defined by MaxEnt takes the form as

$$p(s_k|x, \Lambda) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{m} \lambda_i f_i(x, s_k)\right)$$

(1)

where each selected $f_i(x, s_k)$ is an indicator function, parameterized by $\lambda_i$, with respect to the importance of this feature function; $Z(x)$ is a normalization factor, which ensures that $\sum_{s_k} p(s_k|x) = 1$. Then the probability of the word can be calculated as

$$P_{\text{m}}(w|x, \Lambda) = \prod_{k=1}^{K} p(s_k|x, \Lambda)$$

(2)

In this work, two different tone models are constructed, which respectively take the tone patterns and the tonal syllables as the class labels. Given the syllable $s_k$ associated with tone $t_k$, the tone-model calculates the probability of $t_k$ conditioned on the observation. Nevertheless, the other model is a phone dependent model, i.e., tonal-syllable-model, which calculates the conditional probability of tonal syllables $s_k$. By utilizing Equation 2, the word score given by either of these two models, $P_{\text{pm}}(w|x)$, can be used to measure the tonal correctness of a candidate word.

2.2. Feature Selection

Considering various factors contributing to Mandarin tone variations, not only the universally used acoustic features, including F0, duration and log-energy, but also a number of lexical features are chosen in our models as listed in Table 1.

<table>
<thead>
<tr>
<th>Feature Classes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic features (AF)</td>
<td>A. F0 mean and its deviation of the current final; Quadratic fitting parameters for the F0 contour of the current final.</td>
</tr>
<tr>
<td></td>
<td>B. Log-energy mean and its deviation of the current final; Quadratic fitting parameters for the log-energy contour of the current syllable.</td>
</tr>
<tr>
<td></td>
<td>C. Duration of the current syllable.</td>
</tr>
<tr>
<td>Lexical features (LF)</td>
<td>A. Syllable location within the current word; Number of syllables in the current word.</td>
</tr>
<tr>
<td></td>
<td>B. Current syllable.</td>
</tr>
<tr>
<td></td>
<td>C. Preceding and following syllables; Tone patterns of the preceding and following syllables.</td>
</tr>
<tr>
<td></td>
<td>D. Preceding and following tonal syllables.</td>
</tr>
</tbody>
</table>

To extract pitch related features, firstly the F0 values of the voiced speech are estimated by the cepstrum-based method. Then the linear interpolation is applied to the unvoiced regions to get a continuous pitch contour. Afterwards, the pitch contour is processed by mean subtraction and variance normalization in logarithm scale, and a low-pass filter is used to smooth the pitch 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 of the syllable. Similarly, the quadratic fitting parameters of the log-energy contour within the syllable are also extracted as features. It should be noted that the features of F0 contour, duration and log-energy are quantized according to their priori probability distributions in advance.

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 take the neighboring tone patterns as necessary features. Furthermore, as in Table 1, the lexical features of word length and syllable location within a word are also exploited to capture the lexical information for tone modeling.

3. INTEGRATING TONE MODELS INTO LVCSR

In recent years, the Weighted Finite State Transducers (WFSTs) have been shown to be effective in speech recognition and several WFST-based recognizers have been successfully constructed, such as systems developed by the IBM [8] and AT&T [9]. In this paper, under the framework of WFSTs, a speech recognizer is built for exploiting the potential of tonal cues in improving Mandarin speech recognition.

Since the models described in section 2.1 represent the tonal characteristics of words, they can be employed to evaluate the word hypotheses and assist to prune the active space during Viterbi beam search. At each time frame, a huge number of active states will be generated. Towards the states arriving at word ends, the path scores accumulated from the start to current frame will be adjusted with the tonal scores as follows:

$$\text{score}_{\text{path}}(W) = \log(P_{\text{am}}(O|W)) + \alpha * \log(P_{\text{m}}(W)) + \beta * \sum_{i=1}^{m} \log(P_{\text{pm}}(w_i|x))$$

(3)

where, e.g., for one partial path with $t$ words, $W = w_1 w_2 \cdots w_t$, $P_{\text{am}}(O|W)$ and $P_{\text{m}}(W)$ are the acoustic and language scores, and $P_{\text{pm}}(w_i|x)$ is the tonal score for each word in this path, which is scored by either the tone-model or the tonal-syllable-model as in Equation 2. Then this refined score will be used to cut out the un-promising word-final states. Hence, the tone models have been applied in the one-pass decoding.

4. EXPERIMENTS AND DISCUSSION

This work carries out a series of experiments to test the effectiveness of our models from several aspects, including the improvement of speech recognition performance by employing tonal cues, the advantages of introducing additional lexical features and the superiority of word-based pruning strategy in the first-pass decoding over the rescoring/reordering scheme in the second-pass search.
4.1. Experimental Setup

The acoustic models are context-dependent Initial-Final models. According to the corresponding phonetic structures, the number of states in each model is set to 2 or 3 for initials, and 4 or 5 for tonal finals. Each emitting state is described by 32 Gaussian mixtures. The used 39-dimension feature vector consists of 12 MFCC coefficients, energy, and their first-order and second-order delta. The language model is a word-based trigram built on a vocabulary of 57K entries.

The MaxEnt based tone models are constructed by employing the Maximum Entropy Toolkit, provided by Zhang Le\(^1\). In addition, the integration weights of the tone models are learned according to the criterion of minimum error rate (MER) [10], which has been widely used in machine translation.

The baseline is a WFST-based decoder, which efficiently integrates the acoustic and language models as well as the pronunciation lexicon. Moreover, two new systems are developed by respectively incorporating the tone-model and the tonal-syllable-model into the baseline recognizer during the first-pass decoding.

