TONE FEATURE EXTRACTION THROUGH PARAMETRIC MODELING AND ANALYSIS-BY-SYNTHESIS-BASED PATTERN MATCHING

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

A functional fundamental frequency (F_0) model is applied to extract tone peak and gliding features from Mandarin F_0 contours aiming at automatic prosodic labeling of a large scale speech corpus. Modeling four lexical tones and representing them in a parametric form based on the F_0 model, we first cluster baseline tone patterns using the LBG algorithm, then perform analysis-by-synthesis-based pattern matching to estimate underlying tone peaks and tone pattern types from observed F_0 contours and phonetic labels with lexical tones. Tone gliding features are re-estimated after the determination of tone peaks. 94% of the automatically estimated labels were consistent with the manual labels in an open test of 968 utterances from eight native speakers. Also, experimental results indicate that the proposed method is applicable for F_0 contour smoothing and tone verification.

1. INTRODUCTION

The prosodic properties of speech, such as tone, pitch accent, and intonation, are mainly manifested by fundamental frequency (F_0) contours. In Chinese, tones carried by syllables perform a discrimination function, like phonemes in a word. 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]. Therefore, tone feature extraction becomes essential for prosodic analysis and labeling.

The ToBI labeling system[2], originally developed for English, is now being adapted to other languages. However, it lacks quantitative representation of F_0 contours and suffers from fluctuations between labelers. On the contrary, a language-unspecific and quantitative labeling of prosodic features is possible based on the command-response model (also known as the Fujisaki model) [3]. However, difficulty in the automatic analysis of observed F_0 contours based on this model requires a manual aid for reliable labeling. In addition, several methods exist that make use of tone features for speech information processing, e.g., [4, 5]. However, no automatic estimation yet offers reliable and accurate tone features.

In this paper, we propose an efficient data-driven method to extract tone features from F_0 contours that makes use of a functional F_0 model developed for Chinese (hereinafter referred to as the model)[6]. An advantage of the model, compared to the Fujisaki model, is that it supports automatic analysis of the F_0 contours. The model would bridge the gap between linguistic and acoustic F_0 features, and create constraints to reduce speaker-dependent effects, thus facilitating data-driven learning and feature extraction.

The remainder of this paper explains details of the method. Section 2 contains a description of the F_0 model, the tone modeling and baseline model training. Section 3 describes algorithms for parametric estimation of tone features, including tone peaks, gliding and pattern type. Finally, experimental results comparing automatic estimation results with manual labels are described in Section 4, and comments and future work are given in Section 5.

2. PARAMETRIC MODELING OF THE F_0 CONTOURS

2.1. A functional F_0 model

In the previous work [6], a functional model was proposed to represent an F_0 contour in Mandarin. The vocal-cord vibration system is regarded as a second-order forced vibration system for modeling of the control mechanism of the F_0 contour generation process. The voice register (a frequency register of utterances) of a speaker is transposed to RONDO (RatiO of Natural frequency of the system to that of Driving fOrce) through warping it along with the frequency response curve of the forced vibration system. The RONDO-F_0 contour is then expressed in 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.

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

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

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

Min(z_1, z_2) means the smaller one of both 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 ith mountain-shaped pattern, respectively. Following the modeling of the accent control mechanism in the command-response model [3], $\Lambda_{x_i}(t)$, $x \in \{r, f\}$, is basically expressed as zero-input responses of a critically-damped second-order linear system. 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 [6]. There are then two speaker-dependent but utterance-independent parameters in frequency domain, namely,
$[f_{0_b}, f_{0_t}]$: 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 ith rise/fall component,
$\Delta \lambda_{x_i}$: amplitude of the ith rise/fall component,
(t_{p_i}, λ_{p_i}) : peak of the ith mountain-shaped pattern, $i = 1, ..., n$.

When given the voice register $[f_{0_b}, f_{0_t}]$, the F_0 contour can be converted into and be then treated in the RONDO-time space. Henceforth, let a variable taking $\hat{}$ denote that its value is observed from real speech opposite to that generated by the model. An algorithm for converting $\hat{F}_0(t)$ to $\hat{\Lambda}(t)$ is described as follows.

