ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

PROSODIC MODELING FOR ISOLATED MANDARIN WORDS AND ITS APPLICATION

Hung-Kuang Shih, Chen-Yu Chiang, Yih-Ru Wang and Sin-Horng Chen
Dept. of Communication Engineering, Chiao Tung University, Hsinchu

ABSTRACT

In this paper, a new approach to syllable-based modeling of F0 contour, duration and energy for isolated Mandarin words is proposed. The syllable F0 contour model considers three major affecting factors, including lexical tone, syllable position in a word and inter-syllable coarticulation effect; while both the duration and energy models additionally consider one more affecting factor of base syllable type. Experimental results on a large single-speaker database showed that the method performed very well. Based on the prosodic model, a learning system for Mandarin word prosody pronunciation is designed and implemented for nonnative speakers.

Index Terms— Prosody modeling, inter-syllable coarticulation effect, Mandarin word prosody pronunciation

1. INTRODUCTION

Prosody modeling is an important research topic in text-to-speech (TTS). A well-designed prosodic model is the key to synthesize natural and pleasant speech. Currently, there are three major approaches of prosody modeling: rule-based, neural network-based and statistical model-based. The rule-based approach tries to generalize human's prosody pronunciation rules from the linguistic point of view [1]. But, the rules are often too complicated to be exploited. The neural network-based approach imitates the learning and memorizing function of human brain. It progressively updates its network to learn the linguistics-prosody relationship [2]. It is criticized as a black box which is hard to analysis. The statistical model-based approach learns a model from a large corpus to build the relation between linguistic features and prosodic features [3]. In this paper, we adopt the statistical model-based approach to learn the prosody generation mechanism of isolated Mandarin words.

The paper is organized as follows. In Section 2, the proposed prosody modeling method is presented. In Section 3, experimental results are discussed. Some conclusions are given in Section 4.

2. THE PROPOSED PROSODY MODELING METHOD

In this study, we consider the modeling of three types of prosodic features including syllable log-F0 contour **sp**, syllable duration **sd** and syllable energy level (maximum energy of final) **se**. These three prosodic features are assumed to be independent of each other and their variation is controlled by four main affecting factors: lexical tone **t**, base syllable type **s**, syllable position in a word **w** and coarticulation state **c**. The model can be generally expressed by

$$P(\mathbf{sp},\mathbf{sd},\mathbf{se}\,|\,\mathbf{t},\mathbf{s},\mathbf{w},\mathbf{c}) = P(\mathbf{sp}\,|\,\mathbf{t},\mathbf{s},\mathbf{w},\mathbf{c})P(\mathbf{sd}\,|\,\mathbf{t},\mathbf{s},\mathbf{w},\mathbf{c})P(\mathbf{se}\,|\,\mathbf{t},\mathbf{s},\mathbf{w},\mathbf{c}) \quad (1)$$

where $P(\mathbf{sp}\,|\,\mathbf{t},\mathbf{s},\mathbf{w},\mathbf{c})$, $P(\mathbf{sd}\,|\,\mathbf{t},\mathbf{s},\mathbf{w},\mathbf{c})$ and $P(\mathbf{se}\,|\,\mathbf{t},\mathbf{s},\mathbf{w},\mathbf{c})$ are syllable log-F0 contour, duration and energy models, respectively.

2.1. Syllable F0 contour model

We assume that the F0 contour of the *n*-th syllable (current syllable) in an isolated spoken word is mainly controlled by three affecting factors including the current lexical tone t_n, the current syllable position in a word w_n, and the coarticulations from the two nearest neighboring tones, t_{n-1} and t_{n+1}, conditioned respectively on the coarticulation states, c_{n-1} and c_n, of the syllable junctures on both sides. Here, coarticulation state c_n represents the degree of coupling between the *n*-th and (*n*+1)-th syllables in a word and is treated as hidden to be labeled. Specifically, the syllable F0 contour is represented by

$$P(\mathbf{sp}\,|\,\mathbf{t},\mathbf{s},\mathbf{w},\mathbf{c}) \approx P(\mathbf{sp}\,|\,\mathbf{t},\mathbf{w},\mathbf{c}) \approx \prod_{n=1}^{N} P(\mathbf{sp}_n\,|\,\mathbf{t},\mathbf{s},\mathbf{w},\mathbf{c}) \approx \prod_{n=1}^{N} P(\mathbf{sp}_n\,|\,t_{n-1}^{n+1}, w_n, c_{n-1}^{n}) \quad (2)$$

