CHAPTER 2D

TONE MODELING FOR SPEECH SYNTHESIS

[1]Sin-Horng Chen, [2]Chiu-yu Tseng, and [3]Hsin-min Wang

[1]*Department of Communication Engineering, National Chiao Tung University*
[2]*Institute of Linguistics and* [3]*Institute of Information Sciences, Academia Sinca, Taipei*
E-mail: schen@mail.nctu.edu.tw,cytling@sinica.edu.tw,whm@iis.sinica.edu.tw

Tone modeling for speech synthesis aims at providing proper pitch, duration, and energy information to generate natural synthetic speech from input text. As the speech processing technology progresses rapidly in recent years, some advanced tone modeling techniques for MTTS are proposed. In this chapter, two modern tone modeling approaches for Mandarin speech synthesis are discussed in detail.

1. Introduction

Prosody is an inherent supra-segmental feature of human speech. It carries stress, intonation patterns and timing structures of continuous speech which, in turn, decide the naturalness and understandability of an utterance. For speech synthesis, a general prosody modeling approach is to build a model, in the training phrase, to describe the relationship between the hierarchical linguistic structure of Chinese text and the hierarchical prosody structure of the corresponding Mandarin speech; and in the test phase to first use the model to map from the hierarchical linguistic structure extracted from the input text to the hierarchical prosody structure, and to then generate prosodic features from the prosody structure. In this approach, linguistic features are regarded as affecting factors that control the variations of prosodic features and are organized into different levels by first using the hierarchical linguistic structure and then mapping to the hierarchical prosody structure. The use of hierarchical linguistic structure is owing to the fact that it is the well-known and conventional way to analyze text and there exist many well-developed techniques and tools, such as lexicon, word and POS tagger, parser and so on. A hierarchical linguistic structure can be composed of various levels including character, lexical word, word chunk, phrase, clause, sentence, paragraph and so on. The use of hierarchical prosody structure is to correct the inappropriateness of directly using hierarchical linguistic

structure to control the generation of prosodic features. A hierarchical prosody structure may contain the following levels[2]: syllable, prosodic word, intermediate phrase, intonational phrase/breath group, prosodic phrase group and so on.

There are three main concerns in prosody modeling for Mandarin text-to-speech (MTTS). One is the hierarchical linguistic structure of Chinese text that describes the relationship among linguistic constituents of different levels. Currently, the syntax of sentence is a generally accepted hierarchical linguistic structure. But some other affecting factors, such as semantic and emotional information, and higher-level factors, such as discourse[2], are also needed to be considered. Another concern is the mapping from the hierarchical linguistic structure to the hierarchical prosody structure. This is the major focus of prosody modeling for speech synthesis in recent years. Prosodic phrasing and break labeling are two related problems in this field.[14,22] The other concern is the generation of prosodic features from the hierarchical prosody structure. Currently, the most popular approach is to superimpose patterns of different hierarchical level[11]. The pattern of each level can be obtained by simply assigning a deterministic average pattern extracted from a speech database or by using a linear/nonlinear regression method to combine affections of various affecting factors.

In the early stage of MTTS study, prosody modeling is performed using relatively simple linguistic structures and prosodic structures.[18] Some lower level linguistic features, say syllable and word, are used. A prevalent approach is to find rules to map from lower-level contextual features, extracted from phonetic structure of syllable, tone and word, to syllable/word-level prosodic features including pitch contour pattern, energy level pattern, initial/final or syllable duration pattern, and inter-syllable pause duration. The estimated prosodic feature patterns are lastly superimposed with a sentence-level intonation pitch pattern selected from a pattern pool or assigned by rules. The resulting synthetic speech is usually far away from high naturalness.

As the speech processing technology progresses rapidly in recent years, sophisticated linguistic structures and prosodic structures are available now. Some advanced prosody modeling techniques for MTTS are therefore proposed.[9,11,19-21] In the following, we discuss two statistical prosody modeling methods in detail.

2. A Five-layer Tone Modeling Method for MTTS

In Chapter 2b, it was demonstrated that phrase grouping is essential for characterizing the prosody of fluent Mandarin speech[1]. Fig. 1 shows the hierarchical organization framework for multiple phrase grouping. From the top down, the layered nodes are: phrase groups (PG), breath groups (BG), prosodic

phrases (PPh), prosodic words (PW), and syllables (SYL). These constituents are associated with the break indices B5 to B1, respectively.[2]

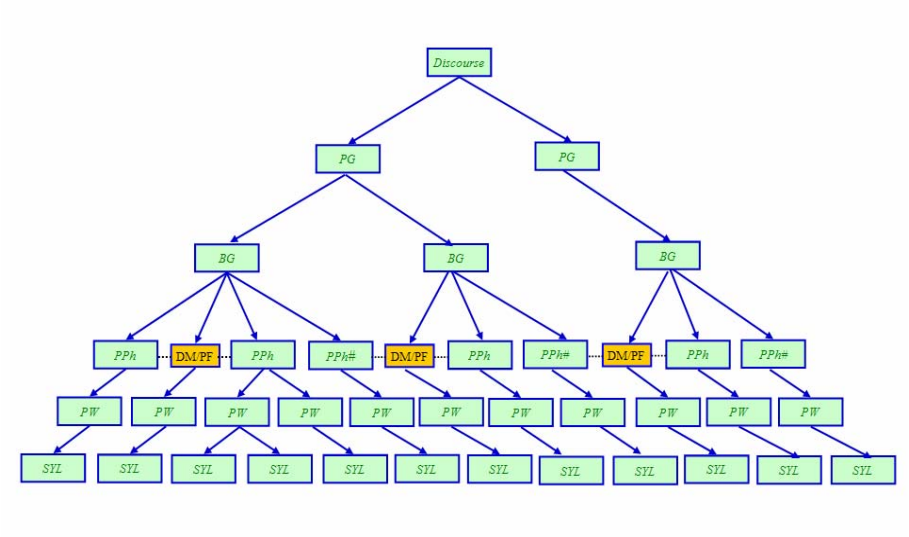


Fig.1. A schematic representation of the hierarchical organization for multiple phrase grouping of perceived units and boundaries.

Evidence of prosodic phrase grouping has been found in both the adjustments of F_0 contours and the temporal allocations within and across phrases. Thus, F_0, duration, and intensity should be considered simultaneously when modeling tone behavior in Mandarin speech. In this section, we discuss tone modeling for Mandarin speech synthesis and the development of a Mandarin TTS system that integrates the prosody processing modules, such as duration modeling, F_0 modeling, intensity modeling, and break predictions.

2.1. *Duration Modeling*

In Tseng[3], the analysis of rhythmic patterns in Mandarin speech reveals that syllable duration is not only affected by the syllable's constitution, but also by the prosodic structures of the upper layers, namely PW, PPh, BG, and PG. These factors allow us to design a layered model for syllable duration.

