# 行政院國家科學委員會專題研究計畫　成果報告

## 中文語音聲學模式及韻律模式之進一步研究(3/3)

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-009-020-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：國立交通大學電信工程學系(所)

計畫主持人：陳信宏

計畫參與人員：江振宇、蕭希群、楊智合

報告類型：完整報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 95 年 9 月 26 日

# 行政院國家科學委員會補助專題研究計畫成果報告

# 中文語音聲學模式及韻律模式之進一步研究(3/3)

Further Studies on Acoustic Modeling and Prosodic Modeling for Mandarin Speech

計畫類別：個別型計畫

計畫編號：NSC－94－2213－E－009－020

執行期間：94 年　8 月　1 日至　95 年　7 月　31 日

計畫主持人：陳信宏

計畫參與人員：江振宇、蕭希群、楊智合

成果報告類型(依經費核定清單規定繳交)：完整報告

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

☑出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢

□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立交通大學電信工程學系

中　華　民　國　95 年　9 月　25　日

# 1. Abstract

This report presents the results of our studies on prosodic modeling for Mandarin speech and the use of prosodic information in automatic speech recognition (ASR). In prosodic modeling, we first propose a statistical pitch contour model to consider some major affecting factors, and then extend the model to further incorporate with the inter-syllable coarticulation and syntactic information. In the use of prosodic information in ASR, a new approach of using temporal information to assist in Mandarin speech recognition is proposed. It incorporates two types of temporal information into the recognition search. One is a statistical syllable duration model which considers the influences of 411 base-syllables, 5 tones, 4 position-in-word factors, and 3 position-in-sentence factors on syllable duration. Another is the timing information of modeling three types of inter-syllable boundary including intra-word, inter-word without punctuation mark (PM), and inter-word with PM. The uses of these two types of temporal information are expected to be useful for improving the segmentation accuracies in both acoustic decoding and linguistic decoding. Experimental results showed that the base-syllable/character/word recognition rates were slightly improved for both MATBN and Treebank database.

**Keywords:** Prosodic modeling, Inter-syllable coarticulation, Syntactic information, Automatic speech recognition

本報告探討中文語音訊號之韻律參數模式及使用韻律信息於語音辨認，在韻律參數模式方面，我們提出一個音節基頻軌跡統計模式，考慮了一些主要影響基頻軌跡變化之因素，以及將音節間的互相影響及語法信息加入模式中。在使用韻律信息於語音辨認方面，我們提出一個使用兩種時間信息來幫助語音辨認的新作法，第一種時間信息是音節長度，將一些主要影響音節長度變化之因素考慮進來，用以協助規範語音辨認之搜尋；另一種時間信息是音節邊界長度信息，考慮句首中尾、詞首中尾、標點符號處之音節邊界停頓長度變化，用以協助規範語音辨認。

**關鍵詞：**韻律軌跡模型、音節互連、語法信息、語音辨認

## 2. Introduction

The technologies of automatic speech recognition and text-to-speech have great progress in recent years. But they are still not widely used in the market. Further studies are needed to make them useful in practical applications. One topic discussed in this report is the study of the dynamic variations of prosodic features for developing high-performance Mandarin text-to-speech systems. Prosodic features to be considered include the fundamental frequency contour, energy contour and duration information of syllable as well as the inter-syllable pause duration. A speech database with syntactic tree labeling is used in this study. Prosodic modeling to describe the relationship of prosodic feature variations and various linguistic features will be exploited. Another topic discussed is the use of prosodic information in automatic speech recognition. We initiate the study from the use of temporal information to assist in Mandarin speech recognition.

The report is organized as follows. Section 3 presents the study of prosody modeling. Section 4 discusses the study of using temporal information in Mandarin ASR. Section 5 gives some conclusions. Section 6 lists the publications of the research.

# 3. Incorporating of Syntactic Information in Pitch Modeling for Mandarin Speech

In this section, a statistics-based syntax-prosody model of *F*0 for Mandarin speech is reported. The model considers three major affecting factors on the syllable pitch contour, including lexical tone, prosodic state and inter-syllable coarticulation effect. The study emphasizes on the incorporation of information extracted from syntactic tree into the model. An explicit relationship of the syntactic information and prosodic state is hence constructed. Experimental results show that the model performed well. By examining the prosodic states labeled automatically by the model, we found that most of them are syntactically meaningful. So it is a promising *F*0 model.

## 3.1 Introduction

Prosody modeling is to explore the information carrying on the prosodic features of human's speech. Many issues are concerned in prosody modeling. They include the labeling of important prosodic cues [2], the construction of prosody hierarchy [3], the modeling of syntax-prosody relationship [6], the prediction of prosodic phrase boundary (break) from text, etc. It can be applied to many fields including speech recognition (SR) and text-to-speech (TTS) [4]. In SR, important prosodic cues can be explored from the input utterance to assist in both acoustic and linguistic decoding. In TTS, a good prosody model can be used to generate appropriate prosodic features from the input text. Among all prosodic features in prosodic modeling, *F*0 is the most important one. This is especially true for Mandarin speech which is known as a tonal language. In this paper, we are interested in syntax-prosody modeling to exploit the relationship of *F*0 contour and linguistic features for Mandarin speech.

In a previous study, a statistics-based pitch model of Mandarin speech which considers three major affecting factors that influence the syllable pitch contour was discussed [1]. Those three affecting factors were the lexical tone and the prosodic state of the current syllable, and the inter-syllable coarticulation between two neighboring syllables. In that study, prosodic state roughly represented the state of the current syllable in a prosodic phrase and was treated as hidden. It was introduced to implicitly account all the effects of higher-level linguistic features on affecting the pitch contour variation. In this paper, we extend the previous prosodic modeling study to incorporate explicit syntactic information into the model to obtain a better syntax-prosody model for F0.