where

$$\mathbf{sp}_n = \mathbf{sp}_n^r + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{w_n} + \boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{c_n,tp_n} + \boldsymbol{\mu}^p \quad \text{for } 1 \le n \le N \quad (3)$$

is the observed log-F0 contour of the *n*-th syllable of an *N*-syllable word and is represented by the first four orthogonally-transformed parameters [4]; $c_{n-1}^n=(c_{n-1},c_n)$ is the coarticulation state of syllable junctures on both sides; $t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$ are tone triple; $\mathbf{sp}_n^r$ is the normalized (or residual) version of $\mathbf{sp}_n$; $\boldsymbol{\beta}_{t_n}$ is the affecting pattern (AP) of tone $t_n \in (1,\cdots,5)$; $\boldsymbol{\beta}_{w_n}$ represents the AP of syllable position-in-word $w_n \in \{(i,j)\,|\,i=1\sim 8, j \le i\}$ with (i,j) standing for the *j*-th syllable of an *i*-syllable word; $\boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}}$ and $\boldsymbol{\beta}^b_{c_n,tp_n}$ are the APs of forward (carryover) and backward (anticipatory) coarticulations contributed from syllable *n*-1 and syllable *n*+1, respectively; tp_n is tone pair $t_n^{n+1}=(t_n,t_{n+1})$; and $\boldsymbol{\mu}^p$ is the AP of global mean. For taking care of word boundaries, two special APs of coarticulation, $\boldsymbol{\beta}^f_{t_1}=\boldsymbol{\beta}^f_{c_0,tp_0}$ and $\boldsymbol{\beta}^b_{t_N}=\boldsymbol{\beta}^b_{c_N,tp_N}$, are adopted to represent the effects of word onset and offset, respectively. Fig. 1 displays the relationship between the APs considered and the observed syllable log-F0 contour. By assuming that $\mathbf{sp}_n^r$ is zero-mean and normally distributed, i.e. $N(\mathbf{sp}_n^r;\mathbf{0},\mathbf{R})$, we have

$$P(\mathbf{sp}_n\,|\,t_{n-1}^{n+1}, w_n, c_{n-1}^{n}) = N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}_{w_n}+\boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}}+\boldsymbol{\beta}^b_{c_n,tp_n}+\boldsymbol{\mu}^p,\mathbf{R}^p) \quad \text{for } 1\le n\le N \quad (4)$$

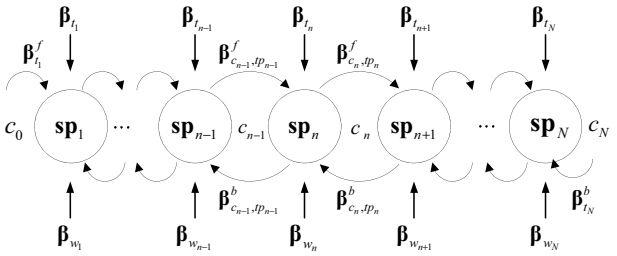

Fig.1: The relationship between the APs considered and the observed syllable F0 contour.

2.2. Syllable duration and energy models

Similar to the syllable F0 contour model, the proposed duration and energy models are expressed by

$$P(sd_n \mid t_{n-1}^{n+1}, s_n, w_n, c_{n-1}^{n}) = N(sd_n; \gamma_{t_n} + \gamma_{w_n} + \gamma_{s_n} + \gamma^{f}_{c_{n-1},tp_{n-1}} + \gamma^{b}_{c_n,tp_n} + \mu^{d}, R^{d}) \quad \text{for } 1 \le n \le N \tag{5}$$

$$P(se_n \mid t_{n-1}^{n+1}, s_n, w_n, c_{n-1}^{n}) = N(se_n; \alpha_{t_n} + \alpha_{w_n} + \alpha_{s_n} + \alpha^{f}_{c_{n-1},tp_{n-1}} + \alpha^{b}_{c_n,tp_n} + \mu^{e}, R^{e}) \quad \text{for } 1 \le n \le N \tag{6}$$

where sd_n and se_n are the observed duration and energy level of the n-th syllable in a word, respectively; γ_x and α_x represent the APs of affecting factor x for syllable duration and energy level, respectively; and $s_n \in (1,\cdots,411)$ is the base-syllable type.

