Hierarchical stress generation with Fujisaki model in expressive speech synthesis

Ya Li [1], *Jianhua Tao* [1], *Keikichi Hirose* [2], *Wei Lai*[1, 3], *Xiaoying Xu*[1, 3]

[1] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] Department of Information and Communication Engineering, University of Tokyo, Japan
[3] Department of Chinese Language and Literature, Beijing Normal University, Beijing, China

yli@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn, hirose@gavo.t.u-tokyo.ac.jp,
laiwei_0508@126.com, xuxiaoying2000@bnu.edu.cn

Abstract

This paper introduces a hierarchical stress generation for expressive speech synthesis. In the previous study, we proposed a novel hierarchical Mandarin stress modeling method, and the text-based stress prediction experiments demonstrates a reliable stress assignment can be obtained from textual features. However, the stress model should be further verified to be an effective and efficient prosody model in a Text-to-Speech system. In this work, Fujisaki model known as an ideal global representation of prosody is adopted to construct the pitch contours. To illustrate the effect of stress model, the Fujisaki model parameters are automatically predicted by the textural feature with and without stress information. The synthetic speech sounds more natural than that without stress modeling. The RMSE of the pitch contour and the feature importance analysis also show stress information can improve the pitch modeling. This work offers a promising method to accurate pitch modeling for Mandarin expressive speech synthesis.

Index Terms: speech synthesis, Fujisaki model, stress, hierarchical modeling, pitch accent

1. Introduction

Expressive speech synthesis has gained a lot of attention recently because people are no longer satisfied with the flat synthetic speech in navigators, automatic call-center, and information broadcasting system etc. Therefore, the accurate modeling of prosody which can express the para-linguistic information, such as emotion, attitude, intentions, speaker characteristics and making the speech sound more vivid becomes particularly important. Stress is the perceptual prominence within words or utterances, and it constitutes the peaks and valleys of the pitch contours, which is an important factor of prosody.

Although previous work on stress realization which based on concatenation system [1, 2] can produce high quality speech, the expressiveness of the synthetic speech still relies on the audio corpus they used. Recently, HMM-based speech synthesis (HTS) draws growing attention for its flexibility in expressive speech synthesis. HTS-based stress generation can be categorized as direct modeling [3] and indirect modeling [4, 6], which mainly refers to prosodic parameter transformation.

The direct modeling is introducing the stress related question into the question set which is used in the HMM models clustering in HTS. However, the speech generated by this approach cannot convey stress clearly in HTS due to the weakness of emphasis/stress cues and statistical averaging effect of HTS [4]. Badino *et.al,* argue that more sophisticate context features should be designed to obtain a clear emphasis/prominence realization [5].

Regarding the indirect modeling, Yamafishi *et. al.*, [6] use speaking style interpolation and adaptation for HMM-based expressive speech synthesis. Maximum Likelihood linear Regression (MLLR) model is adopted in the style adaptation. Yu, *et.al.*, [4] utilize two-pass decision tree model and factorized decision tree model to extract word-level emphasis patterns from natural English speech, and then embed the emphasis model in the HTS framework. Although the speech generated by prosodic parameter transformation can convey stress effectively, it happened sometimes that a few syllables turned out too strong compared with the adjacent syllables, which makes the whole utterance sound unnatural. This indicates that the tradeoff between prominence and naturalness is hard to balance.

The ultimate goal of our work is synthesizing human-like expressive speech with stress. In the previous study, we proposed a novel hierarchical Mandarin stress modeling method [7]. The top level of this model emphasizes stressed syllables, while the bottom level focuses on unstressed syllables for the first time due to its importance in both naturalness and expressiveness of synthetic speech. The text-based stress prediction experiment demonstrates we can get a reliable stress assignment from textual features. However, the stress model should be further verified to be an effective and efficient prosody model in a Text-to-Speech system.

Therefore, generation process model of fundamental frequency contours known as Fujisaki model is adopted to generate pitch contours in this work. The reason of adopting Fujisaki model lies in two aspects. First, Fujisaki model is also a superpositional quantitative model for representing F0 contour of speech, which is perfect match the two-level hierarchical stress model we proposed, and thus we can directly control the hierarchical stress. Although some hierarchical pitch modeling methods [8-10] have already been proposed recently, the hierarchy is implicit modeled. Second, Fujisaki model is not only a parametric F0 contour stylization, but also has physiological and physical basis [11], and can well represent the long-term features than the commonly used frame-by-frame analysis method.