## 3.2 The Proposed Pitch Model

The proposed syllable pitch contour model considers the following three major affecting factors: lexical tone, prosodic state and inter-syllable coarticulation. The model is formulated based on the assumption that all affecting factors are combined additively and can be expressed by

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

Where $\mathbf{x}_{k,n}$ and $\mathbf{y}_{k,n}$ are vectors of four orthogonal expansion coefficients [4] representing, respectively, the observed and normalized (i.e., residual) pitch contours of the *n*-th syllable in utterance *k*; $\boldsymbol{\chi}_{t_{k,n}}$ is the affecting pattern of tone $t_{k,n} \in \{1,2,3,4,5\}$; $\boldsymbol{\chi}_{p_{k,n}}$ is the affecting pattern of prosodic state $p_{k,n} \in \{1,2,\cdots,P\}$; $c_{k,n} \in \{1,2,\cdots,C\}$ is the coarticulation state of the inter-syllable location between syllables $n$ and $n+1$; $tp_{k,n} \in \{(1,1),(1,2),\cdots,(5,5)\}$ 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 $c_{k,n-1}$; $\chi^{b}_{c_{k,n},tp_{k,n}}$ is the backward affecting pattern of the tone pair $tp_{k,n}$ with coarticulation state $c_{k,n}$. Fig. 1 displays the relationship of syllable pitch contours and these affecting factors.

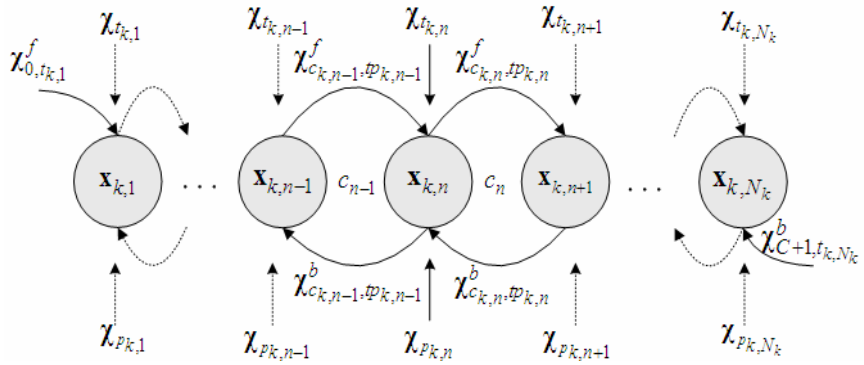


Fig. 1: The relationship of syllable pitch contours and affecting factors used.

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

$$N(\mathbf{x}_{k,n};\boldsymbol{\mu}+\chi_{t_{k,n}}+\chi_{p_{k,n}}+\chi^{f}_{c_{k,n-1},tp_{k,n-1}}+\chi^{b}_{c_{k,n},tp_{k,n}},\mathbf{R}) \tag{2}$$

Here, both the prosodic state, representing the state 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 $P(\mathbf{i}_{k,n} \mid c_{k,n})$ which describes the relationship of 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 $P(s_{k,n} \mid p_{k,n})$ 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 the current study, the model of $c_{k,n}$ involves three features and is expressed by

$$P(\mathbf{i}_{k,n} \mid c_{k,n}) = P(PD_{k,n} \mid c_{k,n})P(PM_{k,n} \mid c_{k,n})P(IW_{k,n} \mid c_{k,n}) \tag{3}$$

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

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 [6]. 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

$$P(\mathbf{s}_{k,n} \mid p_{k,n}) = P(\mathbf{s}_{k,n} = sr_i \mid p_{k,n}) \tag{4}$$

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)<br>· beginning-of-word (BW)<br>· end-of-word (EW)<br>· single-syllable-word (SW) |
| type of the preceding phrase at the same level in the tree | · single-syllable- word (PSW)<br>· 2 or 3-syllable word (PW23)<br>· 4 or more-syllable word (PW4)<br>· phrase boundary without PM (PPB)<br>· phrase boundary with PM (PPBPM) |
| type of the following phrase at the same level in the tree | · single-syllable- word (FSW)<br>· 2 or 3-syllable word (FW23)<br>· 4 or more-syllable word (FW4)<br>· phrase boundary without PM (FPB)<br>· phrase boundary with PM (FPBPM) |
| $sr_i$ | (PSW\| PW23\| PW4\| PPB\| PPBPM)_BW<br>5 combinations |
| | EW_(FSW\| FW23\| FW4\| FPB\| FPBPM)<br>5 combinations |
| | (PSW\| PW23\| PW4\| PPB\| PPBPM)_SW_(FSW\| FW23\|<br>FW4\| FPB\| FPBPM)<br>25 combinations |
| | WW<br>1 combination |

### 3.2.1 The training of the pitch model

To estimate the parameters of the model, a sequential optimization procedure based on the ML criterion is adopted. It first defines a likelihood function expressed by

$$\begin{aligned}
\mathrm{L} &= \log[P(\mathbf{x}, \mathbf{s}, \mathbf{i} \mid \mathbf{p}, \mathbf{c}, \lambda) \\
&= \log \prod_{k=1}^{K} P(\mathbf{x}_k, \mathbf{s}_k, \mathbf{i}_k \mid \mathbf{p}_k, \mathbf{c}_k, \lambda) \\
&\approx \log \prod_{k=1}^{K} P(\mathbf{x}_k \mid \mathbf{p}_k, \mathbf{c}_k, \lambda_{\mathbf{x}}) P(\mathbf{s}_k \mid \mathbf{p}_k, \lambda_{\mathbf{p}}) P(\mathbf{i}_k \mid \mathbf{c}_k, \lambda_{\mathbf{c}}) \\
&= \sum_{k=1}^{K} \sum_{n=1}^{N_k} \log P(\mathbf{x}_{k,n} \mid p_{k,n}, c_{k,n-1}, c_{k,n}, \lambda_{\mathbf{x}}) \\
&\quad + \sum_{k=1}^{K} \sum_{n=1}^{N_k} \log P(\mathbf{s}_{k,n} \mid p_{k,n}, \lambda_{\mathbf{p}}) + \sum_{k=1}^{K} \sum_{n=1}^{N_k} \log P(\mathbf{i}_{k,n} \mid c_{k,n}, \lambda_{\mathbf{c}})
\end{aligned} \tag{5}$$