The analysis was conducted on a corpus of female read speech of 26 long paragraphs or discourses in text. The corpus consisted of 11,592 syllables in total. Initially, the speech data was aligned automatically with initial and final phones using the HTK toolkit, and then labeled manually by trained transcribers to indicate the perceived prosodic boundaries or break indices (BI).

2.1.1 *Intrinsic Statistics of Syllable Duration*

A layered model is used to estimate a syllable's duration. At the SYL-layer, the following linear model is adopted:

$$\begin{aligned} Syllable\ intrinsic\ duration = const_d + CTy + VTy + Ton + PCTy + PVTy \\ + PTon + FCTy + FVTy + FTon \\ + \ 2\text{-}way\ factors\ of\ the\ above\ factor \\ + 3\text{-}way\ factors\ of\ the\ above\ factor, \end{aligned} \quad (1)$$

where $const_d$ is a reference value, which is dependent on the corpus; *CTy*, *VTy*, and *Ton* represent the offset values associated with the consonant type, vowel type, and tone of the current syllable, respectively; the prefixes *P* and *F* represent the corresponding factors of the preceding and following syllable, respectively; the 2-way factors consider the joint effect of two single-type factors; and the 3-way factors consider the joint effect of three single-type factors. There are C_2^9 (=36) 2-way factors in total. The 3-way factors that have a negligible influence on a syllable's duration are not considered. Therefore, only three 3-way factors are considered, namely, the combination of consonant types, the vowel types, and the tones of the preceding, current, and following syllables. As a result, there are 49 factors in total. As reported in Tseng[3], the SYL-layer model can explain about 60% of syllable duration.

2.1.2 *The Effect of the Layered Prosodic Structure*

As shown in Fig. 2, a syllable's duration is affected by its position within a PW. Note that the final syllable in the PW tends to be longer than the other syllables.

$$DurS = Syllable's\ intrinsic\ duration + f_{PW}(PW\ length, position\ in\ PW) \quad (2)$$

The PW-layer speeds up the rhythm by subtracting a value derived from Fig. 2.

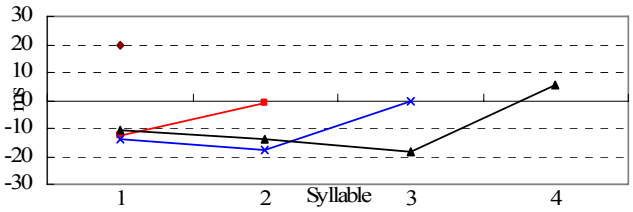


Fig.2: Rhythmic patterns in the PW-layer

The PPh-layer affects a syllable's duration in a similar way to the PW-layer. In the BG-layer and above, the length of the prosodic unit gets longer and more complicated, but the perceived significance only exists in the initial and final PPh units. Therefore, we model the BG-layer's effect as the effect of the initial and final PPhs in that layer. The overall model is thus formulated as:

$$\begin{aligned} DurS = {} & Syllable's\ intrinsic\ duration + f_{PW}(PW\ length, position\ in\ PW) \\ & + f_{PPh}(PPh\ length, position\ in\ PPh) \\ & + f_{IFPPh}(Initial/Final\ PPh\ length, position\ in\ PPh), \end{aligned} \tag{3}$$

where *DurS* denotes the modeled syllable's duration; and f_{PW} (·), f_{PPh} (·), and f_{IFPPh} (·) denote the portions of the syllable's duration affected by the function of the length of PW, PPh, and initial or final PPh in PG, respectively, together with the target syllable's position within them.

2.2 F_0 Modeling

Many F_0 models of sentence/phrasal intonation are proposed in the literature. We use the well-known Fujisaki model as the production model for F_0.[4] The model connects the movements of the cricoid cartilage to the measurements of F_0 and is thus based on the constraints of human physiology. Therefore, it is reasonable to assume that the model can accommodate F_0 output in different languages. Successful applications of the model on many language platforms, including Mandarin, have been reported.[5,6]

In the case of Mandarin Chinese, phrase commands are used to produce intonation at the phrase level, while accent commands are used to predict lexical tones at the syllable level.[7] Phrasal intonations are superimposed on sequences of lexical tones. Therefore, interaction between the two layers causes modifications of the F_0 during production of the final output. The superimposing of a higher level onto a lower level leaves room for even higher levels of F_0 specification to be superimposed and built. Thus, we can implement the hierarchical organization framework of phrase/intonation-grouping in the Fujisaki model by adding a PG layer over phrases.[8,9] The F_0 patterns of phrase grouping can hence be derived.

2.2.1 *Building the Phrasal Intonation Model*

A linear model for the phrase command of the Fujisaki model is adopted as follows:

$$\begin{aligned} Phrase\ command\ Ap = {} & const_{Ap} + coeff1 \times pause \\ & + coeff2 \times pre\_phr + coeff3 \times f_0 min \\ & + f_{PPh}(Phrase\ command\ position\ in\ PPh) \\ & + f_{IFPPh}(Initial/Medial/Final\ PPh). \end{aligned} \tag{4}$$

where $const_{Ap}$ is a reference value, which is dependent on the corpus; *pause* is the preceding speechless portion associated with the current phrase command; *pre_phr* is the accumulated phrase command response of previous phrase commands as the response of the current phrase command reaches its peak; $f_0 min$ is the minimum fundamental frequency of the utterance; f_{PPh} (·) reflects the

position in PPh that the related phrase command is located; and f_{IFPPh} is for the PG intonation which only has a significant effect on the first and last PPh units.

Fig. 3 shows a comparison between the F_0 prediction/production of a PG and the original intonation.

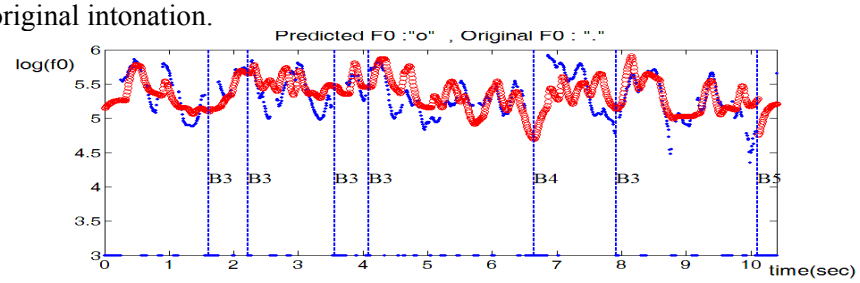


Fig. 3. The simulation result of global intonation modeling of a PG. The thin line composed of dots represents the original F0 contour, while the thick line composed of circles represents the predicted contour