Some work on stress/prominence/focus realization has already been conducted with Fujisaki model. Kiriyama *et al.*, [12] and Chen *et. al.*, [13] implement a rule-base focus control with Fujisaki model. Ochi, *et.al.*,[14] also use Fujisaki model to control focus, but they predict the differences of the Fujisaki model commands between with and without focus utterances. The method is confirmed by the experiments. But, there are some constrains and limitations in the prosodic difference modeling.

10.21437/SpeechProsody.2014-195

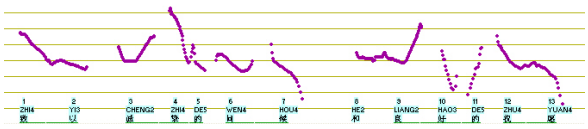
In this work, we want to testify whether the hierarchical stress model could improve the pitch modeling and how much improvement it will achieved. Therefore, we semi-automatically built a 500 sentence corpus with Fujisaki model parameters annotation as the first step. Then we constructed two Fujisaki model parameter prediction models using decision trees, among which, one utilizes the common features used in prosodic model parameter prediction, and the other introduces the stress information compared with the first model. Afterwards, the continuous pitch contours are generated by the two models. Listening test, objective experiment and features importance analysis are carried out. The results show stress model can improve the pitch modeling.

The rest of the paper is organized as follows. Section 2 introduces the hierarchical modeling method, including the hierarchical stress modeling and the superpositional modeling of F0 contour by Fujisaki model. Section 3 shows the details of text-based Fujisaki model parameter prediction with and without the hierarchical stress features. Experimental results and discussions are given in Section 4, and followed by the conclusion and future research in Section 5.

2. Hierarchical modeling

2.1. Hierarchical Mandarin stress model

Mandarin stress can be categorized as sentence stress and word stress from the range of their influence. Considering the importance of unstressed syllable in the naturalness and intelligibility of speech, a novel two-level Mandarin stress modeling method was proposed, in which, word level unstressed syllable investigation are emphasized for the first time, and in sentence level stressed syllables are studied as traditional methods do [7]. Fig. 1 shows a hierarchical stress assignment of a speech sample. First, the sentence stress is assigned onto words, and then each syllable's stress level is assigned within the word. The sentential stressed (denote as *3* in sentence stress) and unstressed syllable (denote as *1* in word stress) are the research focus of this model.



(a) Natural pitch contours

	zhi4yi3	cheng2zhi4de5	wen4hou4	he2	liang2hao3de5	zhu4yuan4
Sentence Stress	2	3	2	1	2	2
Word Stress	3 2	3 2 1	2 3	2	3 2 1	3 2

(b) Hierarchical stress assignment of speech sample (a)

Figure 1: Hierarchical Mandarin stress modeling.

2.2. Hierarchical pitch modeling with Fujisaki model

Fujisaki model is a command-response model that describe F0 contour as the superposition of the outputs of phrase and tone(/accent) commands [11].

Unlike most non-tone languages, which only have positive tone commands, Mandarin has both positive and negative tone commands. The tone command configuration for Mandarin can be found in [11]. The model (for tonal language) can be formulated by the following equations:

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} Ap_i Gp(t - T_{0i}) + \sum_{j=1}^{J} \{A_{1j}[Ga(t - T_{1j}) - Ga(t - T_{2j})] + A_{t2j}[Ga(t - T_{2j}) - Ga(t - T_{3j})]\} \quad (1)$$

$$Gp(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & for\ t \geq 0 \\ 0, & for\ t < 0 \end{cases} \quad (2)$$

$$Ga(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma], & for\ t \geq 0 \\ 0, & for\ t < 0 \end{cases} \quad (3)$$

where F_b , $G_p(t)$ and $G_a(t)$ are base fundamental frequency level, phrase commands and tone commands respectively, the detailed symbolic representations can be found in [11].

2.3. Integrating the two hierarchical models

Figure 2 illustrates the integration and relationship between Fujisaki model and the hierarchical stress model. The sentence level stress is corresponding to the phrase command in Fujisaki model, and the word stress is corresponding to the tone command. Inspired by this relationship, we utilize the sentence stress information to improve the phrase command prediction, and the word stress feature is introduced in the tone command prediction. We expect that the Fujisaki model commands can be estimated more accurately through this manner. Then we can evaluate the prosody of the synthetic speech to verify the hierarchical stress generation method.