The 1997 DARPA/NIST Mandarin continuous speech recognition broadcast news HUB-4 benchmark from LDC\(^2\) is adopted, from which the clean data (full bandwidth speech above 20db SNR) are selected for evaluation. Accordingly, 19113 utterances from the training set are used to train the tone models; 1363 utterances from the development set are used in MER to get the weights of tone models; 654 utterances from the test set, including 230 for male speakers and 424 for female speakers, are recognized in the experiments.

4.2. Experimental Results and Discussion

4.2.1. Improvement from Tonal Cues

Performance comparisons among the baseline and two new systems incorporated with tonal cues are conducted, and significant improvements are achieved as shown in Table 2. The baseline system has a character error rate (CER) of 23.95%. When the tone-model is integrated with the weight of 4.66, a 7.6% relative reduction is obtained. While as to the tonal-syllable-model (the integration weight is 0.484), a more significant 11.1% relative reduction of CER is achieved.

<table>
<thead>
<tr>
<th>System</th>
<th>Err.(%)</th>
<th>Sub.(%)</th>
<th>Del.(%)</th>
<th>Ins.(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>23.95</td>
<td>20.96</td>
<td>2.79</td>
<td>0.19</td>
</tr>
<tr>
<td>tone-model</td>
<td>22.13</td>
<td>19.10</td>
<td>2.84</td>
<td>0.19</td>
</tr>
<tr>
<td>tonal-syllable-model</td>
<td>21.30</td>
<td>18.29</td>
<td>2.87</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Furthermore, we run the statistical significance test to detect the performance improvement, in which the approximate randomization approach [11] is modified to output the significance level, p-value, for the CER metric, and all the produced p-levels are less than 0.001. Therefore, the performance improvements imposed by the utilization of tonal cues are statistically significant. Meanwhile, the experimental results have illustrated that the recognizer benefits more from the tonal syllables than the tone patterns, and it is reasonable to build a word-related tone model.

4.2.2. Feature Comparison

To examine the function of additional lexical features as described in section 2.2, two experiments are carried out.

A tone recognition system is constructed, in which the correct transcription is segmented into words for the extraction of lexical features, and aligned with the speech to get the time-alignment information for the extraction of acoustic features. The tone recognizer is tested on 1363 utterances from the development set. As shown in Table 3, the accuracy of tone recognition is only 54.5% with all acoustic features. But when the lexical features are involved, the accuracy can be greatly improved to 81.3%.

<table>
<thead>
<tr>
<th>Features</th>
<th>AF-A</th>
<th>AF-ABC</th>
<th>AF-ABC+ LF-AB</th>
<th>AF-ABC+ LF-ABC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy(%)</td>
<td>53.1</td>
<td>54.5</td>
<td>77</td>
<td>81.3</td>
</tr>
</tbody>
</table>

With respect to the speech recognition task, the lexical features also exhibit their capability, where the 230 male utterances are adopted for evaluation. As illustrated in Figure 1, the additional lexical features greatly improve both these two models. Especially in the tonal-syllable-model, the acoustic features just bring 0.88% reduction of CER compared to the baseline system. Nevertheless, exploiting both the acoustic and lexical features can achieve 2.56% reduction of CER.

4.2.3. One-pass vs. Two-pass Search

To dig out the superiority of the one-pass search framework, several experiments are performed. We firstly observe the performance variation over different sizes of relative beam in the Viterbi search, as illustrated in Figure 2. It is obvious that both the tone-model and the tonal-syllable-model consistently outperform the baseline.

Under the one-pass search framework, the tone models are used to evaluate the tonal correctness of each word and help to prune the

---

\(^1\)Maximum Entropy Modeling Toolkit for Python and C++ downloaded from http://homepages.inf.ed.ac.uk/s0450736/MaxEnt_toolkit.html

\(^2\)http://www.ldc.upenn.edu/
active state space, therefore, a better hypothesis word list can be preserved at each frame during Viterbi beam search. In order to compare two integration strategies, besides the utilization in the first-pass decoding, the tone models are also used to rescore the word lattice in a second-pass decoding. As illustrated in Figure 3, integrating tone models in the first-pass decoding is superior to that in the second-pass search. This superiority can be further confirmed by the oracle.

![Fig. 2. Performance comparison over different relative beam sizes.](image)

![Fig. 3. Comparison of two integration strategies of tone models](image)

accuracy rate on the $N$-best hypotheses. When $N$ is set to 512, the oracle result of baseline is 89.67%. The incorporation of tone-model and tonal-syllable-model improves the oracle results to 90.31% and 91.32% respectively. Thus, the integration in the one-pass search can not only improve the recognition accuracy, but also generate better word lattice or $N$-best hypotheses, which are important for the subsequent work, such as spoken document retrieval.

5. CONCLUSION

This study presents two explicit maximum entropy based tone models, which respectively capture the tone and tonal syllable information within particular words. Under the one-pass framework, they are employed to prune the active word-final states. The new approach mainly benefits from two aspects. Firstly, besides the acoustic observation, the lexical information is also captured in our tone modeling. Secondly, the one-pass integration strategy leads that a better active space can be preserved by utilizing the tonal cues during Viterbi beam search. Therefore, the significant performance improvements are achieved as illustrated by simulations. Experimental results on the HUB-4 Mandarin Evaluation benchmark show that the tone-model and the tonal-syllable-model bring 7.6% and 11.1% relative reduction of character error rate respectively. And the effectiveness of our approach is also confirmed by a series of further analysis and investigation experiments.

6. REFERENCES


Acknowledgments

The work was supported in part by the National Natural Science Foundation of China-NSFC (60535030, 60605016), the National Key Basic Research Program of China-NKBPC (2004CB318005, 2004CB318105, 2009CB825404), and the National High Technology Research and Development Program of China-NHTRDPC (2006AA010103).