

Tone Pattern Discrimination Combining Parametric Modeling and Maximum Likelihood Estimation

Jinfu Ni and Hisashi Kawai

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, “Keihanna Science City” Seika-cho, Kyoto 619-0288 Japan

{jinfu.ni, hisashi.kawai}@atr.co.jp

Abstract

This paper presents a novel method for tone pattern discrimination derived by combining a functional fundamental frequency (F_0) model for feature extraction with vector quantization and maximum likelihood estimation techniques. Tone patterns are represented in a parametric form based on the F_0 model and clustered using the LBG algorithm. The mapping between lexical tones and acoustic patterns is statistically modeled and decoded by the maximum likelihood estimation. Evaluation experiments are conducted on 469 Mandarin utterances (1.4 hours of read speech from a female native) with varied analysis conditions of codebook sizes and tone contexts. Experimental results indicate the effectiveness of the method in both tone discrimination and detection of the inconsistency between a lexical tone and its F_0 pattern. The method is suitable for the prosodic labeling of a large scale speech corpus.

1. Introduction

Robust tone pattern discrimination is desirable in prosodic labeling and analysis of Chinese, focusing particularly on fundamental frequency (F_0) contours. The first reason for this is that, in Chinese, tones perform a discrimination function, like phonemes in a word. In Mandarin, there are four lexical tones carried by syllables, traditionally called the 1st to 4th tones or Tones 1 to 4, also denoted by H(igh), R(ise), L(ow) and F(all) tones, respectively, which are based on their pitch movements. Within the prosodic complex (basically the F_0 contours), the smallest distinctive configuration is tone. Tone is primary in that larger configurations, including the prosodic complex itself, are specific modifications of a string of one or more tones [1]. From this viewpoint, reliable analysis and labeling of the prosody must have the capability of dealing with the variability of the lexical tones under various conditions.

The second reason is related to the necessity of tone verification arising from the labeling of a large-scale, synthesis-oriented speech corpus. Chinese is written in characters, known as *Hanzi*. Each character corresponds to one syllable (but not vice versa of course), which is normally transcribed in *Pinyin* (a Roman alphabet form, literally meaning ‘putting together the sounds’). While there are literally tens of thousands of Chinese characters, there are only around 1,300 syllables. Because there are no word separators like the space used in English, the conversion of a *Hanzi* string into a *Pinyin* string is mostly done through morphological analysis. The use of auto-converted tones for labeling speech corpora suffers from conversion errors due to the word boundary ambiguity; 6.2% tone errors were found in an auto-converted *Pinyin* script including 229,423 syllables [2]. Tone pattern discrimination is a hopeful

candidate for the detection and correction of potential wrong tone labels.

Tone pattern discrimination deals with similar problems as tone recognition, like how to effectively model a tone. In continuous speech, tone patterns are subject to various modifications of contextual tones and intonation. Several studies associated with tone recognition in continuous Mandarin speech have been conducted in the past few years. HMM [3][4], Neural Networks [5] and decision trees [6] have been used in these studies; tone classification in continuous speech was proven to be a very hard task. In contrast with tone recognition, lexical tones are available for the task of tone pattern discrimination. Ample use of tone information can improve the performance of tone pattern discrimination.

In this paper, we present a data-driven method for tone pattern discrimination based on a functional F_0 model (hereinafter referred to as the model [7]). The idea is to utilize the model instead of dealing directly with the F_0 contour in order to bridge the gap between linguistic and prosodic features. Another advantage of the model is the small number of parameters required to represent F_0 contours, which potentially improves the efficiency of data-driven learning methods. A language-unspecific and quantitative representation of F_0 contours is possible based on the command-response model (also known as the Fujisaki model) [8]. However, difficulty in the automatic analysis of observed F_0 contours based on this model requires a manual aid for reliable work. Compared to the Fujisaki model, the model supports automatic analysis of the F_0 contours [9]. The model would bridge the gap between linguistic and prosodic features, and create constraints to reduce speaker-dependent effects, thus facilitating tone feature extraction and tone pattern modeling. On the other hand, the vector quantization (VQ) has shown good performance for clustering tone patterns [10]. The model and VQ are adopted with the maximum likelihood estimation technique for dealing with tone pattern discrimination.

