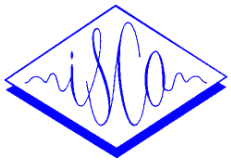

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

INTEGRATION OF TONE RELATED FEATURES FOR MANDARIN SPEECH RECOGNITION BY A ONE-PASS SEARCH ALGORITHM

Pui-Fung WONG, Man-Hung SIU

Dept. of EEE, Hong Kong University of Science and Technology
Clear WaterBay, Kowloon, Hong Kong
eefung@ust.hk, eemsiu@ust.hk

ABSTRACT

How to model Chinese tones and integrate them into an HMM-based recognizer for Chinese recognition has long been an area of interest to researchers. In this paper, we propose the use of a polynomial trajectory model to represent pitch shape. We further propose an efficient one-pass search approach that integrates the tone likelihood into the Viterbi search procedure. We report a number of experimental results on tone classification and tonal syllable recognition in the 863 corpus. While the improvement in the tonal syllable accuracy is small, it nevertheless shows the feasibility of the proposed approaches.

1. INTRODUCTION

It is well-known that Chinese is a tonal language and the tone identity is needed for correct recognition of the Chinese words. Chinese words are composed of one or multiple mono-syllable units called characters and each character is typically associated with one tone. Automatic recognition of Chinese speech requires the accurate recognition of tones. In this paper, we use the term ”tonal syllables” to mean the Chinese syllables with particular tones associated and ”base syllables” to mean syllables without regard to their tones.

Chinese tones are commonly characterized by the pitch contour of the syllable and thus are supra-segmental in nature. Because of this, recognition of Chinese tones using an HMM-based recognizer, which assumes conditional frame independence, can be difficult. Furthermore, Mel-filtered cepstral coefficients (MFCC), which are the most widely used features for speech recognition, explicitly does not focus on the pitch frequency information. Different approaches have been proposed for Chinese speech recognition. The simplest one creates tone-dependent phonetic units, such as tonal syllables [1]. While this does not explicitly use any pitch-related features, apparently the MFCCs still capture enough pitch information for this approach to give reasonable performance. Another approach [2, 3] is to augment the MFCC with pitch-related features such as F0. This has shown to perform better than the using MFCCs alone [5] but it does not capture the supra-segmental nature of tone. It also introduces the problem of representing the F0 features for the unvoiced frames. To fully capture the supra-segmental nature of tone, another approach separates the base syllable recognition and tone recognition into two sequential stages [4]. In the first stage, a HMM-based syllable recognizer searches for N best possible base syllable sequences. In the second stage, tones are identified for the base syllable sequences found in the first stage. One limitation of this approach is that the tone information is not included in the search process.

In this paper, we propose to model pitch contour using a polynomial trajectory model [9] for Mandarin tones and then propose a 1-pass search algorithm that searches for the best tonal syllables in an HMM-based recognition framework. The basic idea is to compute the tone likelihood for each syllable during decoding instead of post-processing. Because tone is a supra-segmental feature, its likelihood can only be computed if we know the syllable segmentation. So, instead of compute the tone likelihood per frame, tone likelihood is computed at all the syllable ends. Before we describe our tone modeling approach, we briefly describe some characteristics of the Chinese syllable structure in the next section. In Section 3, we then describe the tone modeling including the pitch extraction algorithm, the pitch contour and the tone model used. In Section 4, we describe the integration of the tone likelihood into the Viterbi search algorithm. We then describe our experiments in Section 5 and conclude the paper in Section 6.

2. MANDARIN SYLLABLE STRUCTURE

Different from the English language, the basic unit in Chinese is characters. Words in Chinese are made up of a sequence of one or more characters. Each character is monosyllable. In phonetics, it is common to break up a syllable into the onset and rhyme [7]. Similarly, in Chinese recognition, the Chinese syllable onsets are called ”initial” and the syllable rhymes are called the ”final”. Similar to an En-

glish syllable, a Chinese syllable may not have an onset and this is sometimes represented as an "null initial". The Chinese final may or may not have a coda (ending consonant). In Mandarin, there are 408 different base syllables. Different from an English syllable, each syllable can have up to 4 lexical tones and 1 neutral tone. Because pitch is only defined for the voiced regions, it is commonly assumed that the Chinese tone is associated only with the syllable finals. Each tone syllable is characterized by its particular pattern of pitch contour or trajectory. For example, tone 1 has a flat contour while tone 2 is a rising pitch. The simplified of pitch contour for tone 1 to tone 4 are shown in Figure 1.

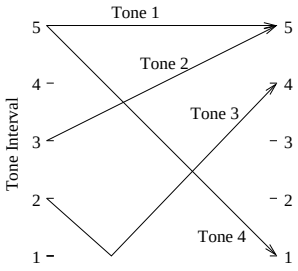

Fig. 1. The four standard tone patterns in Mandarin Chinese.

