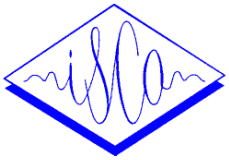

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SYNTHESIS OF FUNDAMENTAL FREQUENCY CONTOURS OF STANDARD CHINESE SENTENCES FROM TONE SANDHI AND FOCUS CONDITIONS

Jinfu Ni * *and Keikichi Hirose* **

* Dept. of Information and Communication Engineering, School of Engineering, University of Tokyo, Japan
** Dept. of Frontier Informatics, School of Frontier Sciences, University of Tokyo, Japan
E-mails: njf@gavo.t.u-tokyo.ac.jp hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

A new method was developed to synthesize fundamental frequency (F_0) contours of Chinese sentences based on a functional model, formerly developed by the authors. The model has an advantage in that decomposition process to phrase and tone components is not necessary. The developed method decides the F_0 contour type for each word based on a set of tone-sandhi rules, and then shapes these word F_0 contour types into the phrasal F_0 contour taking focus conditions into account. The tone-sandhi rules are formulated as 19 bi- and 198 tri-tone-sandhi contouremes in a parametric form, which are obtained by quantitative analysis of 84 di-, 538 tri- and 938 tetra-syllable words. Each contoureme can realize F_0 contours in 3 different ranges: normal, depressed and expanded. The focus conditions decide range type for each word and a peak reference-line for a phrase. When generating phrasal F_0 contour, the model parameter values of contouremes, selected using tone sequence and range type information, are adjusted so that F_0 peaks appear along the peak reference-line. Experimental results confirmed the validity of the proposed method.

1. INTRODUCTION

It is well known that prosodic properties of speech, such as tone and intonation, are mainly manifested by fundamental frequency (F_0) contours. This is especially true for tonal languages such as Chinese. Therefore, generating F_0 contours relevant to input linguistic and also para-/non-linguistic information comes an important issue in Chinese speech synthesis. Although several methods have been already developed for F_0 contour generation, and some realized rather good quality of speech, the generation was syllable-based with limited considerations on tone-sandhi effects, resulting in lack of smoothness when synthetic speech was listened to. In this paper, model-based F_0 contour generation for Chinese speech is addressed.

A command-response model of F_0 contours was proposed assuming an F_0 contour being the superposition of accent and phrase components [1]. Although this model can represent an observed F_0 contour well in any languages (tone components instead of accent components, in the case of tonal languages), analysis of Chinese sentence F_0 contours based on the model comes difficult. This is because the decomposition process of the contours into phrase and tone components involved in the analysis is erroneous due to large undulations in the F_0 contours. From this viewpoint, we formerly have proposed a functional model (the model, henceforth) for Chinese [2], which does not include the decomposition process. The model can generate a contour close to an observed F_0 contour only from F_0 peaks of constituting syllables. Here, a model-based method for Chinese F_0 contour synthesis is developed and discussed. The rest of the paper consists as follows; the proposed method is outlined in section 2, and experimental evaluation is presented with a brief discussion in section 3. The paper is concluded with section 4.

2. OUTLINE OF THE METHOD

In this section, after explaining the model, relations between model parameters and synthesized contours are discussed.

2.1. Model outline

Based on the model [2], a Chinese F_0 contour as a function of time can be expressed by

$$\frac{\ln \mathrm{F_0(t)} - \ln \mathrm{f_{0_b}}}{\ln \mathrm{f_{0_t}} - \ln \mathrm{f_{0_b}}} = \frac{\mathrm{T}(\Lambda(\mathrm{t})) - \mathrm{T}(\lambda_\mathrm{b})}{\mathrm{T}(\lambda_\mathrm{t}) - \mathrm{T}(\lambda_\mathrm{b})}, \text{for } \mathrm{t} \geq 0, \quad (1)$$

where

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

and

$$\Lambda(\mathrm{t}) = \Lambda_{\mathrm{r}_1}(\mathrm{t}) + \sum_{\mathrm{i}=1}^{\mathrm{n}-1} \mathrm{Min}(\Lambda_{\mathrm{f_i}}(\mathrm{t}), \Lambda_{\mathrm{r_{i+1}}}(\mathrm{t})) + \Lambda_{\mathrm{f_n}}(\mathrm{t}) \ . \quad (3)$$