The remainder of this paper explains the details of the method. Section 2 contains a description of the F_0 model, the parametric tone modeling, and the maximum likelihood estimation algorithm. Section 3 presents evaluation experiments and results. Comments and future work are described in Section 4.

2. Outline of the method

2.1. A functional F_0 model

The model [7] is a quantitative functional model that describes F_0 contours in logarithmic scale as a transposition of a range of concatenative mountain-shaped patterns into the voice register (a frequency register of utterances) of a speaker. The voice register of a speaker is normalized in RONDO (RatiO of Nat-

ural frequency of the system to that of Driving fOrce) through warping it along with the frequency response curve of a forced vibration system. The RONDO- F_0 contour is then expressed by the concatenative mountain-shaped patterns lined up in series at the time axis. The F_0 contour as a function of time t is given as follows.

$$\frac{\ln F_0(t) - \ln f_{0b}}{\ln f_{0t} - \ln f_{0b}} = \frac{A(\Lambda(t)) - A(\lambda_b)}{A(\lambda_t) - A(\lambda_b)}, \text{ for } t \geq 0, \quad (1)$$

where

$$A(\lambda) = \frac{1}{\sqrt{(1 - (1 - 2\zeta^2)\lambda)^2 + 4\zeta^2(1 - 2\zeta^2)\lambda}}, \text{ for } \lambda \geq 1, \quad (2)$$

and

$$\Lambda(t) = \Lambda_{r_1}(t) + \sum_{i=1}^{n-1} \text{Min}(\Lambda_{f_i}(t), \Lambda_{r_{i+1}}(t)) + \Lambda_{f_n}(t). \quad (3)$$

$\text{Min}(z_1, z_2)$ means the smaller of z_1 and z_2 . Equations (1) and (2) jointly indicate the transposition of the voice register. Equation (3) expresses the RONDO- F_0 contour $\Lambda(t)$, where $\Lambda_{r_i}(t)$ and $\Lambda_{f_i}(t)$ indicate the rise and fall components of the i th mountain-shaped pattern, respectively. Furthermore, $\Lambda_{x_i}(t)$, $x \in \{r, f\}$, is basically expressed as zero-input responses of a critically-damped second-order linear system [8]. Particularly,

$$\Lambda_{r_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{r_i}(1 - D_{r_i}(t_{p_i} - t)), & \text{for } t \leq t_{p_i}, \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

$$\Lambda_{f_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{f_i}(1 - D_{f_i}(t - t_{p_i})), & \text{for } t \geq t_{p_i}, \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

$$\text{where } D_{x_i}(t) = (1 + \frac{4.8t}{\Delta t_{x_i}}) e^{-\frac{4.8t}{\Delta t_{x_i}}}, \text{ for } t \geq 0. \quad (6)$$

In the model, parameters ζ , λ_t and λ_b can be commonly fixed at 0.237, 1 and 2, respectively [7]. There are then two speaker-dependent but utterance-independent parameters, namely,

$[f_{0b}, f_{0t}]$: top and bottom frequencies of the voice register,

and five utterance-dependent but speaker-independent parameters in the RONDO-time space,

- n : number of mountain-shaped patterns,
- Δt_{x_i} : response time for the i th rise/fall component,
- $\Delta \lambda_{x_i}$: amplitude of the i th rise/fall component,
- (t_{p_i}, λ_{p_i}) : peak of the i th mountain-shaped pattern, $i = 1, \dots, n$.

2.2. Parametric modeling of tone patterns

Previous work [7] indicated that six tone patterns are necessary for representing Mandarin F_0 contours. Figure 1 illustrates the modeling of the four lexical tones based on the model. Basically, there is a pattern of H-2P(eaks) for H tone, two patterns, R-1P(eak) and R-2P(eaks), for R tone, two patterns, L-1P(eak) and L-2P(eaks), for L tone, and a pattern F-1P(eak) for F tone. In Chinese, some syllables (usually neutral syllables) do not have inherent tones, and are known as N(eutral) tone or Tone 0. No extra tone patterns are needed for N tone, because its F_0 values are affected by the surrounding tones (usually the preceding syllable's tone). N tone may take one of these tone patterns according to its actual value.