Combining tones with the base syllables, there are a total 1302 different tonal syllables (not all base-syllable and tone pairs occurs in Mandarin).

3. TONE MODEL

Our tone modeling process includes three major components: 1) Pitch extraction, 2) pitch contour fitting and 3) tone likelihood computation.

The pitch extraction is performed on a frame-by-frame basis using the cepstrum method [8]. To eliminate the pitch doubling or halving errors, the extracted pitch is smoothed using a non-linear median filter [11]. In our other work using CEP, around 85% of estimated pitch is within 10 Hz from that obtained using the laryngograph data.

3.1. Pitch Contour Fitting

As we have discussed before, the syllable tone is related to the shape of the pitch contour across the syllable. Given the shapes of the five tones shown in Section 2, we proposed to represent the pitch contour by a polynomial. The coefficients of the polynomial become the features for tone modeling. There are several advantage in using a polynomial fitting. First, the coefficients are easy to estimate. Second, polynomial trajectory models [6] have been used in speech recognition for modeling the acoustics. That provides a good theoretical and implementation foundation.

Let F is $(F_1, F_2, .., F_N)^{'}$ be a sequence of F0 for a syllable. The objective is to find the polynomial of order $d-1$ with coefficients β_k that best represent F. Let $\hat{F} = (\hat{F_1}, \hat{F_2}, .., \hat{F_N})^{'}$ be an estimated vector of F. Each estimated $\hat{F_i}$ can be expressed as,

$$\hat{F_i} = \beta_0 + \beta_i t_i + \beta_2 {t_i}^2 + \cdots + \beta_{d-1} t^{d-1} \quad (1)$$

where $t_i = \frac{i}{N}$. t_i is a normalized time scale so that durations are normalized and the fitted model is independent of durations. Equation 1 can be also expressed in the matrix form as below,

$$\begin{bmatrix} \hat{F_1} \\ \hat{F_2} \\ \vdots \\ \hat{F_N} \end{bmatrix} = \begin{bmatrix} 1 & t_1 & {t_1}^2 & \dots & {t_1}^{d-1} \\ 1 & t_2 & {t_2}^2 & \dots & {t_2}^{d-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_N & {t_N}^2 & \dots & {t_N}^{d-1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix} \quad (2)$$

$$\hat{F} = T\vec{\beta}$$

Different optimization criteria can be used to estimate $\vec{\beta}$ such as the square error. In this work, we used the least square error criterion and the set of $\vec{\beta}$ can be estimated by the following equation,

$$\vec{\beta} = (T^{'}T)^{-1} T^{'} F \quad (3)$$

To simply our discussion, we call $\vec{\beta_j}$ the tone feature for syllable j.

3.2. Tone Likelihood

The ultimate objective of the modeling is to estimate the tone likelihood such that the tone information can be combined with the syllabic information in recognition. Let $\vec{\beta_j}$ and τ_j be the tone feature and tone class for syllable j respectively $\tau_j \in \{0, 1, 2, 3, 4\}$. By modeling the tone features as a Gaussian mixture, the tone likelihood of $\vec{\beta_j}$ can be expressed as:

$$p(\vec{\beta_j}|\tau_j = i) = \sum_l w_{i,l} \mathcal{N}(\vec{\beta_j}; \mu_{i,l}, \Sigma_{i,l}),$$

where $\mu_{i,l}$ and $\Sigma_{i,l}$ are the mean and covariance of the l-th mixture for tone i and $w_{i,l}$ is the mixture weight.

4. SPEECH RECOGNITION

In this section, we describe the integration of the tone likelihood into the HMM-base recognition process. Viterbi algorithm is used to obtain the single most likely state sequence through a set of models for an utterance. In our system, this searching process is performed with the "Token Passing Model" [12] in which the path metric $\psi_j(t)$ and the partial path together (called a token) are stored for each state at each time instance. The token is recursively updated across time. The use of tokens eliminates the need to perform state trace-back at the end of the Viterbi search. Furthermore, at each time for each state, state alignment information for

past match is also available. The recursive computation of the path metric $\psi_j(t)$ at time t and state j is given by,

$$\psi_j(t) = max_i\{\psi_j(t-1) + \log(a_{ij})\} + \log(b_j(O_t)), \quad (4)$$

where $b_j(O_t)$ is the observation probability of observation O_t at state j. Suppose null node j marks the end of a acoustic unit k (for example, a syllable or a word), then, the log probability is updated using Equation 5, where a_{ij} is the transition probability from state i to state j and i is the last real node of unit k.