2.3 *Intensity Modeling*

Segmental RMS (root mean square) values are first derived using the ESPS toolkit. For each initial and final phone in a syllable, the averaged RMS value is calculated by using 10 equally spaced frames in the target segment time span. To eliminate the difference in levels between paragraphs, possibly caused by slight changes during recording, the RMS values within each paragraph need to be normalized to NRMS (normalized RMS) values. Intensity modeling is much the same as duration modeling[10]:

$$\begin{aligned} IntS &= Syllable's\ intrinsic\ intensity \\ &+ f_{PW}(PW\ length, position\ in\ PW) \\ &+ f_{PPh}(PPh\ length, position\ in\ PPh) \\ &+ f_{IFPPh}(Initial/Final\ PPh\ length, position\ in\ PPh), \end{aligned} \tag{5}$$

where $f_{PW}(\cdot)$, $f_{PPh}(\cdot)$, and $f_{IFPPh}(\cdot)$ denote the portions of syllable intensity affected by the function of the length of PW, PPh, and initial or final PPh in PG, respectively, together with the target syllable's position within them. The *Syllable's intrinsic intensity* is modeled by:

$$\begin{aligned} &Syllable's\ intrinsic\ intensity \\ &= const_i + CTy + VTy + Ton + PCTy + PVTy + PTon \\ &+ FCTy + FVTy + FTon + 2\text{-}way\ factors\ of\ the\ above\ factor \end{aligned} \tag{6}$$

2.4 *The TTS System*

The above duration, F_0, and intensity modeling methods are not only useful for analyzing prosodic patterns in Mandarin speech, but can also be used to predict

prosodic parameters for synthesizing speech according to text input. Given a large annotated speech database, the predicted duration, F_0, and intensity parameters can be used to select appropriate units for direct concatenation or to minimize the signal processing requirement. Many research works on unit selection have been published. As a large annotated speech database is usually infeasible, instead of reviewing existing methods, we present a promising way to rapidly adapt a TTS system to new voices by applying the above statistical analysis and modeling framework.

Since the Chinese writing system consists of mono-syllabic logographic characters and there are only 1,292 distinct tonal syllables, it is reasonable to choose syllables as concatenative units. The duration, F_0, and intensity models described above are based on the PG structure. Therefore, we need a specially designed database so that the TTS system can be implemented to use these models[11]. The time-domain pitch-synchronous overlap-add (TD-PSOLA)[12] method is used to perform prosody modification in the TTS system.

2.4.1 *Speech Database*

The database comprises 1,292*3 Mandarin tonal syllable tokens. Each of the 1,292 syllables is embedded in a phrase of a three-phrase carrier sentence (i.e., a PG of 3 PPhs) in the initial, medial, and final positions, respectively. The speech data was recorded by a native female speaker in a sound-proof room. So, for each syllable, three tokens are collected.

2.4.2 *Duration Adjustment*

Since the TTS database is from a different speaker, the absolute duration predicted by the duration model should be adjusted, while the rhythmic patterns in the PG organization should be kept. Because the initial, medial, and final syllables were originally collected from the same positions of the PG, their duration should not be changed. The duration of the remaining syllables, which were originally the first syllable of a PW at the medial position of a medial PPh of a 3-PPh PG, should be modified to satisfy the rhythmic pattern in PG organization. In this way, to synthesize a PG of m characters (or syllables), the duration of the i-th syllable is given by:

$$DurS_i^* = \begin{cases} OriDur(S_i), & i = 1, m/2, m \\ OriDur(S_i) - DF_i, & 1 < i < m/2, m/2 < i < m, \end{cases} \tag{7}$$

where $OriDur(S_i)$ is the corresponding syllable-token's original duration, and DF_i is an offset factor calculated by:

$$\begin{aligned} DF_i = M^d_{TC} / M^d_{MC} \times [& f_{PW}(PW\ length, position\ in\ PW) - f_{PW}(2,1) \\ & + f_{PPh}(PPh\ length, position\ in\ PPh) - f_{PPh}(11,6) \\ & + f_{IFPPh}(Initial/Final\ PPh\ length, position\ in\ PPh)], \end{aligned} \tag{8}$$

where M^d_{TC} and M^d_{MC} are, respectively, the mean of syllable duration in the TTS corpus and the training corpus; and $f_{PW}(\cdot)$, $f_{PPh}(\cdot)$, *and* $f_{IFPPh}(\cdot)$ are same as those in Eq. (3), which are estimated from the training corpus.

2.4.3 *F_0 Adjustment*

In the implementation of F_0 adjustment, the comparison is confined to the first F_0 peak of the predicted PG intonation and the average F_0 of the first syllable from the carrier sentence. The phrase control mechanism for phrase components in the Fujisaki model is defined as[4]:

$$G_p(t) = \begin{cases} \alpha^2 t \times \exp(-\alpha t), & \text{for } t \geq 0 \\ 0, & \text{for } t < 0. \end{cases} \tag{9}$$

In Eq. (9), the time required to reach the maximum is 1/α. Therefore, the maximum value of the phrase response $Ap \times G_p(t)$ is:

$$P = Ap \times \alpha \times \exp(-1). \tag{10}$$

From Eq. (10), it is clear that P is proportional to Ap when α remains a constant.

We can estimate the adjustment value ΔAp of Ap according to the difference between the average F_0, denoted as P_c, of the first syllable from the carrier sentence and the first F_0 peak, denoted as P_p, of the predicted PG intonation:

$$\Delta Ap = (P_c - P_p) \times \exp \times \alpha^{-1}. \tag{11}$$

Then, every predicted phrase command must be adjusted according to ΔAp. Note that the adjustment does not change the shape of the intonation, but the level moves closer to that of the carrier sentence database.

2.4.4 *Intensity Adjustment*

Intensity adjustment is realized in the same way as duration adjustment. If m syllables need to be synthesized, the intensity of the i-th syllable is given by

$$IntS_i^* = \begin{cases} OriInt(S_i), & i = 1, m/2, m \\ OriInt(S_i) - IF_i, & 1 < i < m/2, m/2 < i < m, \end{cases} \tag{12}$$

where $OriInt(S_i)$ is the corresponding syllable-token's original intensity and IF_i is an offset factor calculated by

$$\begin{aligned} IF_i = {} & M^i_{TC} / M^i_{MC} \times [f_{PW}(PW\ \ length, position\ in\ PW) - f_{PW}(2,1) \\ & + f_{PPh}(PPh\ length, position\ in\ PPh) - f_{PPh}(11,6) \\ & + f_{IFPPh}(Initial/Final\ PPh\ length, position\ in\ PPh)], \end{aligned} \tag{13}$$

where M^i_{TC} and M^i_{MC} are, respectively, the mean of the syllable intensity in the TTS corpus and the training corpus; and $f_{PW}(\cdot)$, $f_{PPh}(\cdot)$, and $f_{IFPPh}(\cdot)$ are same as those in Eq. (5), which are calculated from the training corpus.