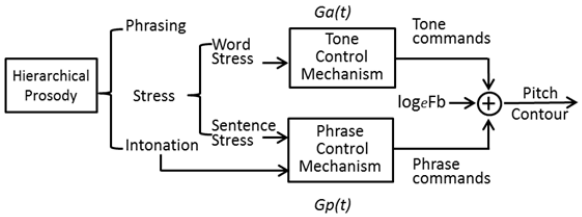


Figure 2: Integration and relationship between hierarchical stress model and Fujisaki model.

3. Fujisaki model parameter prediction

3.1. Corpus construction

In this work, we built a 500 sentences corpus, selected from the stress annotated corpus introduced in [7]. The Fujisaki model parameters are extracted automatically at first and then manually corrected. Fig. 3 is a sample of the Fujisaki model parameter labeling result through the FujiParaEditor [15].

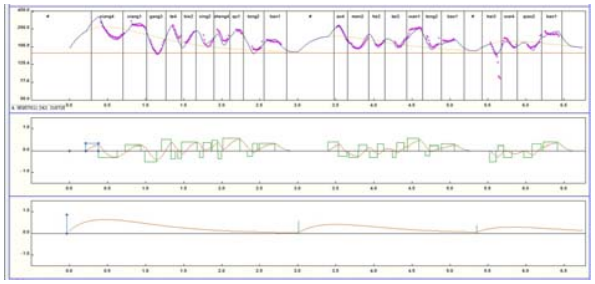


Figure 3: *Fujisaki model parameter annotation for Mandarin.*

The assignment of the phrase command usually coincides with the intonation phrase, but this is not a strict rule in the corpus labeling. The value of α, β, γ are the same as [11]. Ten percent of the data is reserved for testing.

3.2. Feature extraction

To verify the stress generation method, we designed two groups of textural features which are used in Fujisaki model parameter prediction. The first group includes the baseline features without stress information, and the second group includes the stress information as well as the baseline features. The sliding window is used for feature extraction in these two groups and the window size is five. For tone command prediction, the feature extraction unit is syllable, and for phrase command prediction, the unit is word.

As for the baseline feature set, different features are selected for the tone command and phrase command predictions. For the tone command prediction, the features are tone (indicate as *t*, hereinafter), syllable boundary (*bk*), the distance from the current syllable to the beginning/end of the sentence (*db/de*), the length of the word (*len*) to which the syllable belongs and its position in the word (*position*).

For the phrase command prediction, the long range context features are selected, namely, the Part-of-Speech of the word (*p*), the length of the word (*l*), the distance from the current word to the beginning/end of the sentence (*db/de*), the phrase length count by syllable and word respectively (*dis2sbsyl, dis2sbw*), and the index of the word in the intonation phrase to which it belongs (*posinphrase*).

In the second feature group, whether the syllable is unstressed or not is introduced in the tone command prediction and whether the word is sentential stressed is introduced into phrase command prediction. The windows for the stress related feature selection is also five. To verify the stress generation method, the hierarchical stress information is extracted from the annotated corpus rather than the automatic prediction from textual features.

3.3. Decision trees

According to Eq(1)-Eq(3), three parameters, with or without a phrase command, phrase command amplitude, and the command starting time, should be predicted from textual features. For tone command, five parameters, namely, two amplitudes and three timing values are selected as the targets. It should be noted that for tone 1, tone 3 and tone 5 syllables, one amplitude and one time value are set to zero. All the timing values are relative time, and are the offsets from the syllable(/word) starts time.

In the Fujisaki model parameters prediction, with or without a phrase command for each word is a binary classification problem, while the other commands, such as amplitude and command starting time, are continuous. J48 is adopted in the first task, and M5P is utilized for others which are both decision trees and implemented in Weka [16]. M5P combines a conventional decision tree with the possibility of linear regression functions at the nodes. In the preliminary experiments, we have tried several statistical models, M5P is slightly better than others in this study, such as C4.5 tree, J48 and linear regression.

4. Experiments and discussion

4.1. Pitch contours generation

The pitch contours are then generated by superimposing the phrase command and tone command responses. To simplify the work, the syllable duration and the base fundamental frequency level, F_b, are assumed to be the same with those in the training corpus. Then all the predicted time values can be converted to absolute values, and constitute a time sequence. Finally the synthetic speech can be generated by PSOLA algorithm which is implemented in Praat.

4.2. Experimental results

The average classification result for with or without a phrase command at a word's boundary is 73.69%. By introducing the hierarchical stress information, the average classification accuracy increases to 78.4%.

Table 1 shows the prediction results for the rest Fujisaki model parameters with continuous values, which only utilize the textual features. The first two rows are the results for phrase command predictions, including the phrase command amplitude A_{p1} and starting time T_0. The rest rows represent the tone commands prediction. It shows that the stress information can enhance the Fujisaki model parameter prediction, however, the improvement is small.