Symbol $\mathrm{Min}(z_1, z_2)$ means to take the smaller one from z_1 and z_2. Equations (1) and (2) jointly indicate the frequency scale transposition which the model introduced. Symbol f_0 refers to frequency; t refers to time; λ means Ratio Of Natural frequency to Driving fOrce frequency (RONDO) in a forced vibration system, whose frequency responses can be expressed as Eq.(2). ζ denotes damping ratio of the system. RONDO scale is defined upon Eq.(2) for $\lambda \geq 1$. Then, an F_0 contour in the RONDO-time space can be represented by concatenation of mountain-shape contours as indicated in Eq.(3). Each mountain-shape contour is composed of *rise* and *fall* components, which are further expressed by

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

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

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

indicates decaying responses over time of a critically damped second order linear system. Symbol x represents either r(ise) or f(all). The model parameters [2] in equations (1) to (6) indicate

$[f_{0_b}, f_{0_t}]$: bottom and top frequencies of voice register of a speaker,
$[\lambda_b, \lambda_t]$: bottom and top values of voice register on the RONDO scale,
ζ : damping ratio of the forced vibration system,
n : number of mountain-shaped (*rise-(peak)-fall)* patterns,
(t_{p_i}, λ_{p_i}) : ith *peak* coordination,
$\{\Delta t_{r_i}, \Delta\lambda_{r_i}\}$: parameters controlling the ith *rise*,
$\{\Delta t_{f_i}, \Delta\lambda_{f_i}\}$: parameters controlling the ith *fall*,

where $i = 1, ..., n$. It is noted that f_{0_b} and f_{0_t} are mapped onto λ_b and λ_t, respectively, and an F_0 value is suppressed to f_{0_t} if it is larger than f_{0_t}.

Fig.1 illustrates the block-diagram of analysis and synthesis process of a Chinese F_0 contour based on the model. Evaluation experiments in [3] confirmed that the model could represent an observed F_0 contour well when the model parameter values were provided appropriately. Basically, parameters ζ, λ_b and λ_t can be fixed to 0.237, 1.98 and 1.0 respectively for all speakers and utterances. Parameters f_{0_b} and f_{0_t} specify top and bottom values of a frequency register of utterances of a speaker. The other parameters $\Delta t_{r_i}, \Delta\lambda_{r_i}, t_{p_i}, \lambda_{p_i}, \Delta t_{f_i}$ and $\Delta\lambda_{f_i}, i = 1, ..., n$, are tightly related to and thus convey linguistic (and para-, non-linguistic) information of utterances in the RONDO-time space.

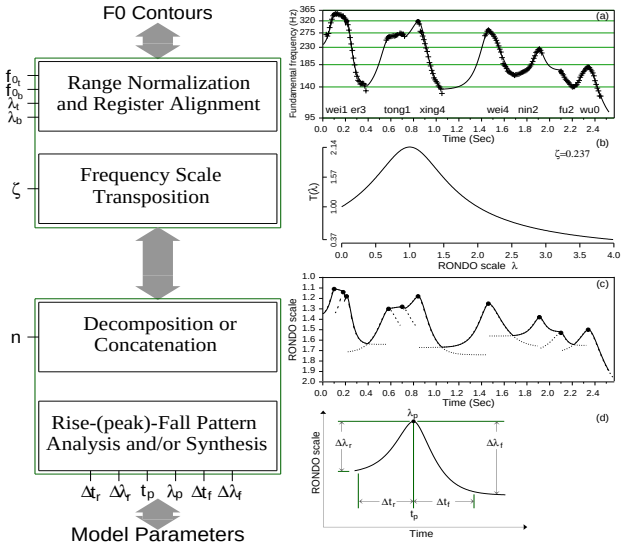

Fig. 1. Block-diagram of analysis and synthesis of a Chinese F_0 contour. On the right side, panel (a) plots the F_0 contour ('+' symbol) and its model approximation (solid lines) in logarithmic frequency, panel (b) displays a fragment of $T(\lambda)$ when $0 \leq \lambda \leq 4.0$, panel (c) plots the model-generated contour (solid lines) and underlying *rise-(peak)-fall* pattern components (dashed lines) in the RONDO-time space, and panel (d) indicates a *rise-(peak)-fall* pattern with its control parameters. Solid circles in (c) and (d) indicate the *peak*s.