$$\psi_t(j) = \psi_t(i) + \log(a_{ij}) \quad (5)$$

4.1. 1 Pass Search Algorithm

To recognize Chinese syllables or words, the tone likelihood is an extra term that contributes at each syllable end. One major advantage of using token passing approach for recognition is that it is possible to find the best syllable begin time from the partial path alignment for a syllable ending at any time instance. At time t, suppose j is the end node of syllable k which has the tone of t_j. Denote $\mathcal{B}(t,j)$ the best word begin time associate with end node j at time j. The tone feature, denoted as $\beta(\mathcal{B}(t,j),t)$ can be computed using the polynomial regression and the tone model likelihood is given by,

$$p(\beta(\mathcal{B}(t,j),t)|t_j) = \sum_l w_{l,j}\mathcal{N}((\beta(\mathcal{B}(t,j),t); \mu_{l,t_j}, \Sigma_{l,t_j})).$$

This tone likelihood can be integrated within the word end score by adding an extra term in Equation 5 that recursively compute the path metric $\psi_t(j)$. The new update equation is:

$$\psi_t(j) = \psi_t(i) + \log(a_{ij}) + p(\beta(\mathcal{B}(t,j),t)|t_j). \quad (6)$$

It is interesting to note that tone likelihood can be thought of as a sort of the word insertion penalty and can be weighted to tune its effect.

4.2. Computation

The above integration requires that the tone feature be extracted at each time t for all possible duration for all word ends. This can be computationally expensive. However, we notice that for two tonal syllables that end at the same time, begin at the same time and have the same tone, their tone likelihoods are the same. This implies that tone likelihoods are computed for each of the five tones at each time for all possible durations only. Also note that in Equation 3, the $(T^{'}T)^{-1}T^{'}$ only depends on the syllable duration and can be pre-computed. By limiting the maximum syllable duration to a reasonable number and pre-computing $(T^{'}T)^{-1}T^{'}$, this can be implemented much more efficiently.

5. EXPERIMENTS

5.1. Acoustic Units

Commonly, syllable initials and finals are selected as the basic acoustic units in Mandarin speech recognition. In our experiments, our acoustic inventory includes 30 initials (including the null initial) [10] and 147 tonal finals plus silence and short pause. The set of initials and finals is summarized in Table 1. "(2-3)" means that the final is expanded into tonal finals for tones 2 and 3. final. Our recognition system is HMM-based that uses a left-to-right topology without skips. Initials are modeled by 3 states and tonal finals are modeled by 5 states. Each state is modeled by using 8 Gaussian mixtures. On the feature extraction, we use 39 features which consist of 12 MFCC plus frame energy and their first and second derivatives. The training of acoustic model is performed using the HTK [14].

Initials	b, c, ch, d, f, g, h, j, k,l, m, n, p, q, r, s, sh, t,x, z, zh, l, m, n, r, _a, _e, _I, _o, _u
Tonal Finals	ia(0-4),ai(0-4),an(0-4),ang(1-4),o(0-4),ar(4), e(0-4),ei(1-4),en(0-4),eng(1-4),er(2-4),i(0-4), iong(1-3),ia(1-4),ian(1-4),iao(0-4),in(0-4), ing(0-4),iu(1-4),o(0-4),ong(0-4),ou(0-4),u(0-4), ua(1-4),uai(1-4),uan(1-4),uang(1-4),ue(1-4), ui(1-4),un(0-4),uo(0-4),v(2-4)

Table 1. Initial and tonal final units for acoustic modeling

5.2. Database

All experiments are performed on the Chinese 1998 National Performance Assessment (Project 863) [13] corpus and its sentences are taken from the People's Daily between 1993 and 1994. In 863, it contains only 2,573 unique sentences as each speakers is read multiple times by different speakers. In this experiment, we select only speakers from Beijing accents. 34 speakers (17 male and 17 female) are selected for training and another 12 for testing. The test set is designed such that there is no overlap between training and test speakers sentence. The details of the database are summarized in Table 2.

863		
Type	Train	Test
Accents	Beijing	Beijing
No. Speakers	17x2	6x2
No. Utterances	18,668	889
No. Syl/Utterances	12.6	12.4

Table 2. Corpus Structure of 863 training and testing

5.3. Baseline Experimental Result

Our baseline experiments used only the MFCC features for training and testing. Since our goal is to investigate the ef-

fect of integrating tone modeling into the recognition process, our experiments did not include any language model. All results are evaluated on tonal syllable accuracy. Our baseline HMM system gives a tonal syllable accuracy of 50.25%.

5.4. Tone Model

Our tone models were trained using tone labels provided in the 863 corpus that should reflect the actual tone pronounced by the speakers instead of the canonical tone marks associated with the character. The latter may be affected by tone changes commonly observed in Mandarin especially for the third tone.