where $\lambda = \{\lambda_{\mathbf{x}}, \lambda_{\mathbf{p}}, \lambda_{\mathbf{c}}\}$; $\lambda_{\mathbf{x}} = \{\boldsymbol{\chi}_t, \boldsymbol{\chi}_p, \boldsymbol{\chi}^f_{c,tp}, \boldsymbol{\chi}^f_{0,t}, \boldsymbol{\chi}^b_{c,tp}, \boldsymbol{\chi}^b_{C+1,t}, \boldsymbol{\mu}, \mathbf{R}\}$ is the parameter set of the syllable pitch model; $\lambda_{\mathbf{p}}$ and $\lambda_{\mathbf{c}}$ are, respectively, the parameter sets of the prosodic state and coarticulation state models; $K$ is the total number of utterances; and $N_k$ is the total number of syllables in utterance $k$. Then, it sequentially updates the three types of affecting factors (i.e., tone, prosodic state, and inter-syllable coarticulation), re-labeling the prosodic state and updating prosodic state model, and re-labeling the coarticulation state and updating coarticulation state model to optimize the likelihood function $L$. The procedure is iteratively executed until a convergence has reached.

The sequential optimization training procedure executes the following steps iteratively. Each step optimally updates a part of parameters.

**Step 0:** Initialization

- Derive the initial affecting patterns $\boldsymbol{\chi}_t$ of five tones by averaging all $F0$ contour samples of each tone.
- Derive initial prosodic state patterns $\boldsymbol{\chi}_p$ and label the prosodic state of each syllable by VQ using the residues $\mathbf{x}1_{k,n} = \mathbf{x}_{k,n} - \boldsymbol{\chi}_{t_{k,n}}$. Find the parameter set $\lambda_p$ of the prosodic model.
- Derive initial coarticulation state patterns $\boldsymbol{\chi}_c$ and label coarticulation state of each inter-syllable location by VQ using the residues $\mathbf{x}2_{k,n} = \mathbf{x}_{k,n} - \boldsymbol{\chi}_{t_{k,n}} - \boldsymbol{\chi}_{p_{k,n}}$. Find the parameter set $\lambda_{\mathbf{c}}$ of the coarticulation model.

**Step 1:** Update the affecting patterns $\boldsymbol{\chi}_t$ of five tones with all other parameters fixed.

**Step 2:** Re-label the prosodic state of all syllables by a Viterbi algorithm using the ML criterion to maximize $L$, i.e.,

$$p^*_{k,n} = \arg\max_{p_{k,n}} \{ \sum_{k=1}^{K} \sum_{n=1}^{N_k} \log P(\mathbf{x}_{k,n} \mid p_{k,n}, c_{k,n-1}, c_{k,n}, \lambda_{\mathbf{x}}) + \sum_{k=1}^{K} \sum_{n=1}^{N_k} \log P(\mathbf{s}_{k,n} \mid p_{k,n}, \lambda_{\mathbf{p}}) \} \tag{6}$$

for $1 \le n \le N_k$ and $1 \le k \le K$. Update the affecting patterns $\boldsymbol{\chi}_p$ of $P$ prosodic states and the parameter set $\lambda_p$ of the prosodic model.

**Step 3:** Re-label the coarticulation state of all syllables by a Viterbi algorithm using the ML criterion to maximize $L$, i.e.,

$$\begin{aligned} c^*_{k,1}, c^*_{k,2}, \cdots, c^*_{k,N_k} = \arg\max_{c_{k,n}} \sum_{n=1}^{N_k} \{ & \\ & \log N(\mathbf{x}_{k,n}; \boldsymbol{\mu} + \boldsymbol{\chi}_{t_{k,n}} + \boldsymbol{\chi}_{p_{k,n}} + \boldsymbol{\chi}^f_{c_{k,n-1}, tp_{k,n-1}} + \boldsymbol{\chi}^b_{c_{k,n}, tp_{k,n}}, \mathbf{R}) \\ & + \log P(PD_{k,n} \mid c_{k,n}) + \log P(PM_{k,n} \mid c_{k,n}) + \log P(IW_{k,n} \mid c_{k,n}) \} \end{aligned} \tag{7}$$

for $1 \le k \le K$. Update the coarticulation state patterns $\boldsymbol{\chi}_c$ of $C$ coarticulation states and the parameter set $\lambda_c$ of the coarticulation model.

### 3.3 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 [5]. The database consisted of 380 utterances with 52192 syllables.

In the simulation, 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 *F*0 were

$$\mathbf{R}_x = \begin{bmatrix} 2869 & -78 & -142 & -53 \\ -78 & 371 & 27 & -48 \\ -142 & 27 & 68 & -66 \\ -53 & -48 & -66 & 63 \end{bmatrix} \Rightarrow \quad \mathbf{R}_y = \begin{bmatrix} 27 & -2.6 & 1.8 & -8.1 \\ -2.6 & 42 & 0.23 & -0.75 \\ 1.8 & 0.23 & 22 & 1.35 \\ -8.1 & -0.75 & 1.35 & 25 \end{bmatrix}$$

$$|\mathbf{R}_x| = 3.47 \times 10^9 \qquad |\mathbf{R}_y| = 5.39 \times 10^5$$

The variances had been reduced significantly by applying the model. This was especially true for the pitch mean. Fig.2 displays the patterns of five tones. They matched very well with our knowledge of standard tone patterns.

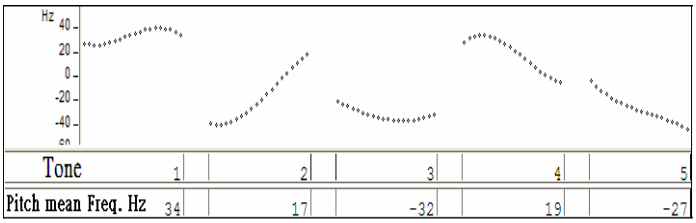



Fig.2: The affecting patterns and their *F0* mean values of five tones.