2.2. Modeling lexical tones in parametric form

There exist four lexical tones in SC, namely, **H** tone (also known as Tone 1) characterized by high-level F_0 contour, **R** tone (Tone 2) by low-rising, **L** tone (Tone 3) by low-level, and **F** tone (Tone 4) by high-falling. In addition, there is a neutral tone **N** (Tone 0), whose F_0 contour loses its original shape and changes depending on the preceding tone. Fig.2 illustrates modeling of these tones using *rise-(peak)-fall* patterns, simply called *basic pattern*s, hereafter. Two types of initial *basic pattern* are available for **L** with either high *peak* 'e' or low *peak* 'f'; a high initial *peak* usually occurs in a syllable with voiced *initial* when it follows to a syllable with **H**, **R** or **L** tone type.

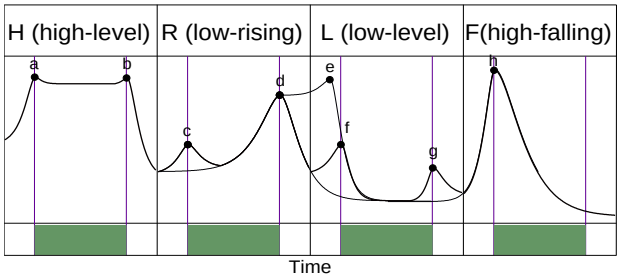

Fig. 2. Modeling the lexical tones. Shading bar indicates perceptually relevant portion of a tone contour.

Basically, two *basic pattern*s are assigned to each tone type for regular formulation in a parametric form except for **F**; no the 2nd *basic pattern* for **F**. Also, for **R** and **L**, *basic pattern*s with the *peak* marked by 'c', 'e', 'f' or 'g' may be absent in certain contexts. Each *basic pattern* is represented in the parametric form:

$$i\text{th } \textit{basic pattern} \Leftrightarrow \{\Delta t_{r_i}, \Delta\lambda_{r_i}, \Delta\lambda_{p_i}, \Delta t_{f_i}, \Delta\lambda_{f_i}\}, \quad (7)$$

where $\Delta\lambda_{p_i}$ indicates a deviation from the reference *peak* value $\hat{\lambda}_{p_i}$, i.e., $\Delta\lambda_{p_i} = \lambda_{p_i} - \hat{\lambda}_{p_i}$. $\hat{\lambda}_{p_i}$ is further represented by the linear equation when t_{p_i} is determined:

$$\hat{\lambda}_{p_i} = \lambda_0 + k * t_{p_i}, \quad i = 1, ..., m, \quad (8)$$

where m denotes number of *pattern*s for a tone group. Parameters λ_0 (*intercept*) and k (*gradient*) are decided so that the line fit to *patterns* of **H** tone combination sequence. Moreover, higher-order linguistic effects (focal effects) and para- and non-linguistic effects can be incorporated into tone-sandhi patterns by properly adjusting these parameters.

2.3. Formulating tone-sandhi rules

As a tonal language, the polysyllabic tone-sandhi contouremes (two, three, and four syllable combinations forming intrinsic tone-sandhi patterns) play an important role in discriminating word meaning in SC. Consequently, to a certain extent, they maintain well-formed invariant patterns. Here, formulation of basic tone-sandhi rules is aimed at to capture various intrinsic tonal variations in tri-tone context and additional anticipatory raising effects. Particularly, when a tri-tone sequence followed by a low-onset tone, **R** or **L**, the mid and back portions of its F_0 contour are clearly raised when compared to those followed by a high-onset tone, **H** or **F**. This is shown in our quantitative analysis of tetra-syllable words.