2.4.5 *Break Prediction*

Prosodic boundaries and break indices are predicted by analyzing the syntactic structure of the text to be synthesized. Basically, the break indices can be predicted according to the punctuation. For a long PPh, we can insert an extra B3 to segment the PPh into two PPh units. PW is a fundamental prosodic unit, while the lexical word (LW) is a basic syntactic unit in the syntactic structure. Therefore, PW prediction is the first step towards building a prosody model from a piece of text. According to Chen[14], only 67.5% of PWs and LWs are coincident in prosodic structure tagged corpora. The accuracy of predicting PWs by grouping LWs using statistical approaches is approximately 90%.

2.4.6 *System Flowchart*

Given a piece of text, the prosodic boundaries and break indices are predicted based on an analysis of the syntactic structure. The PG hierarchical structure and the pronunciation (the syllable sequence associated with the text) are also generated. Then, the duration and intensity of all syllables are assigned by the duration model and the intensity model, respectively, while the F_0 contours of all phrases are generated by the intonation model. The output of text processing is stored in a predefined XML document. Finally, the TD-PSOLA method is used to perform prosody modification, and the TTS system outputs the concatenated waveform.

2.5 *Discussion and Conclusions*

The TTS system introduced in this section attempts to synthesize fluent speech in long paragraphs based on a specially-designed moderate syllable-token database. It is believed that an integrated prosodic model that organizes phrase groups into related prosodic units to form speech paragraphs would significantly improve the naturalness of the output of an unlimited TTS system. How mono-syllables can be collected to provide further prosodic information has been shown.

3. A Tone Modeling Approach Using Unlabeled Speech Corpus

Traditionally, prosody modeling is conducted using well-annotated speech corpora with all prosodic phrase boundaries and break indices being properly labeled in advance. Usually, this is done manually. But it is a labor-intensive work. Besides, the inconsistency of human labeling is also a problem. So, most corpora used in prosody modeling are not large. Some alternative prosody modeling studies for syllable duration and pitch contour of Mandarin speech using unlabeled speech corpora were performed recently.[19-20] In this section, a statistical tone modeling method for Mandarin pitch contour using an unlabeled speech corpus is discussed. The method is an extension of that proposed by Chiang.[21] The basic idea is to use a statistical model to consider some major affecting factors that control the variation of the syllable pitch contour. By this way, the relationship between the observed values of pitch contour patterns in the speech corpus and its major affecting factors can therefore be built automatically.

3.1 *Review of previous works*

Two prosody modeling studies for Mandarin speech using unlabeled speech corpora were proposed recently.[19-21] One is for syllable duration[19] and another is for syllable pitch contour[20]. We briefly review them in the following subsections.

3.1.1 *A Statistical Syllable Duration Model*[19]

The syllable duration model is designed based on the idea of taking each affecting factor as a multiplicative companding factor (CF) to control the compression and stretch of syllable duration. Five major affecting factors including tone, base-syllable, speaker-level speaking rate, utterance-level speaking rate, and prosodic state are considered. Prosodic state is conceptually regarded as the state in a prosodic phrase. The model is expressed by

$$Z_n = X_n \gamma_{t_n} \gamma_{y_n} \gamma_{j_n} \gamma_{l_n} \gamma_{s_n}, \tag{14}$$

where Z_n and X_n are the observed and normalized durations of the *n*-th syllable; γ_p is the CF of the affecting factor *p*; t_n , y_n , j_n , l_n and s_n represent respectively the lexical tone, prosodic state, base-syllable, utterance, and speaker of the *n*-th syllable; and X_n is modeled as a normal distribution with mean μ and variance *v*. The model further considers the three Tone 3 patterns[15,16] of falling-rising, middle-rising and low-falling to increase the number of tones to 7. The model is trained by an EM algorithm with prosodic state being treated as hidden.

The model is validated by using a speech corpus containing paragraphic utterances of 5 speakers. The following conclusions were obtained:

- The variance of syllable duration reduces significantly as the influences of these five affecting factors are eliminated.
- The influences of 7 tones and 411 base-syllables can be directly obtained from their CFs.
- The prosodic state is linguistically meaningful.
- The labeling of the three Tone 3 patterns looks good.
- The CFs of utterance show that long texts are always pronounced fast while short texts are pronounced in arbitrary speed.
- Both initial and final durations are propositional to the syllable duration except when the syllable is largely shortened or lengthened. In the two extreme cases, final is usually compressed or stretched more seriously.

3.1.2 *A Statistical Syllable Pitch Contour Model*[20]

The syllable pitch contour model is built based on the same principle of syllable duration modeling. The mean and shape of syllable pitch contour are separately modeled by considering different set of affecting factors. For pitch mean, affecting factors considered include the tones of the previous, current, and following syllables; the initial and final classes of the current syllable; the prosodic state of the current syllable; and the speaker's level shift and dynamic range scaling factors. For pitch shape, affecting factors considered include lexical tone combinations[15,16], the initial and final classes of the current syllable, the prosodic state for the effects of high-level linguistic features, and the pitch level shifting effect of speakers.

The same 5-speaker speech corpus is used to evaluate the pitch mean and shape models. The following conclusions were obtained:

- The variances of pitch mean and shape parameters reduce significantly as the influences of their respective affecting factors are eliminated.
- Many tone sandhi rules, including the famous 3-3 tone sandhi rule, can be observed from the CFs of tone combinations.
- The prosodic state is linguistically meaningful.
- A change of the prosodic state index, from large to small, indicates a possible phrase boundary. An effective rule-based method to detect minor and major prosodic phrase boundaries is therefore proposed.

3.2 F_0 Modeling

The proposed syllable pitch contour model considers the following three major affecting factors: lexical tone, prosodic state and inter-syllable coarticulation. Here, prosodic state is used to account for the influences of all high-level

linguistic features and can be conceptually regarded as the state of the current syllable staying in a prosodic phrase.