2.3. Training of the proposed model

To estimate the parameters of the model, we first define a log-likelihood function for each isolated word by

$$L^{w} = \log \prod_{n=1}^{N} \left[\begin{array}{l} P(\mathbf{sp}_n \mid t_{n-1}^{n+1}, w_n, c_{n-1}^{n}) \times P(sd_n \mid t_{n-1}^{n+1}, s_n, w_n, c_{n-1}^{n}) \\ \times P(se_n \mid t_{n-1}^{n+1}, s_n, w_n, c_{n-1}^{n}) \end{array} \right] \tag{7}$$

A total log-likelihood L is then calculated by summing over all L^w of training words. Then a sequential optimization procedure based on ML criterion is employed to update the APs and label coarticulation state of each syllable juncture so as to maximize L until a convergence is reached. The sequential optimization procedure is divided into two main parts: *initialization* and *iteration*.

2.3.1 Initialization

Since the coarticulation state of each syllable juncture is treated as a hidden variable to be labeled in the training process, a proper initial determination of coarticulation state for each inter-syllable juncture should be performed prior to the determination of APs. The initialization step is hence divided into two parts: (a) determination of coarticulation state and (b) initialization of APs.

(a) Determination of coarticulation state

In this study, the type of coarticulation state is empirically set to be three, i.e. $c_n \in$(c1,c2,c3) where c1, c2 and c3 represent "strong", "medium" and "weak" couplings between consecutive syllables on syllable juncture, respectively. We first determine c1 junctures by the following rule: a syllable juncture is labeled as c1 if the F0 contours of the two successive syllables are continuous across the juncture. We then use the vector quantization (VQ) technique to divide all other junctures into two classes of c2 and c3. Here, energy-dip level (minimum energy) on syllable juncture is chosen as the feature of VQ so that c2 and c3 are featured as higher and lower energy-dip levels, respectively.

(b) Initialization of APs

Since the observed syllable log-F0 contour, duration and energy level are assumed to be the superimpositions of several APs, the estimation of an AP may be interfered by the existence of the APs of other types. We hence adopt a progressive estimation strategy to first determine the initial APs which can be estimated most reliably and then eliminate their affections from the surface pitch contours for the estimations of the remaining APs. In this study, the order of initial AP estimation is listed as follows: global mean $\{\boldsymbol{\mu}^{p}, \mu^{d}, \mu^{e}\}$, five tones $\{\boldsymbol{\beta}_t, \gamma_t, \alpha_t\}$, coarticulation $\{\boldsymbol{\beta}^{f}_{c,tp}, \boldsymbol{\beta}^{b}_{c,tp}, \gamma^{f}_{c,tp}, \gamma^{b}_{c,tp}, \alpha^{f}_{c,tp}, \alpha^{b}_{c,tp}\}$, syllable position in a word $\{\boldsymbol{\beta}_w, \gamma_w, \alpha_w\}$, and base syllable type $\{\gamma_s, \alpha_s\}$. Lastly, the covariance matrixes $\{\mathbf{R}^{p}, R^{d}, R^{e}\}$ can be obtained via ML estimation.

2.3.2 Iteration

The iteration is a multi-step iterative procedure listed below:

1. Update the APs of five tones.
2. Update the APs of coarticulation.
3. Update the APs of syllable position-in-word.
4. Update the APs of 411 base syllables
5. Update covariance matrixes.
6. Re-label the coarticulation state sequence of each word c_1^{N-1} by the Viterbi search algorithm so as to maximize L^w.
7. Repeat 1 to 6 until a convergence of L is reached.

2.3.3 Decision tree for coarticulation state prediction

Lastly, a binary decision tree for predicting coarticulation states of syllable junctures from the input word in the testing phase is constructed using the training set with all coarticulation states being properly labeled. The CART algorithm is adopted to train the tree.

3. EXPERIMENTAL RESULTS

The performance of the proposed prosody modeling method was evaluated using a read isolated Mandarin word speech corpus of a single female professional announcer. The corpus consists of 107,936 words with 277,218 syllables selected from the NCTU Speech Lab Dictionary. Nine tenth of the corpus was used for training and the remaining for testing. The training process converged after 48 iterations. The variance of the observed data and the MSE of prosody modeling for the inside and outside tests are displayed in Table 1. Notice that the coarticulation states were determined by the Viterbi algorithm for the inside test, while they were determined by the binary decision tree for the outside test.