Based on the lexical tone modeling shown in Fig.2, the basic tone-sandhi rules were summarized as 19 bi- and 198 tri-tone-sandhi contouremes in the parametric form expressed in Eq.(7). These contouremes were determined through inspecting 19 bi-, 59 tri- and 221 tetra-tone-sandhi patterns. These patterns were obtained by a quantitative analysis of 84 di-, 538 tri- and 938 tetra-syllable words uttered by an announcer (Set-1 in Table 1). Furthermore, each contoureme can realize contours in 3 different ranges: normal (Type-A), depressed (Type-B) and expanded (Type-C). The range effect on tone-sandhi patterns was simultaneously incorporated into their parameter search process by the frequency scale transposition [3]. Fig.3 shows examples of the formulated tone-sandhi patterns, where $\lambda_0 = 1.1636$ and k=0.10 giving the *peak* reference-line.

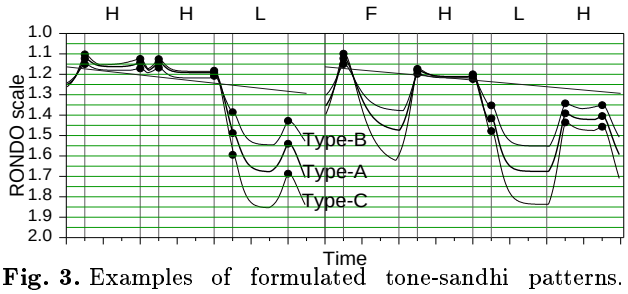

Fig. 3. Examples of formulated tone-sandhi patterns. Each pattern shows contours in 3 different ranges. Solid circles indicate *peak*s. Tilted straight-lines indicate the *peak* reference-lines.

As for the parameter search for contouremes, parameters λ_0 and k were first decided so that the line fit to the **H-HH** and **HHHH** patterns. $\Delta\lambda_{p_i}$ was then calculated as average of selected samples' $\Delta\lambda_{p_i}$. The parameters $\{\Delta t_{r_i}, \Delta\lambda_{r_i}\}$ and $\{\Delta t_{f_i}, \Delta\lambda_{f_i}\}$ independently were estimated from selected samples using Analysis-by-Synthesis (AbS) technique. It should be noted that parameters $\{\Delta t_{r_i}, \Delta\lambda_{r_i}\}$ and $\{\Delta t_{f_i}, \Delta\lambda_{f_i}\}$ should be determined so that tonal discriminative features colud also be properly kept in synthesized phrasal F_0 contours, where the contouremes needed to be adjusted to an assigned *peak* reference-line. For instance, for **RL** sequence shown in Fig.2, parameters $\{\Delta t_{f_2}, \Delta\lambda_{f_2}\}$ values for the 2nd *basic pattern* of **R** should not to change the following **L** feature (low-level) when either 'e' or 'f' *pattern*s are absent in certain contexts.

2.4. Calculating phrasal model parameters taking focus conditions into acount

A focus may be either intended or intrinsic one. Intended focus is controllable by the speaker: placing a focus on important words. Intrinsic focus is decided by the style and structure of the sentence. The focus is utilized to decide range type for each word and to calculate λ_0 and k for the phrasal F_0 contour control.

The F_0 contour type is firstly determined for each word based on the tone sandhi rules. The model parameter values of contouremes, selected using tone sequence and range information, are then adjusted so that F_0 peaks appear along the *peak* reference-line $\hat{\lambda}_{p_i} = \lambda_0 + k * t_{p_i}$, where $i = 1, ..., m$. Particularly, for ith *pattern* with timing t_{p_i}, the *peak* is readjusted by $\lambda_{p_i} = \hat{\lambda}_{p_i} + \Delta\lambda_{p_i}$, while the others remain unchanged. However, the parameters that are responsible for transition between word / phrasal boundaries need to be specially treated. Finally, the sentential F_0 contour generated by these resulting parameter values is transposed from the RONDO-time space into frequency-time space with speaker-specific register alignment.

The basic features relevant to the proposed method are illustrated in Fig.4. This figure shows modeling of an observed F_0 contour of the sentence "nin2 hao3 (hello), huan1 ying0 nin2 shi3 yong4 (welcome you to using) || *wei1 er3 tong1 xin4* (WEIER communication) | *she4 bei0* (devices)" where symbol || indicates the syntactic boundary, and | indicates the boundary of two content words (italic parts). An intended focus is placed to the two content words. The sentence consists of two prosodic phrases separated by a long pause. *Intercept* λ_0 is reset at the onset of each prosodic phrase; λ_0 takes two values 1.17 and 1.09, respectively. Within the 2nd prosodic phrase, content words under the focus take normal range (Type-A), while others take depressed range (Type-B). For an utterance produced in plain statement style like this example, parameter k could be fixed to 0.28 for Type-B parts and to 0.06 for Type-A part.