3.2.1 *The Proposed Syllable Pitch Contour Model*

The model is formulated based on the assumption that all affecting factors are combined additively and can be expressed by

$$\mathbf{x}_{k,n} = \mathbf{y}_{k,n} + \boldsymbol{\chi}_{t_{k,n}} + \boldsymbol{\chi}_{p_{k,n}} + \boldsymbol{\chi}^{f}_{c_{k,n-1},tp_{k,n-1}} + \boldsymbol{\chi}^{b}_{c_{k,n},tp_{k,n}} \tag{15}$$

where $\mathbf{x}_{k,n}$ and $\mathbf{y}_{k,n}$ are vectors of four orthogonal expansion coefficients representing, respectively, the observed and normalized pitch contours of the n-th syllable in utterance k; $\boldsymbol{\chi}_{t_{k,n}}$ is the affecting pattern of the current tone $t_{k,n} \in \{1,2,3,4,5\}$; $\boldsymbol{\chi}_{p_{k,n}}$ is the affecting pattern of prosodic state $p_{k,n} \in \{0,1,2,\cdots,P\}$; $c_{k,n} \in \{0,1,2,\cdots,C,C+1\}$ is the coarticulation state of the inter-syllable location between syllables n and $n+1$ with $c_{k,-1} = 0$ and $c_{k,N_k} = C+1$ representing the states of utterance beginning and ending, respectively; $tp_{k,n} \in \{(1,1),(1,2),\cdots,(5,5)\}$ is the tone pair $(t_{k,n}, t_{k,n+1})$; $\boldsymbol{\chi}^{f}_{c_{k,n-1},tp_{k,n-1}}$ is the forward affecting pattern of the tone pair $tp_{k,n-1}$ with coarticulation state $c_{k,n-1}$; $\boldsymbol{\chi}^{b}_{c_{k,n},tp_{k,n}}$ is the backward affecting pattern of the tone pair $tp_{k,n}$ with coarticulation state $c_{k,n}$; We note here that $\boldsymbol{\chi}^{f}_{c_{k,-1},tp_{k,n-1}} = \boldsymbol{\chi}^{f}_{0,t_{k,1}}$ and $\boldsymbol{\chi}^{b}_{c_{k,N_k},tp_{k,N_k}} = \boldsymbol{\chi}^{b}_{C+1,t_{k,N_k}}$. Notice that we directly assign the prosodic state $p_{k,n} = 0$ for those syllables whose F0 can not be detected. Fig. 4 displays the relationship of syllable pitch contours and these affecting factors.

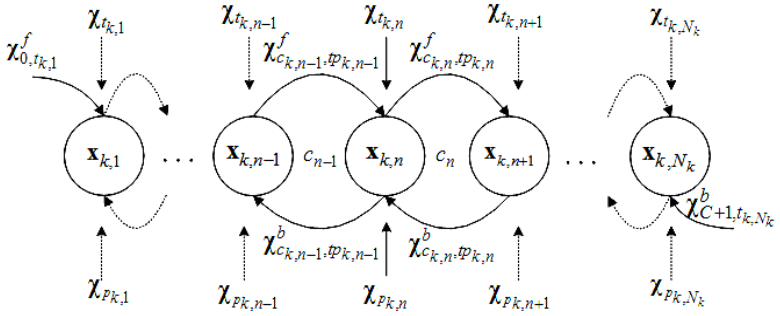


Fig. 4. The relationship of syllable pitch contours and affecting factors used.

The normalized pitch shape $\mathbf{y}_{k,n}$ is modeled as a Gaussian distribution $N(\mathbf{y}_{k,n};\boldsymbol{\mu},\mathbf{R})$, or equivalently $\mathbf{x}_{k,n}$ is modeled by

$$N(\mathbf{x}_{k,n};\boldsymbol{\mu}+\boldsymbol{\chi}_{t_{k,n}}+\boldsymbol{\chi}_{p_{k,n}}+\boldsymbol{\chi}^{f}_{c_{k,n-1},tp_{k,n-1}}+\boldsymbol{\chi}^{b}_{c_{k,n},tp_{k,n}},\mathbf{R}) \tag{16}$$

Here, both the prosodic state, representing the prosodic feature variation in a prosodic phrase, and the coarticulation state, representing the degree of coupling between two consecutive syllables, are treated as hidden. To help determining them, two additional probabilistic models are introduced. One is the coarticulation state model $P(\mathbf{i}_{k,n}\,|\,c_{k,n})$ which describes the relationship of the coarticulation state $c_{k,n}$ and a set of acoustic/linguistic features $\mathbf{i}_{k,n}$ extracted from the vicinity of the inter-syllable location following syllable *n*. Another is the prosodic state model $P(s_{k,n}\,|\,p_{k,n})$ which describes the relationship of the prosodic state $p_{k,n}$ and a set of syntactic features $s_{k,n}$ extracted from the syntactic tree of the sentence containing syllable *n*.

In this work, the model of $c_{k,n}$ involves four features and is expressed by

$$P(\mathbf{i}_{k,n}\,|\,c_{k,n})=P(PD_{k,n}\,|\,c_{k,n})P(PM_{k,n}\,|\,c_{k,n})P(IW_{k,n}\,|\,c_{k,n})P(IT_{k,n}\,|\,c_{k,n}) \tag{17}$$

where $\mathbf{i}_{k,n}=(PD_{k,n},PM_{k,n},IW_{k,n},IT_{k,n})$; $PD_{k,n}$ and $PM_{k,n}$ are, respectively, the pause duration and punctuation mark following syllable *n*; $IW_{k,n}$ indicates whether the inter-syllable location between syllables *n* and *n*+1 is an inter-word or intra-word; and $IT_{k,n}$ is the general type of consonant of the *n*+1 syllable.

The prosodic state model describes the relationship of $p_{k,n}$ and some features representing the role of the current syllable *n* in the syntactic tree[17]. In this study, 31 syntactic features determined based on the contextual information of the syllable are chosen. They are categorized according to the position of the current syllable in a word: beginning-of-word (BW), within-word (WW), ending-of-word (EW), and single-syllable-word (SW). They are listed in Table 1. The model is then expressed by

$$P(\mathbf{s}_{k,n}\,|\,p_{k,n})=P(\mathbf{s}_{k,n}=sr_i\,|\,p_{k,n}) \tag{18}$$

where sr_i is a syntactic role of the current syllable.

Table 1: The syntactic roles used in the modeling of $p_{k,n}$

position in a word	• within-word (WW) • beginning-of-word (BW) • end-of-word (EW) • single-syllable-word (SW)
type of the preceding phrase at the same level in the tree	• single-syllable- word (PSW) • 2 or 3-syllable word (PW23) • 4 or more-syllable word (PW4) • phrase boundary without PM (PPB) • phrase boundary with PM (PPBPM)
type of the following phrase at the same level in the tree	• single-syllable- word (FSW) • 2 or 3-syllable word (FW23) • 4 or more-syllable word (FW4) • phrase boundary without PM (FPB) • phrase boundary with PM (FPBPM)
sr_i	(PSW| PW23| PW4| PPB| PPBPM)_BW 5 combinations
	EW_(FSW| FW23| FW4| FPB| FPBPM) 5 combinations
	(PSW| PW23| PW4| PPB| PPBPM)_SW _(FSW| FW23| FW4| FPB| FPBPM) 25 combinations
	WW 1 combination

3.2.2 *Experimental Results*

Performance of the proposed pitch modeling method was evaluated using a Mandarin speech database. The database contained the read speech of a single female professional announcer. Its texts were all short paragraphs composed of several sentences selected from the Sinica Tree-Bank Corpus[17]. All sentences of the Tree-Bank corpus were manually parsed to extract their syntactic tree structures. The database consisted of 380 utterances with 52,192 syllables.