To train the tone model, syllable boundarys were used and were obtained by aligning the syllable labels with the acoustics using trained HMM models. Since we are mostly interested in the shape of the pitch contour, pitch values with a syllable are normalized to have zero mean and unit covariance.

To measure the effectiveness of the tone models and select the suitable polynomial order, we performed tone classification experiments. Syllable segments were estimated by aligning the acoustic with the syllable labels using the Viterbi algorithm. Different orders of polynomials were estimated for each syllable. A simple classifier that selects the highest likelihood was used to classify the syllable tones and the results for using a single full-covariance Gaussian tone model with different orders of polynomial are tabulated in Table 3. We also investigated the effect of using diagonal Gaussian mixtures for tone models. However, it does not perform as well as the single full-covariance model.

Polynomial Order d	3	4	5
Tone Classification Acc. (%)	53.9	49.7	41.3

Table 3. Tone accuracy vs. order of polynomial

5.5. Recognition

We performed tonal syllable recognition using the one-pass integration of polynomial-fitted tone model likelihood into the Viterbi search process. Based on the results from the tone classifier, we used only the quadratic polynomial with a single full covariance Gaussian resulting in a tonal syllable accuracy of 50.34% as compared to 50.25% obtained in the baseline. While the increase in accuracy is small, it does demonstrate that the integration framework is feasible and can be explored more fully using more powerful models.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we propose the use of a polynomial trajectory model for tone modeling. We then propose a one-pass algorithm that can directly integrate the tone likelihood into the Viterbi recognition framework. The proposed framework improves the tonal syllable recognition accuracy slightly. In these experiments, we believe that the tone model can be further improved. For example, we plan to further explore the use of mixtures as well as a better polynomial fitting using the likelihood criterion. Furthermore, the acoustic realization of a phonetic unit is affected by its neighboring units. This contextual effect is also very noticeable on Mandarin [4] and the pitch period of each tones can be different due to neighboring context. We plan to explore the use of context dependent tone model to capture contextual effects.

7. ACKNOWLEDGEMENT

This work is partially supported by the Hong Kong RGC under the grant under HKUST 6049/OOE. The authors would like to acknowledge Wilson Tam for sharing his recognizer.

8. REFERENCES

[1] C.H. Lin et al. An initial Study on Mandarin Syllable Recognition Based on Sub-syllable Units. *Proc. ICCPCOL-91*, Taipei, Taiwan, pp.302-306, Aug, 1991.

[2] Hank C.H. Huang et al. Pitch Tracking and Tone Features for Mandarin Speech Recognition. In *Proc. ICASSP2000*, Vol.3 pp.1523-1526.

[3] Y.W. Wong and Eric Chang. The Effect of Pitch and Lexical Tone on Different Mandarin Speech Recognition Tasks. In *Proc. Eurospeech*, 2001, pp.1517-1521.

[4] Hsin-Min Wang, et "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data", In *IEEE Trans. on ASSP* VOL.5. NO.2 1997.

[5] Pui-Fung, Wong and Man-Hung Siu. Integration of Tone Related Feature for Chinese Speech Recognition. to appear in *IEEE Conference on Multimodel Interfaces 02*.

[6] Gish, H, Ng, K. Parametric trajectory models for speech recognition. In *Proc. ICSLP*, 1996 Vol 1. p466-469

[7] Henry Rogers. Theoretical and practical phonetic, Chapter 16, p.271, Copp Clark Pitman Publishing, 1991

[8] A.M. NOLL, "Cepstrum Pitch Determination" *J. Acoustic Soc. Am.*, Vol. 41, pp. 293-309, 1967

[9] C. Patavee, J. Somchai, A. Visarut, M. Ekkarit, F0 Feature Extraction by Polynomial Regression Function for Monosyllabic Thai Tone Recognition, In *Proc EUROSPEEH*, 2001.

[10] J. Zhang et al. Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition. *Proc. EUROSPEECH*, pp 1617-1620, 2001

[11] L.R. Rabiner et al. Applications of a non-linear smoothing algorithm to speech processing. In *IEEE Trans. ASSP*, VOL.23 pp.552-557, Dec 1975

[12] S. J. Young, N.H. Russel, J.H.S. Thornton: Token Passing: a Simple Conceptual Model for Connected Speech Recognition. Cambridge University Engineering Department 1989 pp 1-14.

[13] National performance assessment of speech recognition systems for Chinese. In *Proc of Oriental COCOSDA workshop '99*, Taipei 1999, pp.41-44.

[14] S. Young, *et al*, The HTK book (for HTK Version 2.2, *Entropic Ltd., 1999*