Algorithm 1 Conversion of $\hat{F}_0(t)$ to $\hat{\Lambda}(t)$ given $[f_{0_b}, f_{0_t}]$

λ=λ_t.
if $\hat{F}_0(t) \geq f_{0_t}$, go to Outlet.
Loop : $F_0(t) = e^{\frac{A(\lambda)-A(\lambda_b)}{A(\lambda_t)-A(\lambda_b)} * (\ln f_{0_t} - \ln f_{0_b}) + \ln f_{0_b}}$.
If $F_0(t) \leq \hat{F}_0(t)$, go to Outlet.
$\lambda = \lambda + 0.0001$.
Go back to Loop.
Outlet : $\hat{\Lambda}(t) = \lambda$.

2.2. Tone modeling

In Mandarin Chinese, every syllable carries one of the five lexical tones, traditionally called the 0th to 4th tones, also denoted by N(eutral), H(igh), R(ise), L(ow) and F(all) tones, respectively, which are based on their pitch movements.

Tone modeling needs to consider not only the main tone features, including peak and gliding (rise, level and fall) features, and tonic F_0 pattern (henceforth, tone pattern) type, but also acoustic requirements for generating smooth F_0 contours. Consequently, six tone patterns are necessary for Mandarin F_0 contours. Figure 1 illustrates the modeling of tone patterns based on the model; accordingly, they are expressed in parametric form in equations (7) and (8) and relocated via a setting parameter t_{p_i}. Basically, there is a pattern 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. There needs no extra tone pattern for the N tone, because its F_0 values depend on its contextual tones; it may take one of these tone patterns according to its actual value.

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

$$\text{2P(eak) case} : < \Delta t_{r_i}, \Delta \lambda_{r_i}, \lambda_{p_i}, \Delta t_{f_i}, \Delta \lambda_{f_i}, \Delta t_{r_{i+1}}, \Delta \lambda_{r_{i+1}}, \lambda_{p_{i+1}} - \lambda_{p_i}, \Delta t_{f_{i+1}}, \Delta \lambda_{f_{i+1}}, t_{p_{i+1}} - t_{p_i} > . \quad (8)$$

2.3. Training baseline tone models

A set of baseline models was trained as the prototypes of tone patterns, which are used in Analysis-by-Synthesis (AbS) based pattern matching for speaker-independent estimation of the peak parameters (discussed in Section 3.1). Since speaker-dependent effects are largely reduced with the transposition of voice registers,

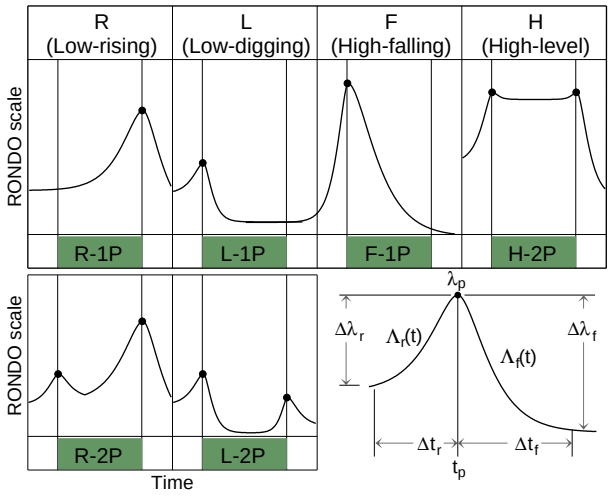

Fig. 1. Modeling of Mandarin tones using mountain-shaped patterns. A mountain-shaped pattern with its control parameters is also superimposed on this figure. Solid circles indicate peaks.