In the evaluation, we set the numbers of prosodic states and coarticulation states to be 16 and 8, respectively. After well training, the covariance matrices of the original and normalized syllable pitch feature vectors were

$$\mathbf{R}_x = \begin{bmatrix} 2487 & -64 & -145 & 8 \\ -64 & 373 & 27 & -40 \\ -145 & 27 & 69 & -71 \\ 8 & -40 & -71 & 19 \end{bmatrix} \Rightarrow \mathbf{R}_y = \begin{bmatrix} 35 & -11 & 2 & -1 \\ -11 & 82 & 7 & -5 \\ 2 & 7 & 32 & 1 \\ -1 & -5 & 1 & 11 \end{bmatrix}$$

$$|\mathbf{R}_x| = 7.93 \times 10^8 \qquad |\mathbf{R}_y| = 8.90 \times 10^5$$

The determinant of the covariance matrix of the normalized pitch feature vector was reduced significantly as compared with that of the observed vector.

Fig. 5 shows the affecting patterns and their F0 mean values of 16 prosodic states. Table 2 displays the state transition probabilities. Fig. 6 displays three typical examples. As shown in Fig. 5, States 1, 2, 3 and 4 have low and flat patterns and hence tend to be located at the trail of a prosodic phrase (because of the declination effect of F_0). High probabilities of $P(EW\_FPB \mid p)$ and $P(EW\_FPBPM \mid p)$ for p=2, 3 and 4, observed from the prosodic state model, also confirm that they appear at the ending parts of syntactic phrase and sentence very often. Moreover, high transition probabilities of 2-2, 2-3, 3-2, 3-3, 4-3 and 4-4 observed from the state transition table (Table 2) show that the low and flat trail pattern of prosodic phrase (see Fig.6(c)) is common to appear. On the other hand, States 15, 14 and 12 have high and rising-falling patterns and hence tend to be located at the beginning of a prosodic phrase (to show the reset phenomenon). This finding can be further confirmed by the high probabilities of $P(PPB\_BW \mid p)$ and $P(PPBPM\_BW \mid p)$ for p=15, 14 and 12 which show that they appear at the beginning parts of syntactic phrase and sentence very often. Moreover, high transition probabilities of 15-10, 15-9, 15-13, 14-10, 14-9, 14-7, 12-9 and 12-7 show that the rising-falling reset pattern (see Figs.6(a) and (b)) of prosodic phrase is common to appear. From closely examining some frequently occurred prosodic state pairs or triplets, we find that most of them form prosodic words.

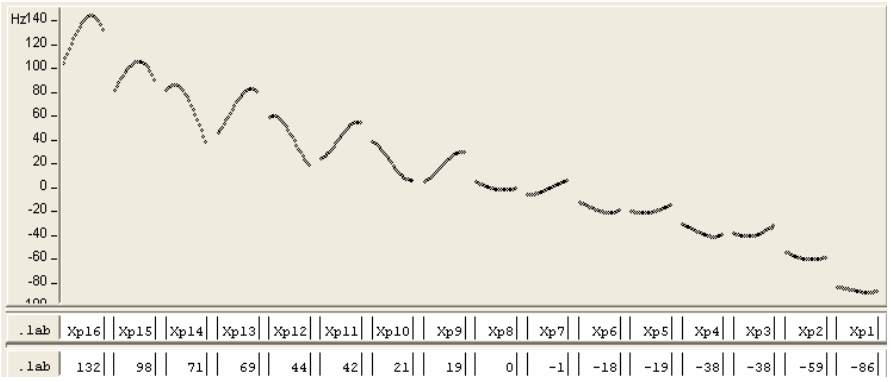
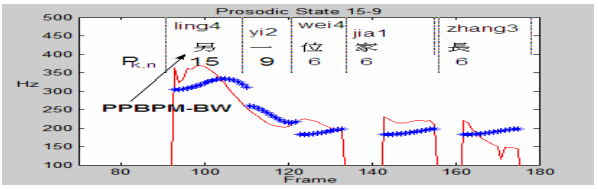


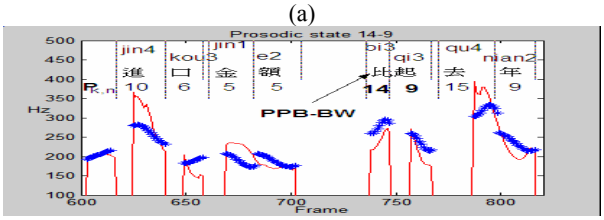
Fig.5: The affecting patterns and their F0 mean values of 16 prosodic states.

Table 2: Prosodic state transition probability $P(p_n \mid p_{n-1})$

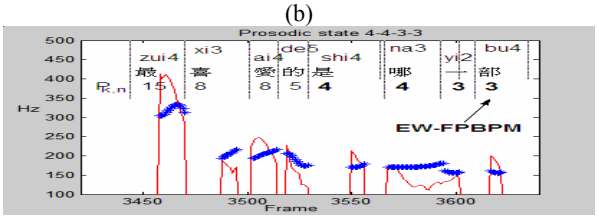
Pn-1\Pn Bigram	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	0.02	**0.68**	0.02	0.02	0.05	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
1	0.08	**0.12**	**0.17**	0.10	0.05	0.10	0.02	**0.12**	0.02	0.10	0.01	0.04	0.01	0.02	0.00	0.01	0.01
2	0.00	0.08	**0.24**	**0.14**	0.09	**0.12**	0.02	0.10	0.02	0.08	0.01	0.05	0.01	0.02	0.01	0.01	0.00
3	0.00	0.04	**0.22**	0.09	**0.22**	0.09	0.10	0.06	0.04	0.03	0.02	0.02	0.02	0.02	0.01	0.00	0.00
4	0.00	0.02	0.09	**0.18**	0.04	**0.20**	0.02	**0.17**	0.01	**0.11**	0.02	0.06	0.02	0.02	0.01	0.01	0.00
5	0.00	0.02	**0.11**	0.07	**0.23**	0.09	**0.19**	0.06	0.08	0.04	0.04	0.03	0.02	0.01	0.01	0.01	0.00
6	0.00	0.01	0.06	**0.17**	0.03	**0.21**	0.02	**0.19**	0.02	0.11	0.02	0.06	0.03	0.03	0.02	0.02	0.01
7	0.00	0.01	0.04	0.04	**0.14**	0.07	**0.21**	0.05	**0.17**	0.05	0.07	0.04	0.04	0.03	0.02	0.01	0.00
8	0.00	0.01	0.04	**0.11**	0.06	**0.19**	0.07	**0.16**	0.06	0.10	0.03	0.07	0.03	0.03	0.02	0.02	0.01
9	0.00	0.01	0.02	0.02	0.06	0.05	**0.15**	0.06	**0.18**	0.07	**0.12**	0.06	0.07	0.05	0.05	0.04	0.01
10	0.00	0.01	0.04	0.07	0.09	**0.14**	**0.12**	**0.12**	0.10	0.09	0.05	0.05	0.04	0.04	0.02	0.02	0.01
11	0.00	0.00	0.02	0.02	0.04	0.04	0.09	0.05	**0.15**	0.05	**0.16**	0.06	**0.11**	0.06	0.07	0.07	0.02
12	0.00	0.01	0.03	0.06	0.08	**0.11**	**0.14**	0.09	**0.14**	0.08	0.09	0.05	0.04	0.04	0.02	0.03	0.01
13	0.00	0.00	0.01	0.01	0.03	0.03	0.08	0.02	**0.17**	0.03	**0.19**	0.04	**0.15**	0.04	**0.10**	0.06	0.03
14	0.00	0.00	0.02	0.03	0.05	0.08	**0.13**	0.09	**0.17**	0.07	**0.13**	0.05	0.07	0.03	0.03	0.03	0.01
15	0.00	0.00	0.01	0.01	0.01	0.03	0.06	0.04	0.13	0.04	**0.21**	0.05	**0.16**	0.05	**0.12**	0.06	0.03
16	0.00	0.00	0.02	0.02	0.01	0.02	0.03	0.03	0.08	0.03	**0.19**	0.06	**0.19**	0.02	**0.14**	**0.11**	0.04