These tone patterns can be represented in the parametric form below according to the number of peaks of mountain-shaped patterns to be used; a pattern is relocated via a setting parameter t_{p_i} .

$$1\text{P(eak) pattern: } \Delta t_{r_i}, \Delta \lambda_{r_i}, \lambda_{p_i}, \Delta t_{f_i}, \Delta \lambda_{f_i} \quad (7)$$

$$N_{r_i}, N_{f_i} \quad (8)$$

$$2\text{P(eak) pattern: } \Delta t_{r_i}, \Delta \lambda_{r_i}, \Delta t_{f_i}, \Delta \lambda_{f_i}, \Delta t_{r_{i+1}}, \Delta \lambda_{r_{i+1}}, \lambda_{p_{i+1}}, \quad (9)$$

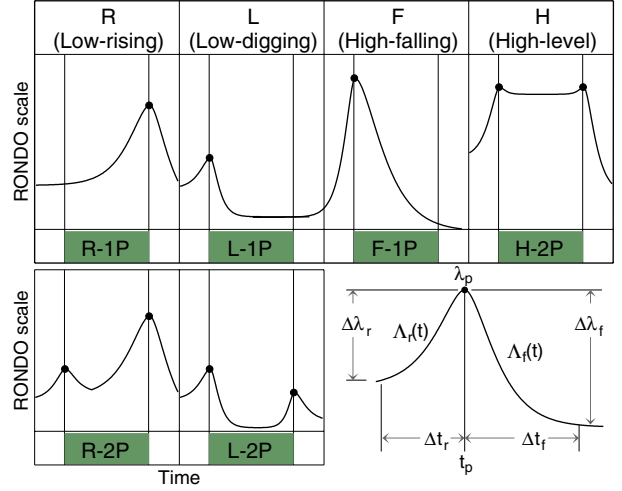


Figure 1: Modeling of Mandarin tones using mountain-shaped patterns. A mountain-shaped pattern with its control parameters is also superimposed on this figure. Solid circles indicate peaks.

$$\Delta t_{f_{i+1}}, \Delta \lambda_{f_{i+1}}, t_{p_{i+1}} - t_{p_i}, \lambda_{v_i} - \lambda_{p_i}, \lambda_{v_i} - \lambda_{p_{i+1}} \quad (9)$$

$$N_{r_i}, N_{f_i}, N_{r_{i+1}}, N_{f_{i+1}} \quad (10)$$

where N_{x_i} indicates the normalized number of voiced frames covering the i th rise / fall component, and λ_{v_i} indicates the value of the valley between the i th peak and the next. The parameters included in equations (7) and (9) are closely related to the model parameters. The parameters in equations (8) and (10) indicate which portions of a tone pattern are inherent for representation of the tone. When the number of voiced frames for N_{x_i} is less than 4, the Δt_{x_i} , $\Delta \lambda_{x_i}$ and N_{x_i} are all fixed at 0.2, 0.25 and 0, respectively, thus emphasizing the inherent tone features.

Tone patterns are clustered using the vector quantization technique. Two codebooks are made for representation of typical tone patterns observed from a training set using the LBG algorithm, i.e., codebooks for 1P(eak)- and 2P(eak)-patterns.

2.3. Maximum likelihood estimation algorithm

Let a tone sequence $T = t_1, t_2, \dots, t_n$ be represented by the observed sequence of (tone pattern) parameter vectors $O = o_1, o_2, \dots, o_n$, where o_i is the parameter vector for the i th tone $t_i \in \{N, H, R, L, F\}$, $i = 1, \dots, n$. Thus, the most probable tone sequence T is then regarded as that of computing

$$\arg \max_T \prod_{i=1}^n p(o_i | t_i) p(t_i | t_{i-1}, t_{i+1}). \quad (11)$$

The symbol $p(o_i | t_i)$ indicates the probability of observation o_i at tone t_i , and $p(t_i | t_{i-1}, t_{i+1})$ indicates the tone transition probability (tri-gram case). These conditional probabilities are estimated from the frequencies in a set of training data, and the likelihood is actually calculated in its logarithmic form.