Table 1. Note for training baseline tone models

Tone	Tone pattern	Primary model parameters representing the tone pattern	Codebook size	Tone count
N	n/a			
H	H-2P	$\lambda_{p_i}, \Delta t_{f_i}, \Delta \lambda_{f_i}, \lambda_{p_{i+1}}, t_{p_{i+1}} - t_{p_i}$	128	727
R	R-1P	$\Delta t_{r_i}, \Delta \lambda_{r_i}, \lambda_{p_i}$	128	462
	R-2P	$\lambda_{p_i}, \Delta t_{f_i}, \Delta \lambda_{f_i}, \Delta t_{r_{i+1}}, \Delta \lambda_{r_{i+1}}, t_{p_{i+1}} - t_{p_i}, \lambda_{p_{i+1}}$	143	1089
L	L-1P	$t_{p_i}, \lambda_{p_i}, \Delta t_{f_i}, \Delta \lambda_{f_i}$	128	833
	L-2P	$\lambda_{p_i}, \Delta t_{f_i}, \Delta \lambda_{f_i}, \Delta t_{r_{i+1}}, \Delta \lambda_{r_{i+1}}, t_{p_{i+1}} - t_{p_i}, \lambda_{p_{i+1}}$	131	165
F	F-1P	$\lambda_{p_i}, \Delta t_{f_i}, \Delta \lambda_{f_i}$	256	2541

200 read-utterances (0.7 hour speech) from a speaker FR were used for setting up the training data set. The alignment between a tone and its underlying tone pattern was manually checked. To emphasize the tone features, only the parameters listed in Table 1 were extracted from the training data, while the other parameters Δt_x and $\Delta \lambda_x$ were all fixed at 0.2 and 0.25, respectively.

The LBG algorithm was used to cluster the samples assigned to a tone pattern, and form a codebook for the tone pattern. The number of tone samples and resultant codebook size for individual tone patterns can be found in Table 1.

3. PARAMETER ESTIMATION

Taking advantage of the model that gives the peak parameters (t_{p_i}, λ_{p_i}), $i = 1, ..., n$, the other parameters can then be easily estimated from the F_0 contour with minimal distortion. The first step in analyzing the F_0 contour is tone peak detection.

3.1. Estimating t_{p_i} and λ_{p_i} for peak feature

The model parameters t_{p_i} and λ_{p_i} are estimated by performing AbS-based pattern matching of a tone pattern with a specific F_0 fragment. Here, we at first assume that the tone pattern type is

given; Section 3.3 discusses how to ascertain the tone pattern type from the candidates. The time scope of the F_0 fragment is determined by syllable boundaries according to given phonetic labels. AbS technique is used to locally adjust the peak parameters when performing the pattern matching. In the 1P(eak) case, searching λ_{p_i} over [$\bar{\lambda}_{p_i}$-0.1, $\bar{\lambda}_{p_i}$+0.1] steps to 0.02, where $\bar{\lambda}_{p_i}$ indicates the parameter value set by the prototype used. In the 2P(eak) case, the adjustment is carried out on $\lambda_{p_{i+1}}$ and $t_{p_{i+1}}$ by similar methods. The time scope limit to search for t_{p_i} is determined by analysis of the curve shape of the F_0 fragment. The parameter estimation is then performed by the following steps.

Step 1 : Initialize/update t_{p_i} and the other parameters if need be.
Step 2 : Calculate the mean square error (MSE) between the observed and model-generated F_0 fragments. If the MSE is less than the existing minimal MSE, perform the AbS-based peak adjustment. Otherwise, go back to Step 1.
Step 3 : Terminate the current search if the MSE is less than a given threshold or the search has covered the time scope.
Step 4 : Turn to the next prototype, if any, and repeat from Step 1.

3.2. Estimating Δt_{x_i} and $\Delta \lambda_{x_i}$ for gliding features

Parameters Δt_{x_i} and $\Delta \lambda_{x_i}$ are jointly estimated by the iteration process from the observed (rise/fall) component $\hat{\Lambda}_{x_i}(t)$ given its peak (t_{p_i}, λ_{p_i}). For convenience, let $\hat{\Lambda}_{x_i}(t_j)$ denote the *j*th sample of $\hat{\Lambda}_{x_i}(t)$, $j = 0, ..., N_{x_i}$, and assume $t_0 = t_{p_i}$. Timing $t_{N_{x_i}}$ is defined at the lowest valley between two adjacent peaks. Accordingly, $t_j = t_{p_i} - j \times tint$ for the rise component, and $t_j = t_{p_i} + j \times tint$ for the fall component, where *tint* indicates a time interval between adjacent F_0 frames. An AbS-based iteration algorithm is described below for the parameter estimation.

