Hierarchical Stress Modeling in Mandarin Text-to-Speech

Ya Li [1], Jianhua Tao[2], Xiaoying Xu [3]

[1, 2] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] Dept. of Chinese Language & Literature, PSC Center, Beijing Normal University, Beijing, China
{yli, jhtao}@nlpr.ia.ac.cn xuxiaoying2000@bnu.edu.cn

Abstract

Automatic stress prediction is helpful for both speech synthesis and natural speech understanding. This paper proposes a novel hierarchical Mandarin stress modeling method. The top level 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. Maximum Entropy model is adopted to predict stress structure from textual features. Experiments show that the modeling method could capture the macro- and micro-characteristics of stress successfully. The F-score of two-level stress predictions are 73.3% and 78.7%, respectively, which are satisfactory compared to other prosody predictions.
Index Terms: Text-to-Speech, prosody, stress, Mandarin

1. Introduction

Prosody is a super-segmental feature of speech and consists of rhythm, stress and intonation, among which, stress is a hot topic in recent years due to the growing demand for expressive Text-to-Speech (TTS). Stress (or pitch accent) is the perceptual prominence within words or utterances. When a syllable is stressed, its pitch goes higher and the duration becomes longer. It servers as one important feature in forming the ups and downs in a pitch contour, which makes the speech sound more expressive. However, Mandarin stress processing is a complicated problem. The difficulties lie in three aspects. First, Mandarin stress perception and labeling are tough work. In English, ToBI [1] is a well-accepted annotation system for prosody patterns of utterances. It defines prominence from the word's pitch movements or configurations and is easy to practice. Nevertheless, it is not the case for tonal language, such as Mandarin, in which each syllable has a tone and a relative steady pitch contour. Due to the frequent perceptual conflict among tone, intonation and stress, it is quite tough to find an exact definition of Mandarin stress, which makes stress labeling more difficult compared to other corpus labeling work. Li et al. [2] build a read corpus ASCCD with four-level stress annotation, and they report that the consistency in stress labeling is about 66%. However, the corpus is a discourse corpus and is not phonetic balanced, so it is not suitable for TTS. Wang et al. [3] have annotated stress for 300 independent utterances selected from Microsoft large-scale TTS speech corpus, nevertheless, the 300 utterances is insufficient for some machine learning methods for stress processing. The second obstacle for Mandarin stress processing involves that the automatic stress prediction from raw text is not up to our expectation. Although we can utilize acoustic features of speech or combine textual and acoustic features together to build an excellent stress detector [4-6], a prediction module just using textual features is still required for a TTS system. Shao et al. [4] utilize Artificial Neural Network (ANN) model to predict Chinese sentential stress using acoustic, textual features and both of them. The F-score for predicting stressed syllable just using text-based ANN model is only 36.1%. Ni et al. [5] compare the performance of Support Vector Machine (SVM), Classification and Regression Tree (CART) and AdaBoost with CART model for Mandarin stress prediction with acoustic, textual features and both of them. They argue that AdaBoost with CART can achieve favorable results than a single decision tree, and SVM is inferior to CART model. Similarly, Wightman [7] and Hirschberg [8] adopt decision trees to predict pitch accent in English and obtain encouraging results, which are above 80% in precision. The third bottleneck for Mandarin stress processing is even if we can assign stress from textual features appropriately, how to balance perceptual prominence and naturalness in synthesizing speech. Several latest investigations are done for this purpose, e.g., [9].