Table 1: The performance of the prosody modeling

	F0 Contour (log-Hz)		Duration (ms)		Energy (dB)	
	variance	MSE	variance	MSE	variance	MSE
inside	0.0434	0.0119	8372	1637	22.77	8.28
outside	0.0436	0.0126	8389	1729	22.97	8.56

3.1. An analysis of APs: some findings

Fig. 2 displays the APs of F0 for five lexical tones. As shown in the figure that the APs of the first four tones matched well with the well-known standard tone patterns discovered by Chao [5]. As for the APs of syllable duration for five tones, it is found that Tone 2 is the longest and Tone 5 is much shorter than all others. For APs of energy, Tone 1 is the largest and Tone 5 is again the smallest.

Fig. 3 displays the coarticulation APs, $\boldsymbol{\beta}^{f}_{c,tp}$ and $\boldsymbol{\beta}^{b}_{c,tp}$, of F0. Generally, it can be found from the figure that the APs of c3 were close to flat lines to show its characteristic of "weak" coupling, while most APs of c1 were bent more seriously in the beginning or ending parts to compensate the mismatch of connecting two F0 patterns across "tightly" coupled juncture. We also find that forward coarticulation APs were generally more seriously bended than backward coarticulation ones. This conforms with the findings of [6]. Moreover, we find that the well-known 3-3 tone *sandhi* rule was properly learned by the proposed model. As shown in Fig. 3(b), the backward coarticulation APs of (3,3) were upward bended drastically. This can make the reconstructed F0 pattern of the first tone-3 in (3,3) resemble the standard tone-2 pattern. As for the coarticulation APs of duration and energy, they were all small and less influential.

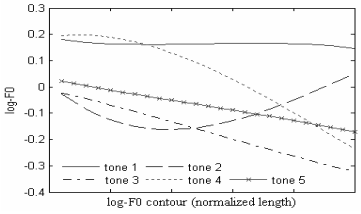

Fig. 2: The lexical tone APs of F0.

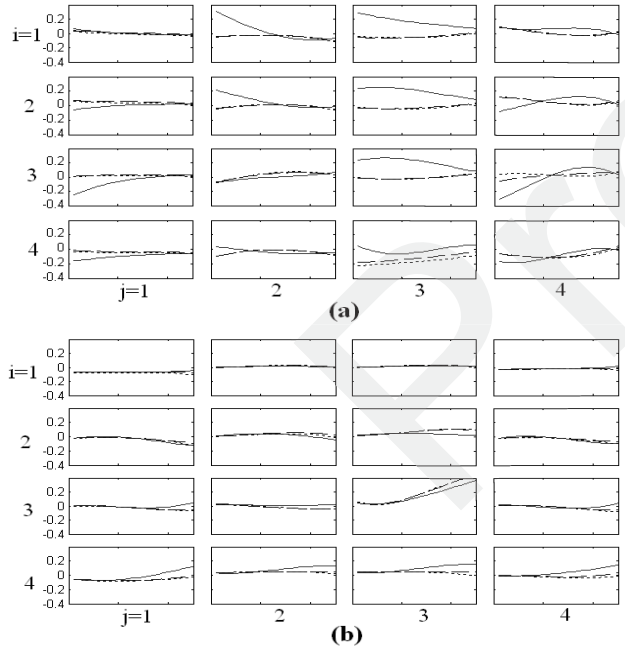

Fig. 3: The (a) forward and (b) backward coarticulation APs of F0 for 16 tone pairs and 3 coarticulation states. c1: solid line, c2: dash line and c3: dotted line;. tp_n=(i,j): tone pair.

Fig. 4 displays the syllable position-in-word AP of F0. It can be clearly observed from Fig. 4(a)~(d) that the syllable pitch level in all four types of polysyllabic word decreased to show the declination effect. We also find that the dynamic range of syllable pitch level increased as the word length increased. Moreover, we find that the patterns of APs for the first two syllables in 3- to 5-syllable words had similar shape with that of bi-syllabic word.