Algorithm 2 Estimating Δt_{x_i} and $\Delta \lambda_{x_i}$ from an observed component $\hat{\Lambda}_{x_i}(t_j), j = 0, ..., N_{x_i}$

$\Delta t_{x_i} = 4 \times tint$ (*tint* = 0.01 in this experiment). $\bar{\epsilon}$ = 999.
Loop 1 : $\Delta \lambda_{x_i}$ = 0.04.
Loop 2 : $\epsilon = \sum_{j=0}^{N_{x_i}} \{\frac{1}{2}|\Lambda_{x_i}(t_j) - \hat{\Lambda}_{x_i}(t_j)| + \frac{1}{8}\sum_{k=1}^{5} |\Delta_k \Lambda_{x_i}(t_j) - \Delta_k \hat{\Lambda}_{x_i}(t_j)|\}$.
If $\epsilon < \bar{\epsilon}$, $\Delta\bar{\lambda}_{x_i} = \Delta\lambda_{x_i}$; $\Delta\bar{t}_{x_i} = \Delta t_{x_i}$; and $\bar{\epsilon} = \epsilon$.
$\Delta \lambda_{x_i} += 0.01$.
If $\Delta\lambda_{x_i} < 1$, go back to Loop 2.
$\Delta t_{x_i} += tint$.
If $\Delta t_{x_i} \leq 0.4$ (terminal threshold), go back to Loop 1.
Outlet : $\Delta t_{x_i} = \Delta\bar{t}_{x_i}$ and $\Delta\lambda_{x_i} = \Delta\bar{\lambda}_{x_i}$.

Here, $\Delta_k \Lambda(t_j)$ indicates the *k*-interval difference expressed as $\Delta_k \Lambda(t_j) = \Lambda(t_{j-k}) - \Lambda(t_j)$ for both $\Lambda(t_{j-k})$ and $\Lambda(t_j)$ being voice frames. Otherwise, $\Delta_k \Lambda(t_j) = 0$. The distortion ϵ between observed and model-generated components covers both the absolute error $\sum_{j=1}^{N_{x_i}} |\Lambda_{x_i}(t_j) - \hat{\Lambda}_{x_i}(t_j)|$ and the error concerning the multiple *k*-interval differences $\sum_{j=1}^{N_{x_i}} \sum_{k=1}^{5} |\Delta_k \Lambda_{x_i}(t_j) - \Delta_k \hat{\Lambda}_{x_i}(t_j)|$. The AbS-based iteration process minimizes the distortion to seek an optimal value pair for Δt_{x_i} and $\Delta \lambda_{x_i}$ relevant to the amplitude and response time of the component.

3.3. Ascertaining parameter n and tone pattern type

In the framework of tone modeling, the parameter n for an F_0 contour is determined by estimating the sequence of tone patterns for the F_0 contour's underlying tone specification. The tone shapes often deviate from the expected canonical shapes, even in read speech. The situation is particularly difficult in speech where a tone can be realized with a shape opposite to the underlying specification in isolated words. In read speech, the unexpected tone shape is largely a consequence of tone sandhi, e.g., L→R|_L, and contextual tone changes in the syllables like *yi1* (one), *qi1* (seven), *ba1* (eight) and *bu4* (not), as well as tone neutralization due to weak stress [1, 7].

Table 2. Specification of tone pattern candidates for lexical tones

Lexical tone	H	R	L	F	N
Tone pattern candidates	H-2P R-1P F-1P	R-1P L-1P R-2P	L-1P R-1P L-2P R-2P	F-1P R-1P	R-1P F-1P H-2P

We treat the issue of tone variations by searching multiple tone patterns for a lexical tone. Table 2 lists the primary tone pattern candidates for each lexical tone. Furthermore, to suppress peak insertion errors, we search a path over the candidate tone pattern matrix with minimal F_0 mismatch errors for the tone sequence. The parameter estimation is described in the following steps.