Fig. 3 shows the affecting patterns of 16 prosodic states. As shown in the figure, States 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 boundary 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 show that the low and flat trail pattern of prosodic phrase (see Fig.4(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 boundary 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 Fig.4(a) and (b)) of prosodic phrase is common to appear. Typical examples are displayed in Fig.4.

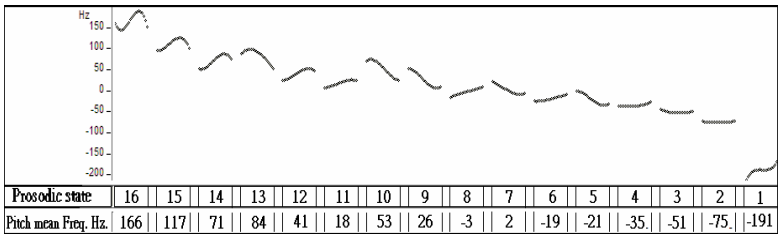
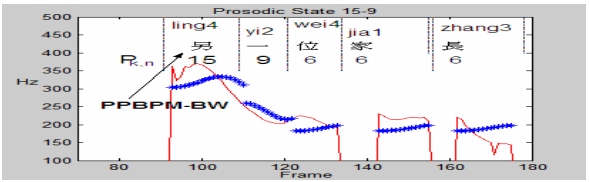
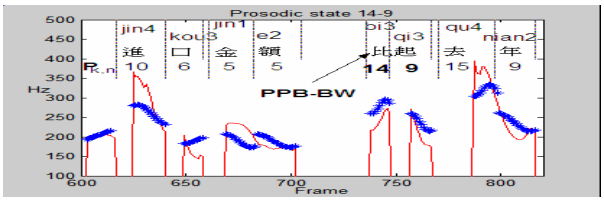


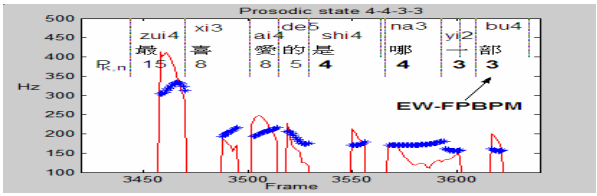
Fig.3: The affecting patterns and their *F0* mean values of 16 prosodic states.




(a)




(b)




(c)

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

Fig. 5 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 States 8, 11, 12, 14 and 15 were located at the beginning of sentence while States 2, 3, 4 and 5 were at the end of sentence. Fig.6 displays the autocorrelations of the means of the original syllable *F*0 and the prosodic-state affecting patterns. With the excluding of the local affections of tone and inter-syllable coarticulation, the prosodic-state affecting patterns shows higher autocorrelation.

Table 2 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.

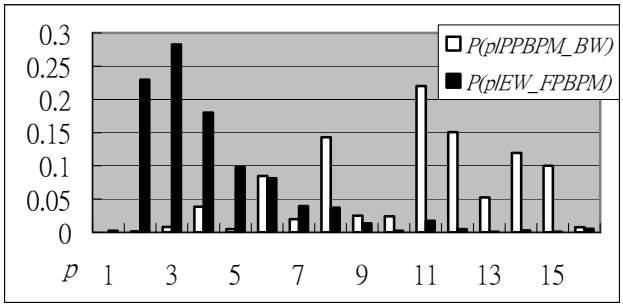



Fig. 5: The distributions of prosodic states at the beginning and end of sentences.

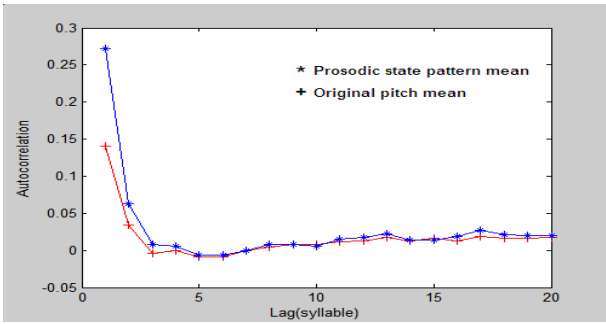



Fig. 6: The autocorrelations of the means of the original syllable pitch and the prosodic-state affecting patterns.

Table 2: 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.7 displays a typical example of the reconstruction of 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 coaticulation affecting pattern. Fig.8 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 matched its original counterpart well. 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.