The algorithm for model training and the estimation of underlying lexical tones from observed F_0 contours with phonetic labels can be simply described as the following modules.

Feature extraction Extraction of the tone feature parameters, i.e., those shown in equations (7) and (9), for individual utterances was done by the analysis-by-synthesis based pattern matching method in [9]. If the pattern type estimated for H tone is either R-1P or F-1P, an extra peak may be automatically

inserted through analysis of the estimated pattern features. After the estimation of parametric tone patterns, the parameters shown in equations (8) and (10) were then derived from the observed F_0 contours.

Making codebooks Two codebooks with given size were then made from a set of training data using the LBG algorithm, i.e., codebooks for 1P(eak)- and 2P(eak)-patterns. In the following experiments, all of the parameter vectors extracted from a set of 469 utterances were used for making the two codebooks.

Tone model training Using the VQ technique, the parameter vectors of the tone patterns were coded with one of the codebooks according to the peak number of the individual patterns. The discrete conditional probabilities required for calculating equation (11) were then estimated from a set of training data.

Tone pattern discrimination The observed parameter vectors for every tone in an utterance were quantized by the use of corresponding codebooks at first. The underlying tone sequence was estimated by computing the equation (11).

3. Experiment evaluation and result

Four experiments were run on 469 utterances, 1.4 hours of read-speech from a female native, to evaluate the method. The potential errors of lexical tones due to the auto-conversion of Chinese character strings to *Pinyin* strings were manually checked.

3.1. Search of tone pattern types using pattern matching

As a reference for tone pattern discrimination, the analysis-by-synthesis-based pattern matching technique was used to search for an appropriate tone pattern type from a limited number of candidates. An experiment was conducted with the following setup. The first 200 utterances were used for training the baseline models as the prototypes of tone patterns using the LBG algorithm. The codebook size was fixed at 128 or 256 for individual tones [9]. The candidates of tone pattern types assigned for individual tones are listed below, taking into account the rules of tone sandhi, e.g., $L \rightarrow R|_L$, and contextual tone changes in the syllables like *yīl* (one), *qīl* (seven), *bāl* (eight) and *bū4* (not), as well as tone neutralization due to weak stress [1] [11]. Namely, H-2P, R-1P and F-1P were assigned as candidates for a search of Tone 1; R-1P, L-1P and R-2P for Tone 2; L-1P, R-1P, L-2P and R-2P for Tone 3, F-1P and R-1P for Tone 4; and R-1P, F-1P and H-2P for Tone 0. The search of tone pattern types was then run over all of the 469 utterances, including 14,709 tones.

Figure 2 shows the distribution of tone pattern types for individual tones. Lexical tones and their sample count are listed on the horizontal axis. It can be seen from this figure that (1) though there exist notable variations in tone patterns, the tone is strong enough to keep its basic pattern against the effects of the other intonational factors, (2) The method for searching for tone pattern types is not good for tone pattern discrimination because a tone pattern type may correspond to several tones.

3.2. Tone discrimination with maximum likelihood estimation

Two experiments were conducted, focusing particularly on an estimation of the underlying lexical tones; the higher the accuracy of lexical tone discrimination, the higher the performance of tone pattern discrimination. The first was done with the aim of examining the effectiveness of the method for tone pattern discrimination. In this experiment, all of the 469 utterances were used as the training and test data (close test). The codebook size was fixed at 512. A tri-gram was used to model the

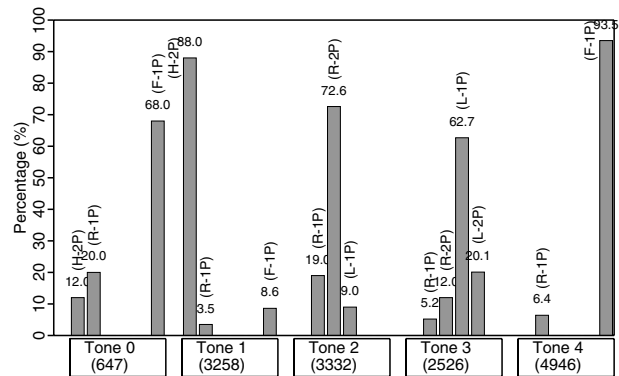


Figure 2: Distribution of tone pattern types for individual tones.