Step 1 : Estimate the parameters for each tone pattern assigned for every tone in an utterance, using the method in section 3.1.
Step 2 : If large errors present, report mismatches of tone patterns with a tone, then search the remaining tone patterns for it.
Step 3 : Select 2 tone patterns according to the MSE for each tone.
Step 4 : Find out the tone pattern sequence with minimal MSE for the utterance. In the pruning process, parameters $\Delta t_{x_i}$and $\Delta \lambda_{x_i}$, $i = 1, ..., n$, are all re-estimated by Algorithm 2.

4. EXPERIMENTAL RESULTS

To test the performance of the proposed method, experiments were conducted on 968 utterances adopted from eight native speakers in three Mandarin speech corpora, HKU-96, USTC-96, and a synthesis-oriented corpus recently built up at ATR. There were 268 utterances taken from the speaker FR, whose another 200 utterances were used for training the baseline tone models. The rest were picked up from the others (100 utterances for each). Experimental evaluation is mostly based on comparison of the automatically estimated results, along with the manual labels, and error analysis.

An example result of the tone feature extraction for speaker LW is illustrated in Figure 2 and Table 3. Table 4 lists the number of samples analyzed and the experimental results for each speaker. The column PbAM indicates the *p*ercentage of consistency *b*etween the *a*utomatically estimated results and the *m*anual labels under two constrained conditions. The first is that both tone pattern types are the same, except for R-1P and R-2P (as well as L-1P and L-2P), which were regarded as the same thing. The other is that parameter t_{p_i} is allowed to flutter within three frames around a manual label. The next two columns μ_{e_m} and μ_{e_a} indicate mean of errors between observed and re-synthesized F_0 contours, where e_m indicates frame errors $|\hat{F}_0(t) - F_0(t)|$ calculated with the manual labels, and e_a indicates those with the automatically estimated results. The last two columns indicate the percentage of frames with errors of $e_a \leq \mu_{e_m}$ and $e_a \leq 2\mu_{e_m}$, respectively. On the other hand, Table 5 lists the distribution of tone pattern candidates for individual tones produced by the male speaker LW and the female

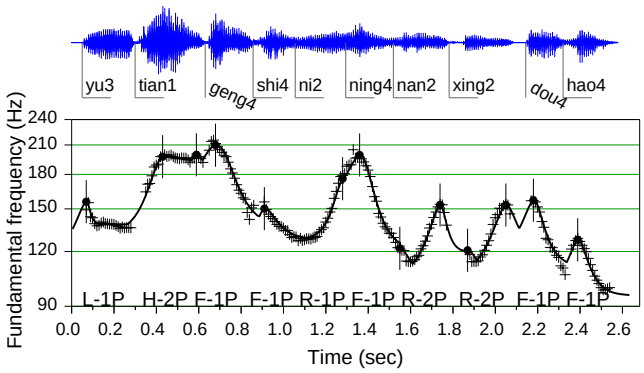

Fig. 2. Example of re-synthesized F_0 contours (lines) using the results estimated by this method (partly listed in Table 3). Vertical bars mark peaks to clarify them, and "+" denotes observed F_0.

Table 3. Partial parameter values for the example shown in Fig. 2.

Syllable	Tone	Pattern type	i	t_{p_i} (Sec)	λ_{p_i}	Δt_{r_i} (Sec)	$\Delta\lambda_{r_i}$	Δt_{f_i} (Sec)	$\Delta\lambda_{f_i}$
yu3	L	L-1P	1	0.06	1.38	0.20	0.25	0.05	0.10
tian1	H	H-2P	2	0.42	1.16	0.23	0.36	0.38	0.04
			3	0.59	1.15	0.20	0.25	0.20	0.25
geng4	F	F-1P	4	0.67	1.07	0.20	0.25	0.26	0.42
	...	...				...			

speaker FR, whose samples analyzed here include both the open test set and those from the training set (200 utterances).