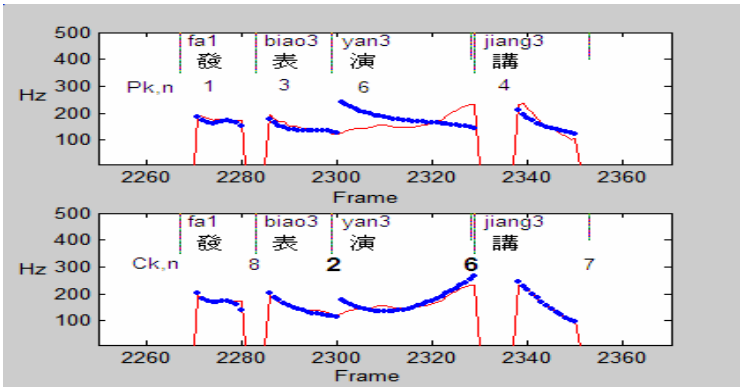
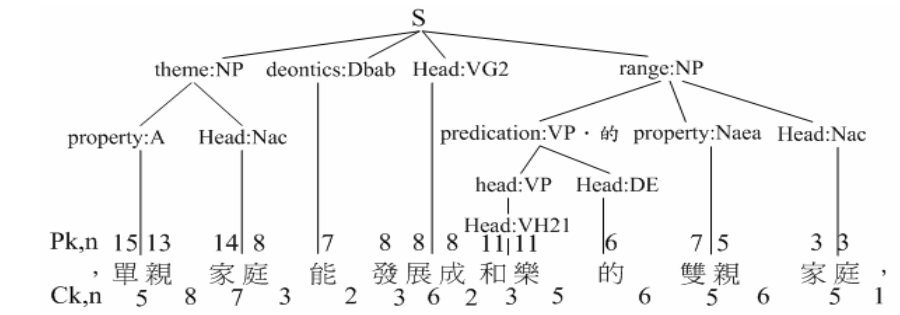
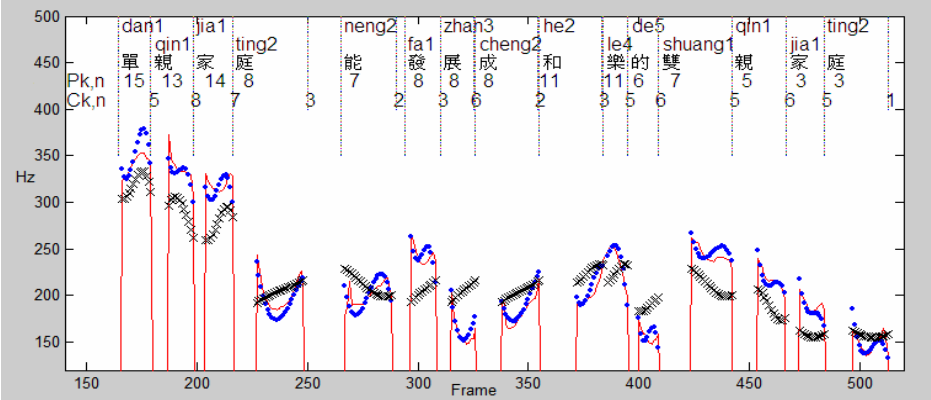



Fig. 7: A typical examples of the reconstructed 3-3 tone patterns: (a) without and (b) with using coarticulation affecting patterns.




(a)




(b)

Fig. 8: A typical example: (a) the syntactic tree of a sentence and (b) the original (—) and reconstructed (···) pitch contours, and mean+prosodic-state patterns (xxx).

### 3.4 Conclusions

A new statistics-based syntax-prosody model of syllable *F*0 contour for Mandarin speech was discussed in this paper. Experimental results showed that the model performed well. Many prosodic cues which are syntactically meaningful can be found by the model. With the construction of explicit relationship of syntactic information and prosodic features, the model can be applied to assist in both ASR and TTS to improve their performances.

# 4. The Use of Temporal Information in Mandarin Speech Recognition

In this section, a new approach of using temporal information to assist in Mandarin speech recognition is discussed. It incorporates two types of temporal information into the recognition search. One is a statistical syllable duration model which considers the influences of 411 base-syllables, 5 tones, 4 position-in-word factors, and 3 position-in-sentence factors on syllable duration. Another is the timing information of modeling three types of inter-syllable boundary including intra-word, inter-word without punctuation mark (PM), and inter-word with PM. The uses of these two types of temporal information are expected to be useful for improving the segmentation accuracies in both acoustic decoding and linguistic decoding. Experimental results showed that the base-syllable/character/word recognition rates were slightly improved for both MATBN and Treebank database.

## 4.1 Introduction

A real-world speech signal always contains rich temporal information ranging from lower-level information, such as phone/syllable/word duration, to higher-level rhythmic information, such as the final-syllable lengthening of prosodic phrase [8]. The temporal information is known to be very helpful for human beings to understand the speech more easily. However, in automatic speech recognition (ASR), the use of temporal information is still primitive. The most basic approach is to incorporate explicit state/phone/syllable duration models or durational constraints into the recognition search for improving the recognition accuracy [9-11]. Another approach is to invoke an embedded phone/syllable/word segmentation in the recognition search process to provide additional acoustic cues to assist in the recognition [12]. But, in all those studies only lower-level durational information, such as HMM state duration or syllable/word duration, was used. No higher-level temporal information was used.

In this paper, a preliminary study of more sophisticatedly using temporal information to improve the ASR for Mandarin speech is discussed. It first extends the conventional base-syllable duration model to consider the influences of three additional affecting factors including tone, position-in-word, and position-in-sentence. With this extension, some higher-level temporal information is invoked in the recognition search. Secondly, it incorporates explicit timing information into the recognition search via constructing models for three types of inter-syllable boundary. These three types of inter-syllable boundary include intra-word, inter-word without punctuation mark (PM), and inter-word with PM. The timing information is expected to be helpful on the segmentation of correct word sequence in the recognition search.

## 4.2 The Proposed Approach

We consider the criterion of speech recognition

$$\begin{aligned} W^*, \Upsilon^* &= \arg\max_{W,\Upsilon} p(W, \Upsilon | X_s, X_p, \Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) \\ &= \arg\max_{W,\Upsilon} p(X_s, X_p, W, \Upsilon | \Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) \\ &= \arg\max_{W,\Upsilon} p(X_s, X_p | W, \Upsilon, \Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) p(W, \Upsilon | \Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) \end{aligned} \tag{8}$$

where $W$ is a word sequence candidate which is composed of words and PMs; $\Upsilon$ is a segmentation candidate which is composed of an HMM state sequence $\Phi_{state}$ and classes of inter-syllable boundaries $\Upsilon_b$; $\Lambda_a$, $\Lambda_t$, $\Lambda_d$ and $\Lambda_l$ denote respectively base-syllable (initial/final) acoustic model (AM), tone model (TM), syllable duration model (DM) and language model

(LM); and $X_s$ and $X_p$ represent the spectral feature vector sequence and the prosodic feature vector sequence of the input utterance, respectively. In this study, we consider three classes of inter-syllable boundary including intra-word, inter-word without major PM and inter-word with major PM. Here, only major PMs belonging to {、, ，, 。, ；, ：, ？, ！} are considered. We denote them as Intra, Inter and Inter-PM, respectively. The first term in Eq.(8) is generally known as the score of acoustic decoding and the second one is the score of segmentation and language decoding.

