WORD-LEVEL TONE MODELING FOR MANDARIN SPEECH RECOGNITION

Xin Lei, Mari Ostendorf

Department of Electrical Engineering
University of Washington, Seattle, WA 98195
{leixin,mo}@ee.washington.edu

ABSTRACT

Standard HMM-based Mandarin speech recognition systems do not exploit the suprasegmental nature of tones, but explicit tone models can be incorporated with lattice rescoring. This work extends previous approaches to explicit tone modeling from the syllable level to the word level, incorporating a hierarchical backoff. Word-dependent tone models are trained to explicitly model the tone coarticulation within the word. For less frequent words, tonal-syllable-dependent tone models or plain tone models are used as backoff. More generally, context-dependent tone models can be used as backoff. The word-dependent tone modeling framework can be viewed as a generalization of the traditional context-independent and context-dependent tone modeling. Under this framework, different types of tone modeling strategies are compared experimentally on a Mandarin broadcast news speech recognition task, showing significant gains from the word-level tone modeling approach.

Index Terms— tone modeling, word prosody, Mandarin speech recognition.

1. INTRODUCTION

Tone modeling plays an important role in Mandarin speech recognition. Most state-of-the-art Mandarin speech recognition systems adopt embedded tone modeling, where tonal acoustic units are used and F_0-related features are appended to the spectral feature vector [1]. Although embedded tone modeling can significantly improve the recognition performance, it does not exploit the suprasegmental nature of tones. A tone spans a syllable which has a variable length instead of a fixed window. Therefore, explicit suprasegmental tone models can be used to post-process the first pass recognition output to further improve the recognition performance.

From the oracle experiments in our earlier work [2], we found that by rescoring the first pass recognition output lattices of the embedded tone modeling with perfect tone information, around 30% relative improvement could be achieved. Using a neural network, a simple syllable-level 4-tone model improves the recognition performance by 4% relative in a Mandarin broadcast news (BN) experiment. Inspired by the word duration modeling approach [3, 4] and other word-level prosody modeling techniques [5], in this work we propose to generalize the syllable-level tone modeling to a word-level tone modeling framework with a hierarchical backoff: word-level tone models (word prosody models) are trained for the frequent words, and tonal syllable (TS) or plain tone models are used as backoff for the infrequent and unseen words. In addition, context-dependent tone models can be used as backoff. These prosody models represent both duration and F_0 characteristics of a word. The word prosody models and the backoff tone models can then be used in word lattice rescoring as a complementary knowledge source. There are several advantages of the proposed approach. First, the tone coarticulation within the word is more explicitly modeled. Second, the different backoff strategies offer the flexibility to model the dependencies between the tone and different linguistic units. Finally, the word prosody models are less susceptible to tone labeling errors in the pronunciation dictionary as long as the errors are consistent between the training and decoding dictionaries.

The rest of the paper is organized as follows: In Section 2, we introduce the word prosody models and the modified decoding criteria. In Section 3, different backoff strategies for infrequent and unseen words are described. In Section 4, experiments are carried out and the recognition results are discussed. Finally, we summarize the key points and propose future work in Section 5.

2. WORD PROSODY MODELS

In a Chinese sentence, there are no word delimiters such as blanks between the words. Longest-first match or maximum likelihood based methods can be used to do word segmentation [1]. A segmented Chinese word is typically a commonly used combination of one or multiple characters. As illustrated in Figure 1, for a word $w_i = c_{i1}c_{i2}\cdots c_{iM}$ which consists of M characters, we denote the corresponding tonal syllable sequence as $s_{i1}s_{i2}\cdots s_{iM}$ and the tone sequence as $t_{i1}t_{i2}\cdots t_{iM}$.[1] In this study, we focus on the tone-related prosodic features. In all our experiments, the feature f_{ij} for

[1] Each Chinese character has a pronunciation of a tonal syllable.