Three points are clear from Figure 2 and Tables 3 to 5. Firstly, the proposed method is vital and speaker-independent; it shows quite good performance for all of the speakers analyzed. Secondly, 94% accuracy is obtained on average for the tone labeling. It improves the performance by around 10% compared to the F_0 peak-detection based method [8]. Lastly, a few parameters can capture the tone features: two parameters for tone peaks and two other parameters for tone gliding. There exist more than 83% of the observations closely waving around the re-synthesized F_0 contours. In addition, the results, showing low average frame error, which slightly depends on speakers, and the ability to select a tone pattern from candidates, implied that this method is applicable for F_0 smoothing and tone verification for a large scale speech corpus.

5. COMMENTS AND FUTURE WORK

We presented a method for the efficient extraction of tone features from F_0 contours that makes use of a functional F_0 model. The model bridges the gap between linguistic and acoustic F_0 features. Therefore, it not only facilitates data-driven learning and parameter estimation, but also makes it possible to use few parameters to capture the tone features. The parametric form of tone features is simple and agrees with the findings in recent tone research [1, 7]. Experimental results have confirmed the effectiveness of the proposed method in real Mandarin speech with multiple speakers.

Some work is worthwhile doing in future. Firstly, the size of the codebook for baseline tone models could be reduced via clustering a codebook of parameters Δt_x and $\Delta\lambda_x$. Secondly, the method may be extended to other languages.

Table 4. Speech samples analyzed and experimental results.

Speaker (sex)	$[f_{0_b}, f_{0_t}]$ Hz Hz	Tone count	PbAM %	μ_{e_m} Hz	μ_{e_a} Hz	$e_a \leq \mu_{e_m}$ %	$e_a \leq 2\mu_{e_m}$ %
FR (F)	[95 , 365]	7,178	94.8	4.46	4.85	64.7	81.9
WL (F)	[120,440]	1,589	95.0	3.38	4.00	60.4	81.9
ZYG (F)	[95 , 360]	1,192	93.3	3.16	3.39	63.0	83.7
LWX(F)	[140,395]	1,030	93.0	3.58	3.43	64.4	86.4
LW (M)	[85 , 210]	1,007	92.8	1.97	2.36	63.1	82.2
LYF (M)	[95 , 220]	992	94.6	2.12	2.12	65.3	83.3
GYQ(M)	[95 , 280]	910	94.6	2.08	2.19	64.7	83.5
SFW(M)	[65 , 170]	1,004	94.7	1.73	1.78	65.5	83.1
Average			94.1			63.9	83.3

Table 5. Distribution of tone pattern candidates for each tone.

Tone	Sample count for FR and LW		H-2P %	R-1P %	R-2P %	L-1P %	L-2P %	F-1P %
H	3,114	(FR)	87.6	3.7				8.7
	196	(LW)	79.9	10.2				9.9
R	3,196	(FR)		18.7	72.4	8.8		
	199	(LW)		38.2	50.2	10.6		*1.0*
L	2,402	(FR)		5.0	11.6	62.7	20.1	*0.5*
	127	(LW)		6.3	9.4	64.6	19.7	
F	4,726	(FR)	*0.2*	6.2			*0.3*	93.8
	478	(LW)		11.5				88.5
N	619	(FR)	12.8	19.9				67.2
	23	(LW)	4.3	30.4				65.2

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

6. 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] K.E.A. Silverman *et al* , "TOBI: A Standard for Labeling English Prosody," *Proc. ICSLP*, pp.867-870, 1992.

[3] 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.

[4] 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.

[5] L-S Lee, C-Y Tseng and C-J Hsieh, "Improved Tone Concatenation Rules in a Formant-based Chinese Text-to-Speech System", *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No.3, pp. 287-294, 1993.

[6] J. Ni and K. Hirose, "Experimental Evaluation of a Functional Modeling of Fundamental Frequency Contours of Standard Chinese Sentences," *Proc. ISCSLP*, pp.319-322, Beijing, 2000.

[7] Y. Xu, "Contextual Tonal Variations in Mandarin," *Journal of Phonetics*, 25, pp. 61-83, 1997.

[8] J. Ni and K. Hirose, "Automatic Prosodic Labeling of Multiple Languages Based on a Functional Modeling of Fundamental Frequency Contours", *Proc. of the Workshop on MSC2000*, pp.65-68, Kyoto, 2000.