(a)



(b)



(c)

Fig.6: Typical examples: (a) State pair 15-9 at the beginning of sentence, (b) 14-9 at the beginning of phrase, and (c) 4-4-3-3 at the end of sentence.

Fig. 7 shows the probabilities of prosodic state given syllables before and after comma and period, i.e., $P(p \mid PPBPM\_BW)$ and $P(p \mid EW\_FPBPM)$. It can be found from the figure that the beginning syllables of sentence stay at States 8, 11,

12, 14 and 15 with high probabilities while the beginning syllables were probably associated with States 2, 3, 4 and 5.

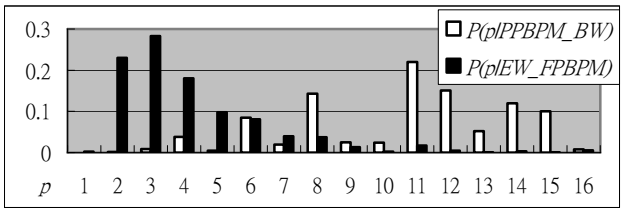


Fig. 7: The distributions of prosodic states at the beginning and ending locations of sentence.

Fig. 8 displays the autocorrelations of the means of the original syllable pitch contour and the prosodic-state affecting patterns. With the excluding of the local affections of tone and inter-syllable coarticulation, the prosodic-state affecting patterns have higher autocorrelation.

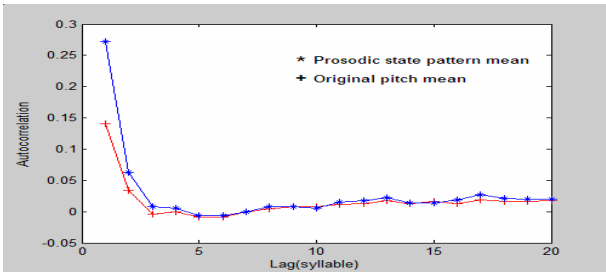


Fig. 8. The autocorrelations of the means of the original syllable pitch and the prosodic-state affecting patterns

Table 3 shows some statistics of eight coarticulation states. It can be found from the table that the first two states have higher hit rates to PM (comma and period) and have longer pause. So they correspond to major and minor breaks with no- or loosely-coupling coarticulation. On the other hand, the last four states have higher probabilities of intra-word and shorter pause durations. So they correspond to states of tightly-coupling coarticulation.

Table 3: Some statistics of eight coarticulation states.

C_n	1	2	3	4	5	6	7	8
P(inter| C_n)	0.85	0.72	0.70	0.67	0.48	0.38	0.32	0.35
P(intra| C_n)	0.15	0.28	0.31	0.33	0.52	0.62	0.68	0.65
P(comma| C_n)	0.32	0.07	0.04	0.04	0.02	0.03	0.02	0.02
P(period| C_n)	0.09	0.02	0.01	0.02	0.01	0.01	0.00	0.01
P(non-PM | C_n)	0.58	0.90	0.95	0.94	0.97	0.97	0.98	0.98
Average Pause duration (ms)	225	76	48	48	28	23	23	23

Fig. 9 displays a typical example of the reconstructed 3-3 tone pattern. It can be seen from the figure that the second Tone 3, which had been changed to a

sandhi Tone 2, was well-reconstructed via the use of coarticulation affecting pattern.

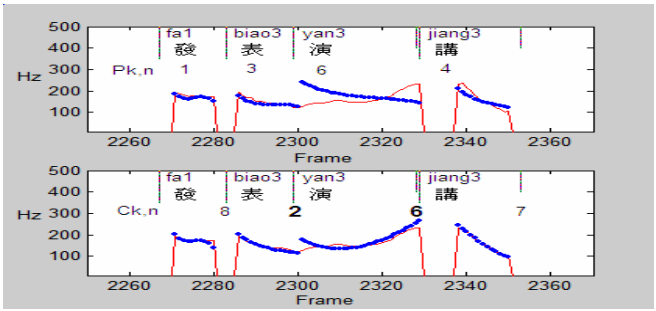
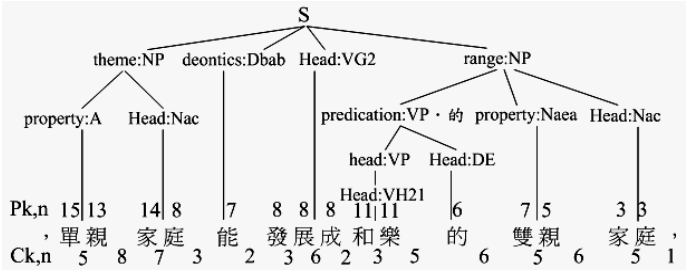
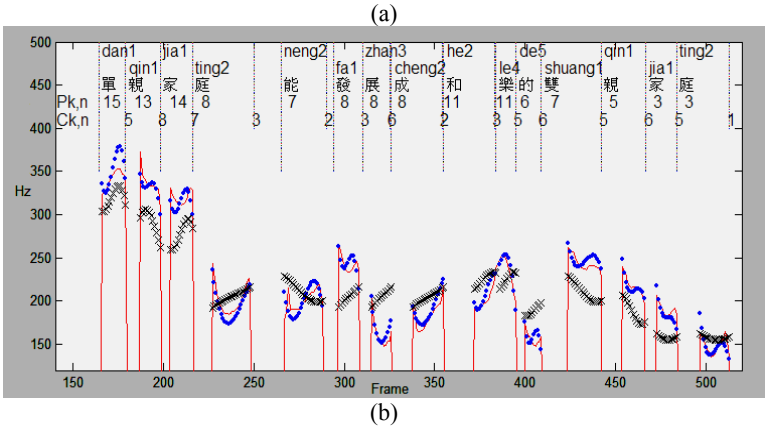


Fig. 9. A typical examples of the reconstructed 3-3 tone pattern: (a) without and (b) with using coarticulation affecting patterns.