1-4244-0728-1/07/$20.00 ©2007 IEEE

each character c_{ij} is a 4-dimensional vector: the syllable duration plus 3 F_0 values sampled from the syllable F_0 contour. The feature f_i for word w_i is obtained by concatenating the feature vectors of all the M characters within the word: $f_i = [f_{i1}; f_{i2}; \ldots; f_{iM}]$.

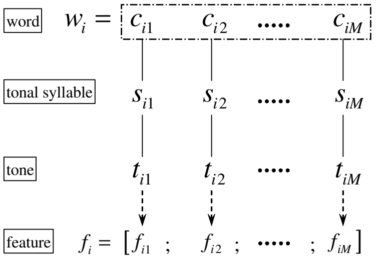
word
$w_i = c_{i1} \quad c_{i2} \quad \cdots\cdots \quad c_{iM}$
tonal syllable
$s_{i1} \quad s_{i2} \quad \cdots\cdots \quad s_{iM}$
tone
$t_{i1} \quad t_{i2} \quad \cdots\cdots \quad t_{iM}$
feature
$f_i = [f_{i1} \; ; \; f_{i2} \; ; \; \cdots\cdots \; ; \; f_{iM}]$

Fig. 1. Backoff hierarchy of Mandarin tone modeling.

By including the tone-related prosodic features, the standard equation of maximum *a posteriori* probability (MAP) decoding can be modified as

$$W^* = \underset{W}{\operatorname{argmax}}\, P(W|O_A, F) \tag{1}$$

$$= \underset{W}{\operatorname{argmax}}\, P(O_A, F|W)P(W) \tag{2}$$

$$= \underset{W}{\operatorname{argmax}}\, P(O_A|W)P(F|W)P(W) \tag{3}$$

where the word sequence $W = \{w_1, w_2, \ldots, w_N\}$ is composed of N lexical words, O_A are the acoustic features, and $F = \{f_1, f_2, \ldots, f_N\}$ are the prosodic features for the word sequence. Equation 3 relies on the approximation that the acoustic features O_A and prosodic features F are conditionally independent given the word sequence.

Assuming the prosody feature f_i only depends on its corresponding word w_i, then the prosody model can be written as

$$P(F|W) = \prod_{i=1}^{N} P(f_i|w_i) \tag{4}$$

where $P(f_i|w_i)$ is the prosody likelihood of word w_i. In our experiments, we have used Gaussian mixture models (GMMs), where the number of Gaussians depends on the available training data for each model.

As with the traditional syllable-level tone models, the word prosody models can be used to rescore the recognition hypotheses in an N-best list or a word lattice. We choose to rescore lattices since a lattice is a much richer representation of the entire search space.

3. BACKOFF STRATEGIES

With a whole-word prosody model, the F_0 contour and duration of the syllables within the word are explicitly modeled. For unseen words or infrequent words that appear less than a certain amount of times in the training data, we use the product of syllable-level models. The particular syllable model is chosen according to a hierarchical backoff illustrated in Figure 1. Within this framework, there are several different backoff strategies that we can take. We investigate both context-independent (CI) and context-dependent (CD) tone models as backoff alternatives.

3.1. Context-independent tone models

To compute the prosody likelihood $P(f_i|w_i)$ of the word w_i with context-independent component models, we use:

$$P(f_i|w_i) \underset{C(w_i)<C_t}{\Longrightarrow} \prod_{j=1}^{M} P(f_{ij}|s_{ij}) \tag{5}$$

where "⇒" denotes backoff, $C(w_i)$ denotes the frequency of the word w_i in the training corpus and C_t is the frequency count threshold. Depending on the amount of training data for the particular TS s_{ij}, the actual tone model used may be TS dependent or simply tone dependent. The backoff strategy in this case is

$$P(f_{ij}|s_{ij}) \underset{C(s_{ij})<C_t}{\Longrightarrow} P(f_{ij}|t_{ij}). \tag{6}$$

When the frequency count of a tonal syllable is larger than the count threshold, an explicit TS-dependent tone model is trained. Otherwise, the likelihood computation is backed off to tone models. For simplicity, we have used the same count threshold for training all tone models including word and CI or CD tonal syllable models.

