MAIN VOWEL DOMAIN TONE MODELING WITH LEXICAL AND PROSODIC ANALYSIS FOR MANDARIN ASR

Shilei Zhang[1], *Qin Shi*[1], *Stephen M. Chu*[2], *and Yong Qin*[1]

[1]IBM China Research Lab, Beijing 100193, China
{slzhang, shiqin, qinyong}@cn.ibm.com

[2]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
schu@us.ibm.com

ABSTRACT

The tone is a distinctive discriminative feature in Mandarin Chinese. Often functional, yet seldom thorough are most large-scale Mandarin speech recognition systems in treating tone modeling. In particular, many lack the necessary sophistication to deal with the myriad variations arising from the combination of acoustic and lexical contexts. This paper reports an attempt to account for these variabilities and to bring richer tone modeling into the IBM Mandarin broadcast transcription system. In particular, we describe a system that combines the embedded approach and a novel explicit tone modeling technique characterized by *a.* robust tone tracking in the main-vowel domain, and *b.* context-dependent models with lexical and prosodic contexts. The proposed method is validated on a connected-digits set and subsequently evaluated on a large-vocabulary broadcast transcription task. It is shown that 14.8% and 5.4% relative reductions in character error rate are achieved respectively.

Index Terms **–** tone models, decision tree, main vowel, tone domain, lattice rescoring.

1. INTRODUCTION

The tone plays an important lexical role in tonal spoken languages like Mandarin Chinese, and naturally, provides distinctive discriminative information in automatic speech recognition (ASR) systems developed for these languages. A variety of approaches to incorporating tonal information [1-6] in Mandarin ASR have been investigated. Arguably the most prevalent in state-of-the-art large vocabulary systems is *embedded tone modeling*, where the fundamental frequency, F0, and/or its derivatives are joined to conventional spectral features, and the tonal identity is typically a part of the acoustic units. Although the straightforward embedded approach can noticeably improve the recognition performance, it does not exploit the supra-segmental nature of tones.

To overcome this limitation, explicit supra-segmental tone models can be used to post-process the recognition hypotheses given by an embedded system to further improve recognition performance. Evidently, this hybrid approach allows us to take advantage of methods for detecting and modeling tonal cues developed for non-ASR applications, such as text-to-speech (TTS), for tone modeling in recognition.

Generally, Mandarin tone modeling and recognition is difficult due to variations in the acoustic, prosodic, and lexical levels, e.g., the overall pitch drift within an utterance, the presence of phrase boundaries, tone co-articulation, and tone *sandhi* [8]. This work aims to address these problems by looking into a robust tonal parameter estimation technique, a decision-tree based tone modeling method that takes contextual information into account, and ways to best bring tonal cues into recognition.

First, parameterized tone models using polynomial coefficients are considered in the work. To further make the parameterization robust to pitch estimation and boundary errors, we propose the use of main vowel phone units as the domain of tone in estimating pitch contour. Second, an effective decision-tree based method is used to construct tone models with higher level and longer range features. In particular, we describe a novel method of building explicit tone models that exploit both prosodic and lexical features, as well as the syllabic contextual information. In this case, the decision tree serves as a probability estimator based on score normalization before it can be combined with the acoustic likelihood. Finally, explicit tone models are applied in lattice rescoring to improve speech recognition performance.

The rest of the paper is organized as follows. In Section 2, we discuss in detail the proposed tone modeling paradigm; Section 3 describes the lattice rescoring step; Experimental results are presented in Section 4, followed by conclusions in Section 5.

2. CONTEXTUAL TONAL MODELING IN THE MAIN VOWEL DOMAIN

2.1. Contextual Tonal Variation in Mandarin

There are four tones in Mandarin Chinese, each defined by a canonical F0 contour pattern [8]. The F0 contour of a syllable spoken in isolation generally corresponds well with the canonical pattern of its tone, although there exists variability due to vowel intrinsic pitch, perturbation by the initial consonant, and the pitch range of a speaker, as well as other individual differences. In addition to these variations, tone in continuous speech undergo both phonological and phonetic modifications due to tone sandhi and tone co-articulation, which can cause the F0 contours to significantly deviate from canonical forms. Tones can also be influenced by many other linguistic and paralinguistic commands, such as phrase grouping, intonation, etc. Due to these constraints, the exact acoustic realizations of tones are determined not only by the properties themselves, but also by their contexts.