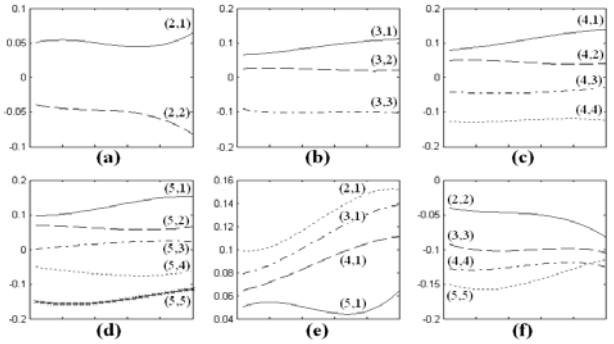

Fig. 4: The syllable position-in-word APs of F0 for (a)-(d) 2- to 5-syllabic words; and (e)/(f) comparisons of APs for the first/last syllables in polysyllabic words. (i,j) means the j-th syllable of i-syllabic word.

Fig. 5 displays the syllable position-in-word APs of duration and energy. It is found from Fig. 5(a) that the ending syllables in all four types of polysyllabic words were much longer than others to show the well-known lengthening effect. We also find from Fig. 5(b) that the syllable energy level of all four types of word tended to decline straightly to show the declination effect.

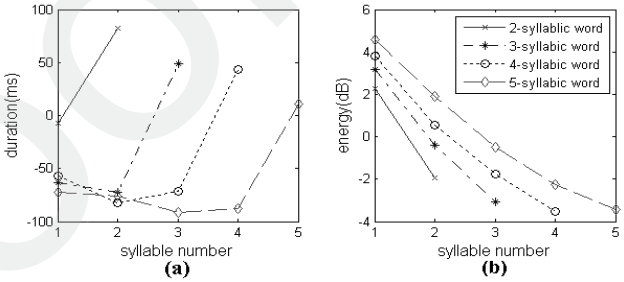

Fig. 5: The syllable position-in-word APs of (a) duration and (b) energy model.

We then analyzed the base syllable APs, γ_s and α_s, of duration and energy by CART. Some interesting phenomena were found: syllables with initial in {b,d,g} are shorter; syllables with nasal ending are longer; syllables pronounced by opening mouth have larger energy; syllables with single vowel have smaller energy; etc. These results match with the prior linguistic knowledge.

Table 2 displays the total residual error (TRE) defined as the ratio of sum-squared values of residual and observed features. It is found that TRE reduced as more APs were considered. Besides, it is found that lexical tone, position-in-word and base syllable had the most significant APs for the three models of syllable log-F0 contour, duration and energy, respectively.

Table 2: The performance (total residual errors, TRE) of the proposed prosody modeling method.

F0 Contour Modeling		Duration Modeling		Energy Modeling	
Affecting factors	TRE	Affecting factors	TRE	Affecting factors	TRE
+ Tone	53.9%	+ Position in word	55.9%	+ Base syllable	71.2%
+ Coarticulation	40.4%	+ Base syllable	35.3%	+ Position in word	46.9%
+ Position in word	28.9%	+ Tone	28.8%	+ Tone	42.5%
		+ Coarticulation	20.6%	+ Coarticulation	37.3%

3.2. Some examples of prosody prediction

Figs. 6 and 7 display some outside-test examples. It can be found from these two figures that most reconstructed features matched reasonably well with their original counterparts. This reveals that the proposed models are effective.

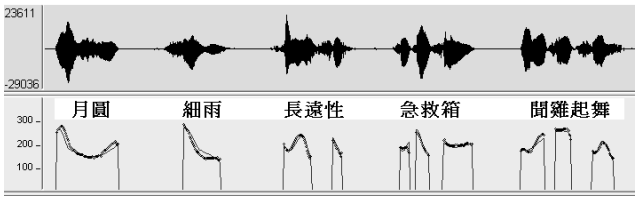

Fig. 6: Some examples of the observed (dots) and predicted (line) F0 contours. The text is "yue4 yuan2, xi4 yu3, chang2 yuan3 xing4, ji2 jiu4 xiang1, wen2 ji1 qi3 wu3".

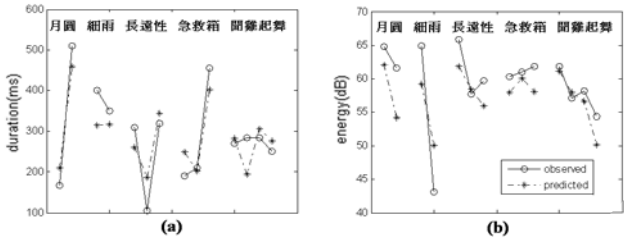

Fig. 7: Some examples of the observed and predicted (a) duration and (b) energy. The text is the same as that of Fig. 6