The score of acoustic modeling can be further simplified and expressed by

$$\begin{aligned} & p(X_s, X_p \mid W, \Upsilon, \Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) \\ & \quad \approx p(X_s, X_p \mid W, \Upsilon, \Lambda_a, \Lambda_t) \\ & \quad = p(X_s \mid W, \Upsilon, \Lambda_a, \Lambda_t, X_p) p(X_p \mid W, \Upsilon, \Lambda_a, \Lambda_t) \\ & \quad \approx p(X_s \mid W, \Upsilon, \Lambda_a, X_p) p(X_p \mid W, \Upsilon, \Lambda_t) \end{aligned} \tag{9}$$

Here, we assume that LM $\Lambda_l$ and DM $\Lambda_d$ are independent of the acoustic decoding. In Eq.(9), the first term is the score of HMM acoustic modeling using spectral features and the second term is the score of prosodic feature decoding. In the current study, we consider 411 base-syllables as the basic acoustic recognition units to further simplify $p(X_s \mid W, \Upsilon, \Lambda_a, X_p)$ as $p(X_s \mid S, \Phi_{state}, \Lambda_a)$, where $S$ is the base-syllable sequence associated with the candidate word sequence $W$ and $\Phi_{state}$ is a candidate of HMM state sequence associated with $S$.

The score of prosodic feature decoding can be further simplified and separated into two terms

$$\begin{aligned} & p(X_p \mid W, \Upsilon, \Lambda_t) \\ & \quad = p(X_t, X_b \mid W, \Upsilon, \Lambda_t) \approx p(X_t \mid T, \Lambda_t) p(X_b \mid \Upsilon_b) \end{aligned} \tag{10}$$

where $X_p = (X_t, X_b)$; $X_t$ is the prosodic features for tone recognition; $X_b$ is the prosodic features for inter-syllable boundary classification; $p(X_t \mid T, \Lambda_t)$ is the score of tone decoding; $T$ is the tone sequence associated with $W$; $p(X_b \mid \Upsilon_b)$ is the score of inter-syllable boundary classification. In this study, $X_t$ consists of 18 parameters including 9 parameters representing, respectively, $F0$ means, $F0$ slopes, and energy means of three uniformly-segmented pitch contour segments of the current syllable; 3 parameters of the last pitch contour segment of the preceding syllable; 3 parameters of the first pitch contour segment of the succeeding syllable; 2 representing pause durations preceding and following the current syllable; and one representing the duration of the current syllable. And $X_b$ consists of the pause duration, the pitch mean and energy level jumps of the preceding and succeeding syllables, and the lengthening factor of the preceding syllable. Here, both $p(X_t \mid T, \Lambda_t)$ and $p(X_b \mid \Upsilon_b)$ are implemented by the neural network-based approach. In each case, a three-layer MLP (multi-layer perceptrons) is employed to generate output discrimination functions for all its classes. We can use these output discrimination functions to perform classification by choosing the class with maximum output as the recognized one. This can check the effectiveness of the MLP classifier. For this application, we transform them into the likelihood scores by

$$P(X \mid \text{Class } i) = \frac{P(\text{Class } i \mid X)}{\sum_k P(\text{Class } k \mid X)} \tag{11}$$

The score of segmentation and language decoding, which is the second term of Eq.(8), can also be further simplified and expressed by

$$\begin{aligned} & p(W, \Upsilon \mid \Lambda_a, \Lambda_t, \Lambda_d, \Lambda_l) \\ & \quad \approx p(W, \Upsilon \mid \Lambda_d, \Lambda_l) \\ & \quad \approx p(\Upsilon \mid W, \Lambda_d, \Lambda_l)\, p(W \mid \Lambda_d, \Lambda_l) \\ & \quad \approx p(X_d \mid W, \Lambda_d)\, p(W \mid \Lambda_l) \end{aligned} \tag{12}$$

where $p(X_d \mid W, \Lambda_d)$ is the score of syllable duration modeling, $X_d$ is the syllable duration sequence derived from the segmentation information $\Upsilon$, and $p(W \mid \Lambda_l)$ is the score of language decoding. Here, we assume that both acoustic model $\Lambda_a$ and tone model $\Lambda_t$ are independent of the segmentation and language decoding. In this study, a word-bigram model $\Lambda_l$ is used.

The syllable duration model adopted is a simple multiplicative model [9] which involves 4 major affecting factors including base-syllable, tone, position-in-word, and position-in-sentence. In the model, the observed duration of syllable *n* is expressed by

$$X_d[n] = Z_d[n]\, \gamma_{sy_n} \gamma_{t_n} \gamma_{wp_n} \gamma_{sp_n} \tag{13}$$

where $Z_d[n]$ is the normalized (or residue) syllable duration and is modeled by a normal distribution $N(\mu, \sigma^2)$ with mean $\mu$ and variance $\sigma^2$; $\gamma$'s are affecting factors; $sy_n$, $t_n$, $wp_n$ and $sp_n$ represent, respectively, the base-syllable, tone, word position, and sentence position of syllable *n*. In the study, we consider 411 base-syllables, 5 tones, 4 types of position-in-word, and 3 types of position-in-sentence. The 4 types of position-in-word are mono-syllabic word, and the beginning, intermediate and ending syllables of a word. The 3 types of position-in-sentence are the beginning, intermediate and ending syllables of a sentence which is ended with a major PM.