Table 1. *Text-based phrase command prediction using M5P decision tree.*

Model	Baseline (without stress)		With hierarchical stress feature	
	Corr.	RMSE	Corr.	RMSE
A_{p1}	0.85	0.18	0.85	0.17
T_0	0.81	106 ms	0.81	105 ms
A_{t1}	0.91	0.19	0.92	0.19
A_{t2}	0.94	0.11	0.94	0.11
T_1	0.58	43 ms	0.58	42 ms
T_2	0.80	34 ms	0.81	32 ms
T_3	0.57	54 ms	0.60	52 ms

Table 2. *Average RMSEs of utterance pitch contour predicted by models with stress and without stress.*

Experiment	RMSE (Hz)
without stress	46
with hierarchical stress	45

To further check the hierarchal stress generation, we align all the Fujisaki model parameters obtained by the automatic prediction, and combine the base frequency (F_b) to generate a continuous pitch contour. Table 2 shows the RMSEs between the natural speech and the F0 predicted by models with and without hierarchical stress. It also indicates that with stress information, pitch contour can be more accurately modeled. Fig. 4 shows pitch contours of two synthetic utterances generated by Fujisaki model. In this figure, the 8th syllable "*shang4*" is unstressed; the pitch is lower in the proposed pitch contours. On contrary, the 9th and 10th syllables"*fei1 chang2*" (In fact, they constitute a word, which means *very*.) are stressed, and "*fei1*" gets the final stress assignment through the

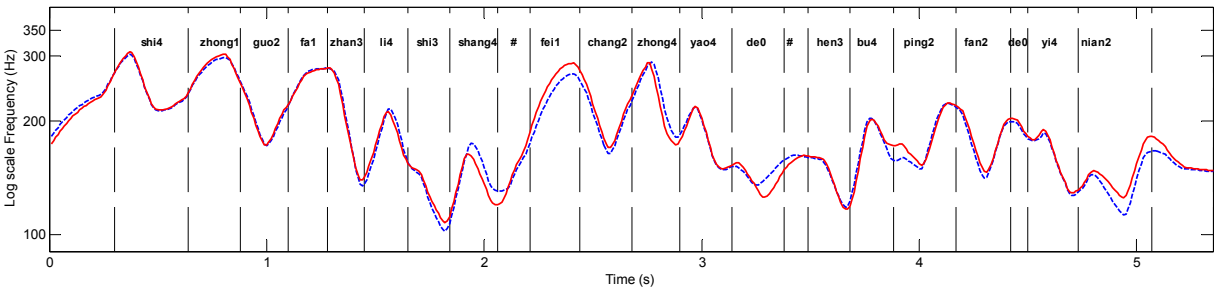


Figure 4: *Pitch contours generated by Fujisaki model with stress and without stress (line : with stress, dot: without stress).*

hierarchical mechanism, therefore, its pitch is higher in the proposed pitch contour. The informal listening test also shows the synthetic speech with stress model which using PSOLA algorithm sounds more natural. The over-salience syllables are hardly found in the synthetic speech. However, it should note that not all Mandarin tones are fully realized in connected speech, thus the variation of tone commands is complex and difficult to model using a decision tree which combines linear regression model in the nodes. Because of the tone command prediction error, some syllables sound unnatural, which makes the whole utterance sounds weird if there are too much tone command prediction errors. In such cases, the naturalness of the synthetic speech is worse than that generated by HTS. Nevertheless, the prominence and naturalness is well balanced in this method because the nature of command-response mechanism.

To evaluate the effect caused by hierarchical stress information in Fujisaki model parameter prediction, we also conducted a feature importance analysis by correlation feature selection [17]. Table 3 shows the feature selection results for each Fujisaki model parameter prediction. In this table, symbols, such as *p, l, t, bk,* represent the features introduced in Subsection 3.2. *s* represents the stress information. The number at the end of each feature denotes the offset in feature extraction. For example, *p_1* represents the Part-of-Speech of the previous word, and *s1* represents the stress information of the next syllable/word. Table 3 clearly illustrates stress information is indeed important in almost every parameter prediction.