To address these problems, syllable based prosodic features, including F0 contour features, average log-energy and duration, are used in this study. Other possible factors contributing to tone

pattern variations are taken into consideration. The instantiation of a lexical tone in continuous speech depends on the neighboring syllables. Moreover, because of co-articulation effect, the contour shape of a syllable may be affected by the F0 contour patterns of neighboring syllables. Therefore, the neighboring tone types, the location of the syllable within a phrase and utterance, average pitch value of syllable and utterance, and speech rate should also be considered as necessary features in the modeling process.

2.2 Tonal Feature

2.2.1. Outlined features

Tonal features can be classified as *detailed features* which use the entire F0 curve, and *outlined features* which capture the main structure of the F0 curve. In this paper, detailed pitch features are first extracted using the YIN [10] algorithm. The pitch values in the unvoiced segments are replaced through piecewise cubic Hermite interpolating polynomial using samples in the adjacent voiced regions. To reduce the number of parameters and improve robustness, outlined features are then extracted from the output from the previous step. The outline F0 features indicate an averaged pitch trajectory during the carrier of the tone. For parameterization, a second-order polynomial $f_t = a + bt + ct^2$ is employed to fit the continuous F0 contour on tone domain using the least square criteria, and the fitting coefficients $\{a, b, c\}$ are treated as the F0 contour pitch features for the each syllable.

2.2.2. Tone Domain

An important issue for tone modeling is on which part of the F0 contour of a syllable carries tone information. Tone is carried by perceivable pitch in the voiced part of a syllable, and no pitch is perceived in the unvoiced region. As a result, the syllable is not the appropriate domain since the whole F0 trajectory is discontinuous at the junctions between neighboring voiced and unvoiced segments. From a modeling point of view, it seems more advantageous to extract tone features from the F0 contour of the syllable rhyme, because the large perturbation at the early portion of a syllable is less relevant to the current tone, and is likely to introduce noise [8]. However, the exact domain of rhyme is dependent on tonal phonetic units of speech recognition engine.

The Pinyin system maps Mandarin phonetics using a set of initials and finals. By grouping the glides with the consonant initials into premes, each syllable can be decomposed into demisyllables. Note that the number of phonemes can be drastically reduced if we assume that the pitch information on the main vowel is sufficient to determine the tone of a whole syllable, which is verified in [9]. In this paper, we focus on the use of main vowel based pitch contour extraction in which tone information is carried by the F0 contour in the main vowel effectively. Because main vowel is bounded by initial consonants or glides and codes on both sides, this part of the F0 contour pattern is stable and can minimize the impact from contextual tonal variation.

Using the main-vowel method, the total number of phonemes can be further reduced compared to the demisyllable structure. In fact, in Mandarin, there are 9 main vowels. For example, the Chinese word for "rescue" is made up of two characters: 救援. The tonal pinyin can be written as "jiu4/yuan2", with the tone represented by a number. The demisyllable structure is described as "JI ***OU4***/YU ***AN2***", while the main vowel structure is represented as

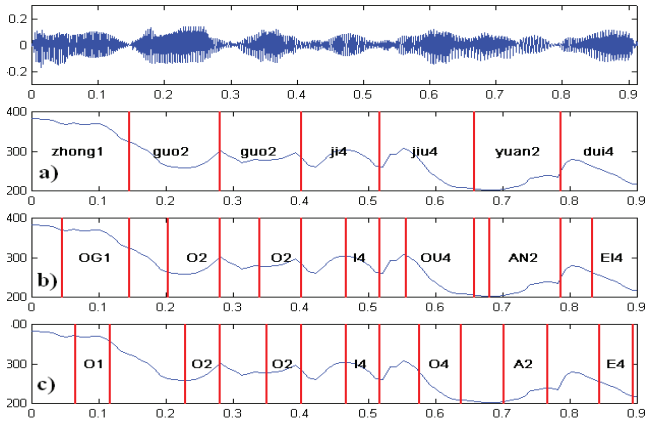

Fig. 1 The F0 contour under three possible domains for tone modeling: *a*. syllable, *b*. demisyllable, and *c*. main vowel.