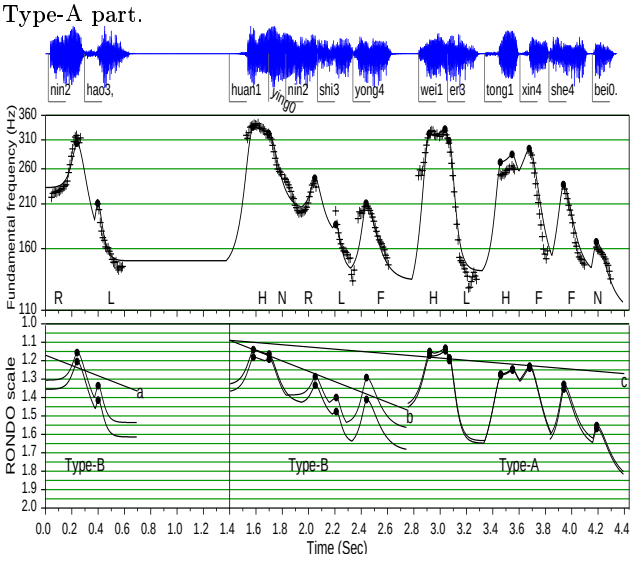

Fig. 4. Modeling of an observed F_0 contour via concatenating formulated tone-sandhi contouremes taking the focus condition into account. In the middle panel, '+' symbols indicate measured F_0's, and solid lines indicate modeled contours, both plotted in log-frequency *vs.* time space. In the bottom panel, the thin and thick curves respectively indicate formulated patterns and those adjusted to the reset *peak* reference-lines a, b and c, which are given by taking focus conditions into account. Solid circles indicate the *peak*s, whose timings are measured directly from the observed F_0 contour.

3. EXPERIMENTAL EVALUATION

The analysis and evaluation experiments were mostly conducted on a total of 1730 speech samples produced by 3 female announcers, as listed in Table 1. These samples were classified into three sub-sets: 1560 isolated words (Set-1), 37 digit-strings (Set-2) and 48 sentences (Set-3). In the AbS process of either model parameter estimation or parameters λ_0 and k search, $[f_{0_b}, f_{0_t}]$ were fixed to values

listed in Table 1. Tone grouping and underlying *peak* timings t_{p_i} were manually determined as well.

Category	Type	Subject	Count	Token	f_{0_b} (Hz)	f_{0_t} (Hz)
Set-1	disyllable	FL	84	1	95	350
word	trisyllable	FL	538	1	95	350
	tetrasyllable	FL	938	1	95	350
Set-2	digit string	WL	37	1	115	410
	digit string	WJ	37	1	110	360
Set-3	Sentence	FL	48	2	95	365

Table 1. Speech samples for the experiments.

3.1. Experiment-1 run on isolated word samples

Experiment on Set-1 was done to make it sure that the formulated tone-sandhi rules can reliably capture the intrinsic tone-sandhi phenomena. Here, only Type-A was used and k was fixed to 0.10. Inspection through comparison of modeled contours with observed ones confirmed the rule's effectiveness. Fig.5 shows examples including the formulated tone-sandhi patterns appeared in Fig.3 and also in Fig.6.

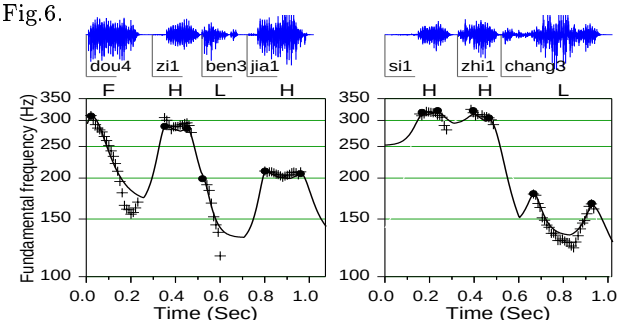

Fig. 5. Representing isolated word F_0 contours ('+' symbols) by tone-sandhi patterns (solid lines).