Similar to the word prosody models, these syllable-level models are trained as GMMs except with fixed 4-dimensional features.

3.2. Context-dependent tone models

More generally, the word prosody models could be backed off to CD syllable-level models such as tone-context-dependent TS models, bitones or tritones. As we have found in [2], the carry-over coarticulation effect from the left context is much more significant than from the right context. Therefore, we have used left-tone context-dependent tone models as follows:

$$P(f_i|w_i) \underset{C(w_i)<C_t}{\Longrightarrow} \prod_{j=1}^{M} P(f_{ij}|t_{i(j-1)}, s_{ij}). \tag{7}$$

Again, depending on the amount of training data for the particular CD models, a backoff model may be used, where here we follow the strategy

$$P(f_{ij}|t_{i(j-1)}, s_{ij}) \underset{C(t_{i(j-1)},s_{ij})<C_t}{\Longrightarrow} P(f_{ij}|t_{i(j-1)}, t_{ij}) \tag{8}$$

$$\underset{C(t_{i(j-1)},t_{ij})<C_t}{\Longrightarrow} P(f_{ij}|t_{ij}) \tag{9}$$

For a reasonably large training corpus, there are enough samples for training all possible bitone models. Therefore, the backoff from left bitone to tone models is usually not used.

For the special case of the first tonal syllable of the word, it is often not straightforward to find the unique left tone context of a word arc in the lattice. We can either use its CI backoff models or expand the lattices according to the crossword left tone context. In our experiments, the former approach has been taken.

4. EXPERIMENTS AND RESULTS

4.1. Baseline system

Training and Test Data: The baseline system in this paper is a Mandarin BN system with embedded tone modeling, as described in [2]. The acoustic models are 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. For testing, we use the CTV and NTDTV shows of NIST RT-04 evaluation set (`eval04`) collected in April 2004. Each show contains around 20 minutes of speech data. The RFA data is not used here because it has a very significant mismatch with the training data.

Features and Models: Standard 39-dim MFCC features with vocal tract length normalization are generated with the frontend 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 [6] to alleviate pitch halving and doubling problems. Then the F_0 smoothing and normalization algorithm described in [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 used in the embedded tone modeling. 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 the neutral tone 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 and 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. Word lattices are generated with the speaker adapted models. The adapted results are evaluated in terms of character error rate (CER).

4.2. Training of prosody models

Forced alignment is performed to align all the training data to tonal syllables. The F_0 features are generated similar to those used in HMM modeling, but without the final step of smoothing since we find it is better for the explicit tone modeling without the low-pass filtering. Based on the forced alignment and the processed F_0 features, the feature vectors for word prosody models and other syllable-level tone models are extracted. The syllable features f_{ij} are mean- and variance-normalized per speaker. GMMs with diagonal Gaussians are trained for all the models that have a frequency count more than the threshold.

4.3. Decoding with prosody models

The prosody models are used to rescore the word lattices from baseline system. For each word arc in the lattice, the new score is computed based on acoustic model (AM), language model (LM) and prosody model (PM) scores,

$$\psi(w_i) = \psi_{AM}(w_i) + \alpha\psi_{LM}(w_i) + \beta\psi_{PM}(w_i), \qquad (10)$$

where α is the language model weight, β is the prosody model weight, and the prosody score $\psi_{PM}(w_i)$ is given by

$$\psi_{PM}(w_i) = \frac{1}{M}\sum_{j=1}^{M} d_{ij} \log P(f_i|w_i), \qquad (11)$$

where d_{ij} is the duration of the j-th character in word w_i. The average syllable duration is used to weight the prosody likelihood, since in practice we find it effective for balancing insertion and deletion errors. To more explicitly control the deletion errors, we can introduce an additive constant proportional to the number of characters in the word, similar to that used in duration rescoring [4]. The weights α and β are determined by a grid search.