"J IM ***O4*** UT/Y YUM ***A2*** NT". Fig. 1 illustrates the F0 contour of different level modeling units with the phrase 中国国际救援队 (China International Search and Rescue Team) as an example. We can see that the main vowels have a stable F0 contour pattern, while the F0 contour of the syllable and the demisyllable depends largely on surrounding acoustic conditions.

2.3. Contextual Tone Models

2.3.1. Feature extraction

As mentioned above, to cope with contextual variation, the predictor and target features used in decision tree based tonal models are expanded to include contextual features as shown in Table 1.

Table 1. Features used for classification and tone modeling

1	*duration and log-energy of the syllable*
2	*curve fitting parameters of the F0 contour of syllable*
3	*2-equal length subsection pitch slopes*
4	*tone types of neighboring syllables*
5	*log-pitch mean of syllable and utterance*
6	*current location within utterance and phrase*
7	*speech rate*

Forced alignment is performed to align all the training data to tonal syllables. Phone-level alignments from a speech recognizer provide durations of syllable and various measures of location and speaking rate.

2.3.2. Decision tree construction

The decision tree is widely used in speech applications partly because of its non-parametric nature and its ability to handle heterogeneous features. It is composed of a set of leaves nodes and a set of non-terminal nodes, each of which consists of a binary question on the features so as to partition the data into two subtrees. If a tree is used as a classifier, for example, for classifying lexical tones, each terminal node is labeled by the dominant class. If a tree is used to estimate class probability, each terminal node represents a particular class distribution.

Our tone modeling decision tree was trained using the *classification and regression tree* (CART) [12] utilizing the gain criterion that minimizes class entropy. Here, the CART, which can

make use of both continuous and discrete features, is built as a classifier to predict the tone type of a syllable. The input features of CART include all the features mentioned in Table 1 except for the curve fitting tonal feature, which will be used to estimate pitch contours probability distribution. The following list highlights the recursive steps of the algorithm in constructing CART tree:

1: at the root node, perform all possible splits on each of the predictor variables, and apply a predefined node measure to determine the reduction in entropy to each split.
2: select the best possible variable to split the node into two child nodes by applying the splitting criteria.
3: repeat steps 1) and 2) for each of the non terminal nodes and produces the largest possible tree.
4: apply pruning algorithm to the largest tree and produces a sequence of sub trees of different sizes from which an optimal tree is selected using k-fold cross validation.

Then all training data can be assigned to terminal nodes in terms of predictor variables.

According to the tonal labels, we partitioned training samples into 5 tonal classes (including the neutral tone) within each terminal node, which are modeled as the prior probability of tone type and unimodal Gaussian density respectively, represented by the mean vector and the covariance matrix of curve fitting tonal features. The mean of a tonal class is obtained by calculating the average of all the vectors belonging to that particular tone. The covariance matrix is obtained by computing the diagonal covariance matrix of the assigned vectors. And the prior probability can be found by calculating the proportion of the number of vectors belonging to a particular tone within the terminal node. Thus, we can get context-dependent tone models by constructing the decision tree.

2.3.3. Likelihood score normalization

In this study, the decision tree serves as a probability estimator. Therefore, it is necessary to convert the decision tree output into a normalized likelihood score before it can be combined with the acoustic likelihood in HMM.

Define f to a 3-demensional curve fitting vector for syllable s with tonal label λ belonging to terminal node j; $\alpha_{ij} \geq 0$ are the prior probabilities, with $\sum_{i=1}^{5} \alpha_{ij} = 1$, and $b_{ij}(u)$ are 3-variate Gaussian densities with mean vector μ_{ij} and covariance matrix Σ_{ij}, which corresponds to tone model of each tone type i within terminal node j of the decision tree.

In order to minimize the syllable dependent variations, the likelihood normalization in log domain using background tone models is computed as:

$$p(f \mid s) = \log(a_{\lambda j} b_{\lambda j}(f)) - \log(1/4 \sum_{\bar{\lambda}} \alpha_{ij} . b_{ij}(f)) \qquad (1)$$

where $\bar{\lambda}$ is the sum of probability densities of other candidate tone types except component λ.