Table 3. Feature importance in each Fujisaki model parameter prediction.(Y/N means with or without phrase command)

Prediction target	Feature importance (descending order)
Y/N	l_1, dis2sbw
A_{p1}	p_1, p0, l_2, l_1,dis2sbw,s_2,s_1,s0
T_0	p_2, p_1, p0, l_2, l_1, dis2sbw, s_2, s_1, s0
A_{t1}	t0, s0
A_{t2}	t0, bk0, db
T_1	t_2, t_1, t0, t1, bk2, bk0, s0
T_1	t0, s_1, s0, s1
T_3	bk2, s_2, s0

5. Conclusions

This paper introduces an attempt in Mandarin expressive speech synthesis by manipulating the stress generation with pitch contour generation process model (Fujisaki model). The Fujisaki model parameters are automatically predicted by the textural features with and without stress information. The hierarchical Mandarin stress modeling method is adopted to control phrase command and tone command correspondingly. And then the continuous pitch contours are generated and further evaluated. The experiments show hierarchical stress information can improve the pitch modeling both in global and local range. The advantage of the proposed method is it can make a good balance between prominence and naturalness compared with the previous direct and indirect stress/prominence modeling in HTS. This work offers a promising method to accurate pitch modeling for Mandarin expressive speech synthesis.

However, the accuracy of automatic Mandarin Fujisaki model parameter perdition needs further improvement, especially for the tone command. As reviewer suggests, a more comprehensive consideration of tones and how they interact with default focus, contrastive stress and different types/degree of stress in a sentence would be necessary. We believe that once this problem is solved, the MOS of synthetic speech can be greatly improved. Moreover the syllable duration variation in stress generation should be taken into consideration too. We will put more effort in these two fields in the future.

6. Acknowledgements

The author would like to thank anonymous reviewers for their valuable comments.

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61273288, No.61233009, No.61203258, No.61305003, No. 61332017, and No.61375027), and partly supported the Major Program for the National Social Science Fund of China (13&ZD189) and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM (CSIDM) Programme Office.

7. References

[1] W. Zhu, "A Chinese Speech Synthesis System with Capability of Accent Realizing," Journal of Chinese Information Processing, vol. 21, pp. 122-128, 2007.

[2] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, et al., "Modelling prominence and emphasis improves unit-selection synthesis," presented at the INTERSPEECH, Antwerp, Belgium, 2007.

[3] Y. Wu, "Research on HMM-based speech synthesis," Doctoral dissertation, University of Science and Technology of China, 2006.

[4] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010, pp. 4238-4241.
[5] L. Badino, J. S. Andersson, J. Yamagishi, and R. Clark, "Identification of contrast and its emphatic realization in HMM based speech synthesis," in INTERSPEECH, Brighton, UK, 2009, pp. 520-523.
[6] J. Yamagishi, T. Masuko, and T. Kobayashi, "HMM-based expressive speech synthesis–towards TTS with arbitrary speaking styles and emotions," in Proc. of Special Workshop in Maui (SWIM), 2004.
[7] Y. Li;, J. Tao;, and X. Xu, "Hierarchical Stress Modeling in Mandarin Text-to-Speech," in INTERSPEECH, 2011, pp. 2013-2016.
[8] Y. Qian, H. Liang, and F. K. Soong, "Generating natural F0 trajectory with additive trees," in INTERSPEECH, 2008, pp. 2126-2129.
[9] H. Zen and N. Braunschweiler, "Context-Dependent Additive log F0 Model for HMM-Based Speech Synthesis," in INTERSPEECH 2009, 2009, pp. 2091-2094.
[10] M. Lei, Y. Wu, F. K. Soong, Z. H. Ling, and L. Dai, "A Hierarchical F0 Modeling Method for HMM-Based Speech Synthesis," in INTERSPEECH, Makuhari, Chiba, Japan, 2010, pp. 2170-2173.
[11] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in Speech Prosody 2004, International Conference, 2004.
[12] S. Kiriyama, K. Hirose, and N. Minematsu, "Prosodic focus control in reply speech generation for a spoken dialogue system of information retrieval," in Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on, 2002, pp. 139-142.
[13] G. P. Chen, Y. Hu, R. H. Wang, and H. Mixdorff, "Quantitative analysis and synthesis of focus in Mandarin," in International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, 2004, pp. 25-28.
[14] K. Ochi, K. Hirose, and N. Minematsu, "Control of prosodic focus in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model," in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, 2009, pp. 4257-4260.
[15] H. Mixdorff, H. Fujisaki, G. P. Chen, and Y. Hu, "Towards the automatic extraction of Fujisaki model parameters for Mandarin," in Eighth European Conference on Speech Communication and Technology, 2003.
[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, vol. 11, 2009.
[17] M. A. Hall, "Correlation-based feature selection for machine learning," The University of Waikato, 1999.