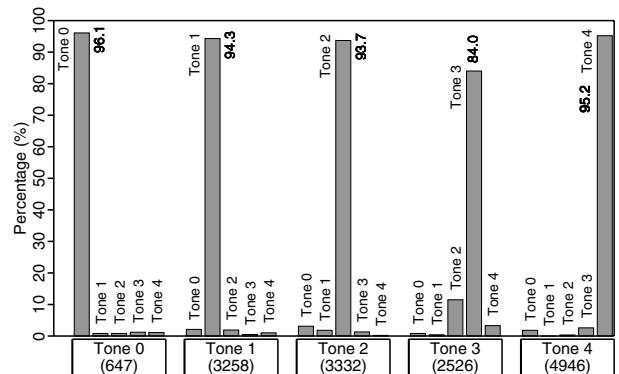


Figure 3: Tone distribution for a close test of tone discrimination with tri-gram model and maximum likelihood estimation.

contextual tone effects [11].

The experiment result is shown in figure 3. Three observations can be made from this figure. Firstly, the parametric modeling of tones is good enough to capture inherent tone features, including Tone 0, provided that its contextual tones are given. Secondly, it is clear that the proposed method is vital for tone pattern discrimination. Lastly, the phenomenon that Tone 3 is partly recognized as Tone 2 can be explained as a tone sandhi effect. The patterns of Tone 3 and Tone 4 are relatively more confusable than the others. In addition, there exists notable tone neutralization. This can be seen from the result that more Tones 1 - 4 were classified into Tone 0 but less vice versa, of course.

Another experiment was designed to investigate the effects of the following factors on the performance of tone pattern discrimination: codebook size, samples outside training set and tone context. Because the sample set was limited, 20-fold cross validation was adopted to simulate an open test. Particularly, the first 460 utterances were equally divided into 20 subsets. Each subset included 23 utterances. The test was done repeatedly. Every subset was in turn used as the test data; the others were then used as the training data. In this experiment, the codebook size was varied from 64 to 1024, combining with mono- and tri-gram context for tone modeling. The experimental results are listed in table 1.

From table 1, firstly, we can see that when the contextual tone effects have been considered (tri-gram model), the recognition accuracy is improved for all of the tone types, especially for Tone 0 and Tone 3. This phenomenon reveals that the patterns of Tones 0 and 3 suffer more contextual tone effects than the others. Secondly, the recognition accuracy increases with

Table 1 Accuracy of tone discrimination with maximum likelihood estimation under varied conditions (open test).

CB size	Context	Tone 0 %	Tone 1 %	Tone 2 %	Tone 3 %	Tone 4 %	Total %
64	Mono-tone	13.9	80.0	76.3	30.6	82.4	68.5
	Tri-gram	43.5	80.2	72.1	48.1	85.7	73.2
128	Mono-tone	31.4	78.6	79.8	41.6	82.4	71.7
	Tri-gram	64.4	82.4	80.3	60.6	87.5	79.1
256	Mono-tone	32.8	79.3	82.3	47.2	85.4	74.5
	Tri-gram	76.9	88.2	86.2	71.1	91.1	85.3
512	Mono-tone	40.2	82.3	84.0	52.3	88.4	77.8
	Tri-gram	92.5	93.7	92.3	81.4	94.1	91.4
1024	Mono-tone	45.7	82.8	84.5	53.3	88.9	78.5
	Tri-gram	93.7	93.6	93.4	83.2	95.4	92.4

Table 2 Result for tone change detection through tone pattern discrimination with maximum likelihood estimation.

Surface Lexical	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
Tone 0		1		1	3
Tone 1	49 (6)		12 (4)	1 (1)	28 (1)
Tone 2	88 (15)	1		10	2
Tone 3	5		259 (3)		
Tone 4	44 (3)	1	10 (1)	1	

the increase of codebook size up to 512, which is a suitable size for tone discrimination. Codebooks with size of 512 are also sensitive to the tone change and neutralization; detailed results are given next. The experiment results also indicate that tone context information plays a more important role than the codebook size in tone classification. This partly convinces us that the larger prosodic complex can not be predicted from observed F_0 patterns unless knowing underlying lexical tones [1]. It is noted that the recognition accuracy decreased around 2% in the open test compared to the close test with the same conditions.