3. LATTICE RESCORING USING TONE MODELS

As mentioned, given the values of predictor variables of a syllable, the tone type is estimated by a mapping from the predictor variables based on the decision tree. And the normalized likelihood score can be computed to rescore word lattices generated by the recognizer, which give a rich representation of the entire search space. The character acoustic likelihood and boundary information are recorded during the first-pass Viterbi search. Then based on the phone segmentation given by the hypotheses and the processed F0 features, the vectors for syllable-level tone models are extracted to evaluate the tone likelihood.

A Chinese word is merely a commonly used sequence of one or multiple characters. For a word $W_i = \{s_{i1}, s_{i2}, \cdots s_{iN}\}$ which consists of N syllables (characters), we denote the corresponding tone sequence as $\lambda_{i1}, \lambda_{i2}, \cdots \lambda_{iN}$. In our experiments, the feature f_{ij} for each syllable is a 3-dimensional vector – the three curve fitting parameters sampled form F0 contour of main vowel. The tone normalized likelihood $P(f_i \mid W_i)$ of the word W_i with context-dependent tone model is,

$$P(f_i \mid W_i) = \frac{1}{N} \sum_{j=1}^{N} \log p(f_{ij} \mid s_{ij}) . \qquad (2)$$

To avoid bias toward longer words, we normalize the word tone score by the number of syllables in each word.

Integrating the tome likelihood into the word acoustic likelihood and language probability, we have

$$\varphi_t(i) = \psi_t(i) + \alpha \cdot P(f_i \mid W_i) , \qquad (3)$$

where $\psi_t(i)$ is the sum of acoustic score and the language score for the word W_i; and α is the weight of tone model optimized in a development set. It is interesting to note that the tone likelihood can be interpreted as a form of word tonal penalty. The hypotheses within lattices are then resorted according to the adjusted total score to give a new "best" hypothesis.

4. EXPERIMENTAL RESULTS

4.1. Connected Digits Experiments

Free from the language model constraints, connected digits provide an excellent platform to validate the proposed tone modeling methods. Digit strings have relatively simple sentence level prosodic structures, making preliminary assessment of the character-level prosodic effects easier. Also, digit strings are usually spoken in phrase groups; thus we can study the dependency of tone expression on the phrase structure. Moreover, digits cover all four lexical tones in Mandarin, and therefore can provide an adequate space to study the contextual effects of tone.

The *spoken connected digit database* was collected in cars under different speed conditions, e.g., parking, medium speed, high speed. The training set consists of 61 hours of digit data from 1,189 speakers. The test set contains 576 utterances with varying lengths ranging from 2 to 5 digits.

For modeling units, there is essentially no difference between the demisyllable and the main vowel approach for digits. Therefore, we only consider the latter here. After the decision tree is constructed with the training data, a total of 58 tone content-dependent patterns are obtained.

Table 2. Contextual tone modeling on the main vowel domain gives clear improvement in both sentence error rate (SER) and character error rate (CER) on the in-car connected digit corpus.

System	SER	CER
baseline	11.3%	2.7%
+tone model	9.2%	2.3%

Table 2 shows that the baseline system with *maximum likelihood* (ML) trained context-dependent acoustic models gets a 2.7% character error rate (CER), and 11.3% sentence error rate (SER). We found that the lattice generated by the baseline system has indeed great potential for the post processing step. In fact, correct hypothesis exists in 87.7% of the lattices of the incorrectly recognized segments. After applying the tone models to rescore the lattices, the CER and SER are reduced to 2.3% and 9.2%, respectively, representing relative error reduction of 19% and 15%. Closer inspection of the results reveals that a large portion of the improvement from tone modeling on digits comes from the reduction of insertions and deletion errors. For instance, the segmentation of vowel-vowel sequence such as "55", pronounced as "wu2 wu3" is extremely difficult without tone due to frequent absence of glottal stops in continuous speech. However, the contextual tone models here are indeed often able to choose the correct segmentation in the lattice by identifying the right tones and taking advantage of the apparent sandhi.