Fig. 10 displays a typical example of the reconstructed pitch contour and prosodic-state patterns of a sentence. It can be found from the figure that the reconstructed pitch contour matches well with its original counterpart. We also found that the trajectory of the prosodic-state patterns was smoother and looked more resemble to a sequence of prosodic-word/phrase patterns. Moreover, a typical prosodic state pair of 15-13 (3-3) was appear at the beginning (end) of the sentence.



(a)



(b)

Fig. 10: A typical example: (a) the syntactic tree of a sentence and (b) the original (—) and reconstructed (···) pitch contours, and mean+prosodic-state patterns (xxx).

3.2.3 *Discussions and Conclusions*

A statistical syntax-prosody model of syllable pitch contour for Mandarin speech was discussed in this section. Experimental results showed that the model performed well on separating the influences of several major affecting factors. Many prosodic cues, which are linguistically meaningful, can be found by the model. Not only the individual prosodic state and coarticulation state but also the state sequences are useful in constructing/analyzing the hierarchical prosody structure of Mandarin speech. It seems better to perform prosodic phrase analysis basing on the prosodic state sequence because the interference from local affecting factors, such as base-syllable and tone, can be eliminated.

With the capability of building explicit relationships of syntactic information and observed syllable pitch contour parameters, the model can be applied to assist MTTS in performing automatic prosodic labeling to obtain large well-annotated training corpora, in predicting prosodic phrase boundary or break, and in generating prosodic information for speech synthesis. This is worth further studying in the future.

4. Conclusions

It is believed that a sophisticated tone model would be effective on providing proper prosodic information to significantly improve the naturalness of the output of unlimited MTTS systems. In this chapter, two modern tone modeling methods for MTTS have been discussed. They constructed computational models to analyze the relationship between hierarchical linguistic structure and prosody structure in a quantitative way. Experimental results confirmed their effectiveness.

References

1. C. Tseng, S. Pin, and Y. Lee, "Speech prosody: issues, approaches and implications," in Fant, G., H. Fujisaki, J. Cao and Y. Xu Eds, *From Traditional Phonology to Mandarin Speech Processing, Foreign Language Teaching and Research Process*, 417-438 (2004)
2. C. Tseng, and F. Chou, "A prosodic labeling system for Mandarin speech database," *in Proc. International Congress of Phonetic Sciences,* 2379-238 (1999).
3. C. Tseng, and Y. Lee, "Speech rate and prosody units: evidence of interaction from Mandarin Chinese," *in Proc. Speech Prosody*, 251-254 (2004).
4. H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan* (E) 5(4), 233-241 (1984).
5. H. Fujisaki, "Modeling in the study of tonal feature of speech with application to multilingual speech synthesis," *in Proc. SNLP-O-COCOSDA* (2002).
6. H. Mixdorff, "Quantitative tone and intonation modeling across languages," *in Proc. Int. Symp. on Tonal Aspects of Languages- with Emphasis on Tone Languages*, 137-142 (2004).

7. H. Mixdorff, Y. Hu, and G. Chen, "Towards the automatic extraction of Fujisaki model parameters for Mandarin," *in Proc. of European Conf. on Speech Communication*, (2003).
8. C. Tseng and S. Pin, "Mandarin Chinese prosodic phrase grouping and modeling - method and implications," *in Proc. Int. Symp. on Tonal Aspects of Languages- with Emphasis on Tone Languages*, 193-19 (2004).
9. C. Tseng and S. Pin, "Modeling prosody of Mandarin Chinese fluent speech via phrase grouping," *in Proc. ICSLT-O-COCOSDA* (2004).
10. C. Tseng and Y. Lee, "Intensity in relation to prosody organization," *in Proc. International Symposium on Chinese Spoken Language Processing*, 217-220 (2004)
11. S. Pin, Y. Lee, Y. Chen, H. Wang, and C. Tseng, "A Mandarin TTS system with an integrated prosodic model," *in Proc. ISCSLP* (2004)
12. M. J. Charpentier and M. G. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," *in Proc. IEEE ICASSP*, 2015-2018 (1986).
13. H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," *in Proc. IEEE ICASSP*, 1281-1284 (2000).
14. K. Chen, C. Tseng, H. Peng, and C. Chen, "Predicting prosodic words from lexical words – a first step towards predicting prosody from text," *in Proc. ISCSLP* (2004).
15. C. Shih, "Tone and Intonation in Mandarin," Working Papers of the Cornell Phonetics Laboratory, no. 3, pp.83-109, June (1988)
16. Z. J. Wu, "Can Poly-Syllabic Tone-Sandhi Patterns be the Invariant Units of Intonation in Spoken Standard Chinese," *in Proc. ICSLP*, pp.12.10.1–12.10.4 (1990)
17. Huang, Chu-Ren, Keh-Jiann Chen, Feng-Yi Chen, Keh-Jiann Chen, Zhao-Ming Gao and Kuang-Yu Chen, "Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface", *in Proc. of 2nd Chinese Language Processing Workshop*, 29-37 (2000)
18. L. Lee, C. Tseng, M. Ouh-young, "The Synthesis Rules in a Chinese Text-to-speech System," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol.37, no.9, pp.1309–1319 (1989)
19. S. H. Chen, W. H. Lai, and Y. R. Wang, " A New Duration Modeling Approach for Mandarin Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 4 (2003)
20. S. H. Chen, W. H. Lai, and Y. R. Wang, "A statistics-based pitch contour model for Mandarin speech," *J. Acoust. Soc. Am.*, 117 (2), pp.908-925 (2005)
21. C. Y. Chiang, S. H. Chen and Y. R. Wang, "On the inter-syllable coarticulation effect of pitch modeling for Mandarin speech", *in Proc. Interspeech*, Sept. (2005)
22. H. Dong, J. Tao and B. Xu, "Prosodic Word Prediction Using the Lexical Information," *in Proc. NLP-KE'05*, 189-193 (2005)