As in training, the feature vector f_i for word arc w_i is extracted from the F_0 features and the time marks in the lattice. However, the speaker-based normalization is not as straightforward as in training, since no oracle transcription is available for getting the syllables and their boundary time marks. There are two options: the first is to use a global mean and variance normalization factor from the training data; the second way is to use the top hypothesis to compute the speaker (cluster) mean and variance normalization factors. For simplicity, in this study we have used global normalization factors in decoding but speaker-based normalization in training.

4.4. Results and discussion

Since both the word-level prosody models and different syllable-level tone models have been trained, we have the flexibility to choose different models and backoff strategies during lattice decoding.

Table 1 shows the decoding results of different models and backoff strategies.[2] The plain tone models can improve the performance slightly, while the word prosody models with backoff provide a much larger improvement. We also find that TS-dependent tone modeling is not significantly different from tone modeling, neither in rescoring directly nor as backoff models.

Table 1. *CER(%) using word prosody models with CI tone models as backoff.*

Model	CTV	NTDTV
Baseline	11.7	19.2
tone	11.5	18.8
TS ⇒ tone	11.5	18.8
word ⇒ tone	11.2	18.2
word ⇒ TS ⇒ tone	11.1	18.4

Table 2 shows the decoding results with CD tone models as backoff. Comparing the left bitone results in Table 2 to the CI tone results in Table 1, we can see that left bitone models are more effective than CI tone models, due to the better modeling of tone coarticulation. However, the results between CI backoff and CD backoff are not significantly different, probably because much of the tone coarticulation has been modeled by the word prosody models. In Table 2, the left-context-dependent TS models perform worse than the bitone models. This might be explained by a lack of dependency between tones and base syllables, or the backoff may not have been properly tuned. With the 3-level CD backoff modeling, performance can be improved by 0.7% absolute on the CTV show and 1.0% absolute on the NTDTV show.

Table 2. *CER using word prosody models with CD tone models as backoff. "l-" denotes left-tone context-dependent models.*

Model	CTV	NTDTV
Baseline	11.7	19.2
l-tone	11.3	18.4
l-TS ⇒ *l*-tone	11.4	18.7
word ⇒ *l*-tone	11.2	18.2
word ⇒ *l*-TS ⇒ *l*-tone	11.0	18.2

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a hierarchical tone modeling framework for lattice rescoring in Mandarin speech recognition. Both word-level and syllable-level tone models are trained and used to rescore word lattices. For infrequent or unseen words, syllable-level tone models are used as backoff. This hierarchical tone modeling framework can be viewed as a generalization of the traditional syllable-level tone models. Experimental results show that word-level tone modeling outperforms syllable-level tone models in a Mandarin BN task.

Preliminary experiments with the RFA data show that acoustic match is needed for effective tone modeling. Future work involves training the prosody models with several hundred hours of speech data, which should alleviate mismatches. We will also investigate the effectiveness of hierarchical prosody modeling in a more complicated Mandarin speech recognition system. In addition, other prosody features such as energy contour may be incorporated to improve the performance.

6. ACKNOWLEDGEMENTS

The authors would like to thank M.Y. Hwang from SSLI lab for help in building the baseline system. Thanks also to M. Siu from Hong Kong University of Science and Technology and T. Lee from Chinese University of Hong Kong 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] M.Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Interspeech*, 2006, pp. 1233–1236.

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

[3] V.R.R. Gadde, "Modeling word durations," in *Proc. ICSLP*, 2000, vol. 1, pp. 601–604.

[4] N. Jennequin and J.-L. Gauvain, "Lattice rescoring experiments with duration models," in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006, pp. 155–158.

[5] D. Vergyri, A. Stolcke, V.R.R. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition," in *Proc. ICASSP*, 2003, vol. 1, pp. 208–211.

[6] M.K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," in *Proc. Eurospeech*, 1997, vol. 3, pp. 1391–1394.

[2] The baseline results are slightly different from our previously published results since a cleaner and more consistent decoding lexicon has been used.