3.2. Experiment-2 run on syntax-free utterances

The 2nd experiment was run on Set-2, 37 digit-strings (from 3 to 7 syllables) uttered by two announcers. It was found that (a) only Type-A sufficed for representing typical F_0 variations, and (b) parameters λ_0 and k could be fixed; the means of k and λ_0 were 0.0632 (variance 0.00080) and 1.18 (variance 0.00174), respectively.

The experiment confirmed that the tone-sandhi rule formulation was valid also for other speakers analyzed and was robust on timing t_{p_i} shift. Fig.6 shows an example. However, two kinds of mismatch, when model-generated contour was compared to observed one, were found in the experiment; one relating to **H** as shown in Fig.6, another relating to **R** when it locates at the last part of a tone group. More examinations are needed on this issue.

3.3. Experiment-3 run on sentence speech samples uttered in plain statement style

The 3rd experiment was run on Set-3, 48 sentences (from 4 to 18 syllables). Analysis results indicated that (a) only Type-A and Type-B were needed for these samples: Type-A for the focused content words, while Type-B for others, and (b) as for parameters λ_0 and k, although more than one target value were needed, they could be fixed when contexts and focus conditions were fixed, as shown in the example of Fig.4. Generally, for tone-sandhi contouremes with Type-A, k took a value similar to that found in Set-1 and Set-2. However, when comtouremes with Type-B were used, k shifted to a value. It could be quantified to either 0.08 or 0.28.

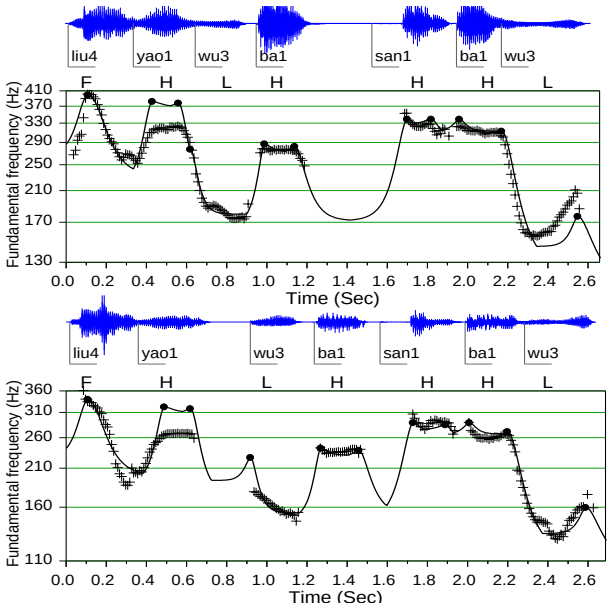

Fig. 6. Modeling F_0 contours by the tone-sandhi rules for the digit-string "6158-385" uttered by speakers WL (top panel) and WJ (bottom panel).

From the three experiments, we could conclude that (1) the tone-sandhi rule formulation is valid regardless of the speakers analyzed, (2) it is reasonable to assume that intended or intrinsic focuses can be utilized to formulate a phrasal F_0 contour (Type-B occurred in Set-3 but not in Set-2. This is probably because Set-3 samples are syntax-constrained but Set-2 samples are syntax-free), and (3) it is convinced that the recalculation, which adjusts tone-sandhi-based model parameters to phrasal F_0 contour contorl, can maintain the tonal discriminative features properly in synthesized F_0 contours.

4. CONCLUSIONS

This paper presents a method for Chinese F_0 contour synthesis. Experiments show that the proposed method is viable, though further work is needed including; construction of rules for representing relation between the prosodic features and higher-order linguistic (and para-/non-linguistic) information, and testing in text-to-speech synthesis.

REFERENCES

[1] 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.

[2] J. Ni and K. Hirose, "A Study on Quantitative Modeling of Sentence Fundamental Frequency Contours in Standard Chinese", *Proc. of 1999 Japan-China Symposium on Advanced Information Technology*, pp.39-46, Tokyo, 1999.

[3] J. Ni and K. Hirose, "Experimental Evaluation of a Functional Modeling of Fundamental Frequency Contours of Standard Chinese Sentences," to appear in *ISCSLP2000*, Beijing, Oct., 2000.