4.2. Broadcast Speech Experiments

The second set of experiments are carried out on a large-vocabulary broadcast transcription task. The acoustic training set consists of 100 hours of audio data released by LDC for the DARPA GALE program. The baseline speaker independent acoustic model has 5K quinphone states modeled by 100K Gaussian densities. The evaluation set is from the 2007 GALE evaluation, referred to as *eval'07* here. It contains 2 hours and 21 minutes of audio, and 40.6K characters in the reference transcript. Results on *eval'07* are further divided into broadcast news (*bn*) and broadcast conversations (*bc*). For contextual tone model training, forced-alignment is performed against the references to get the syllable segmentations information. The decision trees generated from the demisyllable and the main vowel domains have 381 and 395 tone modes, respectively.

Table 3. Character error rates for *eval'07*. Contextual tone modeling on the main vowel gives the best performance.

System	*eval'07*	*bn*	*bc*
baseline	22.4	12.8	34.3
embedded tonal model	21.6	12.1	33.4
+contextual tone (demisyllable)	21.9	12.3	33.9
+contextual tone (main vowel)	21.2	11.9	32.8

For the acoustic models, we compared a system built on the main vowel phone set and one using the demisyllable set, and found that the latter achieves better baseline performance and gives lattices with more correct word candidates. Therefore, the baseline system employs demisyllable phonetic units by a forward Viterbi algorithm to generate a word lattice, which is then forced-aligned for each processed word to get the main vowel based alignment information. On *Eval'07*, the experimental results show a CER of 22.4% without tone processing as in the baseline system. When the tonal cues with demisyllable alignment and main vowel alignment information are incorporated into the post-processing stage, 0.5% and 1.2% absolute reduction are achieved respectively, as shown in Table 3. The results show that main vowel based tone model can achieve higher reduction in CER than that of the demisyllable. It is also confirmed by the experiment that the proposed contextual tone models can further improve recognition performance on top of embedded tonal modeling in large vocabulary continuous speech recognition tasks.

5. CONCLUSIONS

This work aims to bring richer tone modeling, in the form of better treatment for lexical and prosodic contexts, into a modern Mandarin ASR system. Experiments show that the CART decision-tree based technique is effective in capturing the contextual variations in tones. It is further shown that tone tracking on the main-vowel domain is indeed more stable than the demisyllable domain, and gives better performance in explicit tone models.

REFERENCES

[1] Y. Tian, J. l. Zhou, M. Chu and E. Chang, "Tone recognition with fractionized models and outlined features," in *ICASSP'04*, pp. 105-108, 2004.

[2] X. Lei and M. Ostendorf, "Word-level tone modeling for mandarin speech recognition," in *ICASSP'07*, pp. 665-668, 2007.

[3] C. Wang and S. Seneff, "A study of tones and tempo in continuous mandarin digit strings and their application in telephone quality speech recognition," in *ICSLP'98*, 1998.

[4] Y. B. Zhang, M. Chu, C. Huang, and M. G. Liang, "Detecting tone errors in continuous Mandarin speech," in *ICASSP'08*, pp. 5065-5068, 2008.

[5] H. Huang and J. Zhu, "Discriminative incorporation of explicitly trained tone models into lattice based rescoring for Mandarin speech recognition," in *ICASSP'08*, pp. 1541-1544, 2008.

[6] P. F. Wong and M. H. Siu, "Decision tree based tone modeling for Chinese speech recognition," in *ICASSP'04*, pp. 905-908, 2004.

[7] H. Gish and K. Ng, "Parametric trajectory models for speech recognition," in *ICSLP'96*, pp.466–469, 1996.

[8] C. Wang, "Prosodic Modeling for Improved Speech Recognition and Understanding," Doctoral dissertation, MIT, Cambridge, MA, 2000.

[9] C. J. Chen, H. Li, L. Shen, and G. K. Fu, "Recognize tone languages using pitch information on the main vowel of each syllable," in *Proc. ICASSP'01*, vol. 1, pp.61–64, 2001.

[10] A. de Cheveigné and H. Kawahara, "YIN.A fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, 111(4): 1917-1930, 2002.

[11] S. M. Chu, et al., "Recent Advances in the IBM GALE Mandarin Transcription System," *in Proc. ICASSP'08*, pp.4329-4332, 2008.

[12] Y. Yohannes and P. Webb, "Classification and Regression Trees, CART™: A User Manual for identifying indicators of vulnerability to famine and chronic food insecurity," *Microcomputers in Policy Research Series 3*, Washington, IFPRI, 1999.