An iterative sequential optimization procedure is employed to train the syllable duration model. It first initializes the training by estimating all affecting factors independently, i.e.,

$$\gamma = \frac{\sum_{n=1}^{N} X_d[n]\, \delta(\gamma_n, \gamma)}{\mu \sum_{n=1}^{N} \delta(\gamma_n, \gamma)} \tag{14}$$

for $\gamma = \gamma_{sy_n}$, $\gamma_{t_n}$, $\gamma_{wp_n}$, or $\gamma_{sp_n}$,

$$\mu = \frac{\sum_{n=1}^{N} \dfrac{X_d[n]}{\gamma_{sy_n} \gamma_{t_n} \gamma_{wp_n} \gamma_{sp_n}}}{N} \tag{15}$$

and

$$\sigma^2 = \frac{\sum_{n=1}^{N} \left( \frac{X_d[n]}{\gamma_{sy_n}\gamma_{t_n}\gamma_{wp_n}\gamma_{sp_n}} - \mu \right)^2}{N}, \tag{16}$$

where $\delta(\cdot,\cdot)$ is the Kronecker delta function and $N$ is the total number of training syllables. It then sequentially estimate the four types of affecting factors, $\gamma_{sy_n}$, $\gamma_{t_n}$, $\gamma_{wp_n}$ and $\gamma_{sp_n}$, one-by-one based on the ML (maximum likelihood) criterion with objective function

$$L = \sum_{n=1}^{N} \log f(X_d[n]) \tag{17}$$

where

$$f(X_d[n]) = N\left( \frac{X_d[n]}{\gamma_{sy_n}\gamma_{t_n}\gamma_{wp_n}\gamma_{sp_n}}; \mu, \sigma^2 \right) \tag{18}$$

$\mu$ and $\sigma^2$ are also updated using Eqs.(15) and (16), respectively. The sequential optimization step is iteratively executed until a convergence is reached.

To reduce the computational complexity, a two-stage recognition search is adopted in this study. In the first stage, a word-lattice which consists of top-10 candidate words is constructed by using only the acoustic model $\Lambda_a$ and the word-bigram LM $\Lambda_l$. Then, in the second stage the best word sequence is determined from the word-lattice by using the criterion shown in Eq.(8). The two-stage recognition search is realized using the HTK toolkit [14].

### 4.3 Experimental Results

Performance of the proposed approach was examined by simulations using two databases. One was the Anchor set of MATBN (Mandarin Chinese Broadcast News Corpus) [13]. It was uttered by 4 anchors in fast speaking styles and is composed of 175,194 training syllables and 14,906 testing syllables. The acoustic models consisted of 100 3-state right-*final*-dependent (RFD) *Initial* HMM models, 40 5-state context-independent (CI) *Final* models, 19 3-state Particle models, one 3-state Breath model, one 3-state Silence model, one 1-state Short Pause model (tied with the middle state of Silence model), and 3 3-state Garbage models. Another database was the read-speech database of Sinica Treebank [6]. It was uttered by a single female announcer in a normal speed. It was composed of 380 utterances with 52,192 syllables. The acoustic models consisted of 100 RFD *Initial* models and 40 CI *Final* models.

For LM, a general bigram LM was first trained using the following three corpora: (1) Sinorama: a news magazine with 9.87 million words; (2) NTCIR: an IR test bench consisting of several domain with 124.4 million words; and (3) Sinica Corpus: general text corpus collected for the language analysis with 4.8 million words. Here a 60,000-word lexicon was used. For the recognition of MATBN, the general LM was adapted using the texts of MATBN which was composed of 1.31 million words with 23,314 particles and 90,052 breathes.

We first examined the syllable duration model. Table 2 shows some affecting factors (AF) for the two databases. They include: (1) mono-syllabic word (MW), and the beginning (BW), intermediate (IW) and ending syllables (EW) of a word for position-in-word; (2) 5 tones; and (3) the beginning (BS), intermediate (IS) and ending syllables (ES) of a sentence for position-in-sentence. It can be found from the table that IW in position-in-word, Tone 5, and BS in position-in-sentence are much shorter; while ES in position-in-sentence is very long.

Table 2: Some affecting factors (AF) of the syllable duration model for the two databases.

| AF / Database | Position-in-word | | | | Position-in-sentence | | |
|---|---|---|---|---|---|---|---|
| | MW | BW | IW | EW | BS | IS | ES |
| MATBN anchor | 1.05 | 0.97 | 0.84 | 1.02 | 0.85 | 0.98 | 1.34 |
| Sinica Treebank | 1.05 | 0.96 | 0.88 | 1.03 | 0.90 | 0.99 | 1.20 |
| | Tone | | | | | | |
| | T1 | | T2 | T3 | | T4 | T5 |
| MATBN anchor | 1.01 | | 1.05 | 0.98 | | 1.02 | 0.73 |
| Sinica Treebank | 1.03 | | 1.07 | 1.00 | | 1.02 | 0.72 |

We then examined the performances of the MLP tone classifier (see Table 3) and MLP inter-syllable boundary classifier (see Table 4). It can be found from Table 3 that tone recognition rates of 75.1 and 85% were achieved for MATBN and Treebank, respectively. Both Tone 1 and Tone 4 were easier to be recognized while Tone 3 and Tone 5 were not. It can also be found from Table 4 that accuracy rates of 58.8% and 69.1% were achieved in the inter-syllable boundary classifications for MATBN and Treebank databases, respectively. The class of inter-word with PM was easier to be correctly detected.

Table 3: Performance of the tone recognizers. (unit: %)

| | T1 | T2 | T3 | T4 | T5 | average |
|---|---|---|---|---|---|---|
| MATBN Anchor | 77.7 | 74.3 | 66.3 | 83.4 | 42.0 | 75.1 |
| Sinica Treebank | 88.0 | 84.4 | 70.8 | 92.6 | 74.9 | 85.0 |

Table 4: Experimental results of the inter-syllable boundary recognizer. (unit: %)

| | Intra | Inter | Inter-PM | average |
|---|---|---|---|---|
| MATBN Anchor | 51.0 | 64.0 | 70.6 | 58.8 |
| Sinica Treebank | 78.8 | 57.3 | 81.9 | 69.1 |

Lastly, we examined the performance of the proposed method of using temporal information in Mandarin speech recognition. Table 5 displays the experimental results. It can be found from Table 5 that the baseline system which used the acoustic and language models, $\Lambda_a$ and $\Lambda_l$ performs well. Base-syllable/character/word recognition rates were 93.49/91.04/86.29 and 94.01/84.99/75.43 for the MATBN anchor and Sinica Treebank databases, respectively. It is noted that both character and word recognition rates for Treebank were relatively low as compared with those of MATBN because Treebank contained much more proper nouns and DM compound words which were treated as individual characters rather than words. The performances were slightly improved as we incorporated the tone recognizer. The performances were further improved for the proposed method as we used the temporal information in the recognition search.