3.3. Tone change detection

An experiment was conducted to test the capability of the method in tone change detection. It was an open test (as those mentioned above) with tri-gram model and codebook size of 512. There were 550 samples in the 460 utterances whose "surface tones" were inconsistent with the lexical tones. The term surface tones are opposite to lexical tones and are recognized according to tonal F_0 patterns. The surface tone samples here were manually determined by the rules of tone sandhi and contextual tone changes aided with visual inspection of the observed F_0 contours.

Table 2 lists the results for the detection of the inconsistency between lexical and surface tones, i.e., tone change, through the maximum likelihood estimation. In each box, the first indicates the number of tone change samples that were detected, while the second inside the brackets, if any, indicates what was not detected. For instance, there are 262 (= 259 + 3) Tone 3 samples that changed into Tone 2 in the recording, in which 259 samples were detected and 3 samples were not detected. In this experiment, the detection accuracy were 100% for Tone 0, 88.2% for Tone 1, 87.1% for Tone 2, 98.9% for Tone 3 and 93.3% for Tone 4. The detection accuracy for tone neutralization was lower than

the others, at 88.6% vs. 97.1%.

4. Comments and future work

We presented an efficient method for tone pattern discrimination that makes use of a functional F_0 model with vector quantization and maximum likelihood estimation techniques. The model bridges the gap between linguistic and prosodic features and makes it possible to use a few parameters to capture the inherent tone features. The mapping between lexical tones and acoustic patterns can be statistically modeled and decoded by the maximum likelihood estimation. Experimental results have confirmed the effectiveness of the proposed method in tone pattern discrimination and tone change detection. Future work will include an evaluation on a large-scale speech corpus and multiple speakers. The method may also be extended to other languages.

Acknowledgement This research was supported in part by the Telecommunications Advancement Organization (TAO) of Japan.

5. References

- [1] P. Kratochvil, "Intonation in Beijing Chinese," *Intonation Systems, A Survey of Twenty Languages*, ed. by D. Hirst and A. D. Cristo, Cambridge Uni. Press, pp. 417-431, 1998.
- [2] J. Ni and H. Kawai, "Tone Verification for Automatically Labeled Mandarin Speech Corpus", *Proc. Spring meeting of the acoustical society of Japan*, Vol. 1, pp. 373-374, 2003.
- [3] H. M. Wang *et al.*, "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data", *IEEE Trans. o ASSP*, Vol. 5, No. 2, 1997.
- [4] J. S. Zhang and K. Hirose, "Anchoring Hypothesis and Its Application to Tone Recognition of Chinese Continuous Speech", *ICASSP00*, pp. 2741-2744, 2000.
- [5] S. H. Chen and Y. R. Wang, "Tone Recognition of Continuous Mandarin Speech Based on Neural Networks", *IEEE Transactions on ASSP*, Vol.3, No.2, pp. 146-150, 1995.
- [6] Y. Cao, Y. Deng, H. Zhang, T. Huang, B. Xu, "Decision Tree Based Mandarin Tone Model and Its Application to Speech Recognition", *ICASSP00*, pp. 1610-1613, 2000.
- [7] J. Ni and K. Hirose, "Experimental Evaluation of a Functional Modeling of Fundamental Frequency Contours of Standard Chinese Sentences," *ISCSLP00*, pp.319-322, 2000.
- [8] H. Fujisaki and K. Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," *J. Acoust. Soc. Jpn(E)*, Vol.5, No.4, pp.233-242, 1984.
- [9] J. Ni and H. Kawai, "Tone Feature Extraction Through Parametric Modeling and Analysis-by-Synthesis-based Pattern Matching", *ICASSP03*, Hong Kong, 2003.
- [10] S. H. Chen and Y. R. Wang, "Vector Quantization of Pitch Information in Mandarin Speech," *IEEE Transactions on Communications*, Vol.38, pp. 1317-1320, 1990.
- [11] Y. Xu, "Contextual Tonal Variations in Mandarin," *Journal of Phonetics*, 25, pp. 61-83, 1997.