3.3 An application: A prosody pronunciation learning system

To illustrate the usefulness of the prosodic models learned, a learning system for Mandarin word prosody pronunciation is designed and implemented for nonnative speakers. Fig. 8 displays the interface of the system. It is generally acknowledged that to speak Mandarin with correct prosody is one of the major barriers to the learning of speaking Mandarin. In our system, the user can first input a Chinese word. The system will then generate the corresponding synthesized Mandarin speech. The target values of the three prosodic features (i.e., F0 contour, duration and energy) of all syllables will be displayed. The user can learn to speak Mandarin word via imitating the synthesized speech. The system will record the user's speech and on-line extract its prosodic features for display. The system can then generate a new synthesized speech by modifying the user's speech with all prosodic features except pitch level being changed to match the target values. So the user can hear his/her own voice with correct prosody. Lastly, we implement a new function to add a random value corresponding to the modeling error (i.e. the residual variance) to each target prosodic feature for increasing the variety of the synthesized speech.

It is worth to note that the prosodic model is used in the unit selection of the TTS system. The cost function of a candidate syllable is designed as the sum of squared errors between its prosodic features and the target values. The formulation is realized using the APs of our prosodic model. A finer synthesis unit can hence be found because both the coarticulation effect and the position-in-word are taken into consideration. Computational efficiency is another advantage of the method.

The performance of the system was informally evaluated by listening tests. Most synthesized speech sounded fluently and naturally for native mandarin speakers. Further evaluation by nonnative speakers will be done in the future.

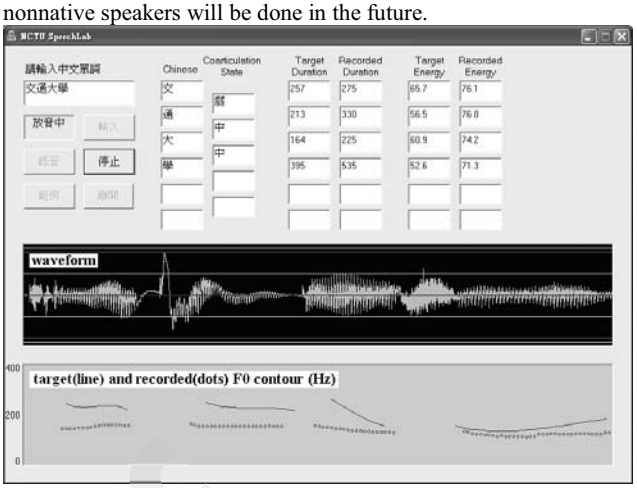

Fig. 8: The interface of the learning system for Mandarin word prosody pronunciation. The prosodic features of both targeted and recorded speech are shown on the screen.

4. CONCLUSIONS

In this paper, a syllable-based prosody modeling method of log-F0 contour, duration and energy for isolated Mandarin words is proposed. Four major APs, including lexical tone, inter-syllable coarticulation effect, syllable position in a word and base syllable type are considered. Experimental results on a large corpus confirmed the effectiveness of the proposed method. A learning system for Mandarin word prosody pronunciation is accordingly constructed for nonnative speaker. Improvement of the learning system is worthy doing in the future.

ACKNOWLEDGEMENTS

This work was supported in part by NSC under contract NSC 95-2221-E-009-057-MY3 and NSC 96-2218-E-002-001.

REFERENCES

[1] L. S. Lee, C. Y. Tseng, and M. Ouh-Young, "The Synthesis rules in Chinese Text-to-Speech system," *IEEE Trans. Acoust, Speech, Signal Processing*, vol.37, no.9, p1309-1319, Sep. 1989.

[2] S. H. Chen, S. H. Hwang, and Y. R. Wang, "An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech," *IEEE Trans. Speech and Audio Processing,* vol.6, no.3, pp.226-239, May 1998.

[3] C. Y. Chiang, H. M. Yu, Y. R. Wang and S. H. Chen, "An Automatic Prosody Labeling Method for Mandarin Speech," *Proc. of Interspeech*, 2007, pp. 494-497.

[4] S. H. Chen and Y. R. Wang, "Vector Quantization of Pitch Information in Mandarin Speech," *IEEE Trans. Communications*, vol. 38, no.9, pp.1317-1320, Sept. 1990.

[5] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkeley Press, University of California, Berkeley, CA, 1968.

[6] Y. Xu, "Contextual Tonal Variations in Mandarin,", *J. Phonetics*, vol. 25, no. 1, pp. 61-83, 1997.