Table 5: The experimental results of the proposed method for Mandarin ASR. (unit: %)

| | | Syllable Recognition. rate | Character Recognition. rate | Word Recog. rate |
|---|---|---|---|---|
| MATBN Anchor | Baseline | 93.49 | 91.04 | 86.29 |
| | Baseline +tone recogn. | 93.59 | 91.15 | 86.51 |
| | Proposed | 93.66 | 91.23 | 86.62 |
| Sinica Treebank | Baseline | 94.01 | 84.99 | 75.43 |
| | Baseline +tone recogn. | 93.89 | 85.21 | 75.73 |
| | Proposed | 94.0 | 85.55 | 75.93 |

### 4.4 Conclusions

A new approach of using a statistical syllable duration model and an inter-syllable boundary model to assist in Mandarin ASR has been discussed in this paper. Experimental results showed that it slightly outperformed the baseline system. Further studies include an analysis of its effectiveness on different type of pronunciation conditions, the use of more sophisticated temporal models, and so on.

## 5. Conclusions

We have proposed a new approach to prosody modeling for Mandarin speech in this project. It employs a statistical model to describe the variation of prosodic features including syllable duration and syllable pitch contour. The model first considers some major affecting factors and then incorporates with inter-syllable coarticulation and syntactic information. Experimental results have confirmed that the proposed approach performed well for the modeling of syllable duration and pitch contour. The prosody model can be applied to prosody labeling, text-to-speech, tone recognition, and speech recognition. A preliminary study of applying the syllable duration model to speech recognition has been conducted. It uses the syllable duration model to provide constraints for assisting in the recognition search. Experimental results have confirmed its effectiveness.

## 6. Publications

1. 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, July 2003.
2. Sin-Horng Chen, Wen-Hsing Lai and Yih-Ru Wang, “A statistical pitch contour model for Mandarin speech,” J. Acoust. Soc. Am., 117 (2), pp.908-925, Feb. 2005.
3. Sin-Horng Chen and Wen-Hsing Lai, “A New Pitch Modeling Approach for Mandarin Speech”, in the proc. of Eurospeech2003.
4. Wei-Chih Kuo, Yih-Ru Wang and Sin-Horng Chen, “A Model-Based Tone Labeling Method for Min-Nan/Taiwanese Speech,” ICASSP2004, Montreal, Canada, May 2004
5. Chen-Yu Chiang, Sin-Horng Chen and Yih-Ru Wang, “On the inter-syllable coarticulation effect of pitch modeling for Mandarin speech”, Interspeech2005, Sept. 2005, Lisbon, Portugal
6. Jyh-Her Yang, Yuan-Fu Liao, Yih-Ru Wang and Sin-Horng Chen, “A New Approach of Using Temporal Information in Mandarin Speech Recognition”, Speech Prosody, May 2006, Dresden, Germany
7. Hsi-Chun Hsiao, Hsiu-Min Yu*, Yih-Ru Wang and Sin-Horng Chen, “Multilingual Speech Corpora for TTS System Development”, Submitted to ISCSLP2006

**REFERENCES**


[1] 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, July

[2] Chiang, Chen-Yu, Wang, Yih-Ru and Chen, Sin-Horng (2005), "On the inter-syllable coarticulation effect of pitch modeling for Mandarin speech", INTERSPEECH-2005, pp. 3269-3272

[3] Colin W. Wightman, Mari Ostendorf, “Automatic Labeling of Prosodic Patterns”, IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, October 1994, pp. 469 – 481.

[4] Tseng, C; Chou, F., 1999, “A prosodic labeling system for Mandarin speech database”, Proceeding of ICPhS 2003, San Francisco, USA, pp. 2379-2382

[5] Sin-Horng Chen, Wen-hsing Lai and Yih-Ru Wang, “A statistics-based pitch contour model for Mandarin speech”, J. Acoust. Soc. Am. 117(2), Feb. 2005, pp. 908 – 925

[6] Huang, Chu-Ren, Keh-Jiann Chen, Feng-Yi Chen, Keh-Jiann Chen, Zhao-Ming Gao and Kuang-Yu Chen. 2000, “Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface”, Proceedings of 2nd Chinese Language Processing Workshop 2000, Hong Kong, pp. 29-37.

[7] Tommy Ingulfsen, Tina Burrows, and Sabine Buchholz, “Influence of syntax on prosodic boundary prediction,” in Interspeech 2005 ,(September 4-8 ,2005) , Lisbon , Portugal, 1817-1820

[8] Tseng, Chiu-yu and Lee, Yeh-lin (2004). “Speech rate and prosody units: Evidence of interaction from Mandarin Chinese,” *Proceedings of the International Conference on Speech Prosody 2004*, (Mar. 23-26, 2004), Nara, Japan, 251-254.

[9] L. Rabiner and B.-H. Juang. Fundamentals of Speech Recognition. Prentice-Hall, 1993, pp 384-385.

[10] A. Anastasakos, R. Schwartz, and H. Shu, “Duration modeling in large vocabulary speech

recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 628–631.
[11] C. Mitchell, M. Harper, L. Jamieson, and R. Helzermam, "A parallel implementation of a hidden markov model with duration modeling for speech recognition," in *Digital Signal Process.*, vol. 5, 1995, pp. 43–57.
[12] W. J. Wang, Y. F. Liao and S. H. Chen, "RNN-based Prosodic Modeling for Mandarin Speech and Its Application to Speech-to-Text Conversion", Speech Communication, 36 (2002), pp.247-265.
[13] Hsin-min Wang, "MATBN 2002: A Mandarin Chinese Broadcast News Corpus" ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003).
[14] Hidden Markov Model Toolkit (HTK) , http://htk.eng.cam.ac.uk