The ultimate goal of our research is synthesizing human-like speech with stress. Therefore, the above mentioned three difficulties must be solved one by one perfectly. This paper addresses the first two questions and leaves the third for next step. Firstly, we built a large stress annotated speech synthesis corpus, and then we proposed a novel hierarchical stress modeling method. Different from the current stress modeling methods which only focus on stressed syllables in a sentence, we enhanced unstressed syllable study in the bottom level, i.e., word level, meanwhile we also emphasized stressed syllable in the top level, i.e., sentence level. The motivation of our method is there are more and more unstressed syllables in natural speech. These unstressed syllables make the surrounding syllables stand out. In addition, when a syllable is unstressed, its word sense, even sentence meaning may change in some cases. Wu [10] and Cao [11] argue that the primary drawback of the current Mandarin speech synthesizer is insufficient handling of unstressed syllables as some unstressed syllables turn out to be too strong, which downgrade the naturalness and intelligibility of synthetic speech. Therefore we strengthen the unstressed syllable study and propose a hierarchical Mandarin stress modeling.

The rest of the paper is organized as follows. Section 2 introduces the hierarchical stress modeling method. Correspondingly, a hierarchical text-based stress prediction model under Maximum Entropy model framework is presented in Section 3. Experimental results and evaluations are given in Section 4, followed by conclusions and future research in Section 5.

2. Hierarchical stress modeling

Mandarin stress can be categorized as sentence stress and word stress from the range of their influence. We analysis the two level stress perceptually and then propose a hierarchical stress modeling method.

2.1. Sentence level stress

Sentence stress is the prominence within a sentence, while word stress describes a syllable's prominence within a prosodic word. Usually, sentence stress is firstly assigned to prosodic words, and then obtained by a syllable in the prosodic word.

From our experience, when hearing a whole utterance, people intuitively recognize a few prominent syllables, and then they will compensate for other less prominent or noise-masked syllables and phonemes using their knowledge of the spoken language, which is similar to phonemic restoration effect [12]. The top of these stressed syllables' pitch contours are often significantly higher in the whole utterance. Therefore, the sentence-level stressed syllables should be fully investigated for speech technology, especially for speech synthesis which aims to produce human-like natural speech.

2.2. Word level stress

In the scope of prosodic word, by contrast, the prominence difference within syllables does exist, but not so distinct. According to the prominence degree, word level stress is classified into three levels, namely, stressed, regular/normal and unstressed syllables.

Over the past decades, linguists pay much attention to stressed syllable and up to now it is still a controversial question. The growing demand for speech technology forces speech researchers switch their focus to non-stressed syllables. Because in daily life, people tend to pronounce words or syllables with little effort, thus bring about more and more weak syllables in real speech. These weak syllables have lower pitch and shorter duration, and they are also important to form the ups and downs in pitch contours. Nowadays there are approximately 20% weak syllables in real Mandarin speech, and even more in Beijing dialect. In this research field, neutral tone and unstressed syllable are two confused terms, because both of them are weak syllables. However, the majority of researchers suggest that unstressed syllable is different from neutral tone syllable. Neutral tone syllables are steadily weakened with a long history and their acoustic realizations are relatively stable regardless of the context. As a result, neutral tone is considered as the fifth tone except the four standard full tones in Mandarin. Most auxiliary words and suffixes, such as "le5, de5", are neutral tone syllables. Due to its stable pattern the neutral tone syllables are comparatively easy for speech processing. On the contrary, many unstressed syllables are quite new but frequently used in real speech. Part of the unstressed syllables only serve as weak syllables which make the surrounding syllables prominent. However, the others occasionally change the word sense when unstressed. For instance, when the second syllable of "dong1 xi1 (*east and west*)" is unstressed, the whole word sense is *thing*, and the acoustic realization also changes a lot. Nevertheless the regular and stressed syllables do not have this semantic function in word sense discrimination. With this reason, the unstressed syllable study should be strengthened. Automatic unstressed syllable prediction would benefit expressive speech synthesis and natural speech processing.

2.3. Two-level stress modeling

In summary, people only address a small number of syllables to make them significantly prominent in upper level (e.g., sentence level), and these sentence-level stressed syllables greatly improve the speech expressiveness. Meanwhile, at lower level (e.g., word level) the prominence syllable is restricted by upper level and cannot be distinctively perceived, thus people weaken other syllables to make that syllable stand out. Through such mechanism, the speech sounds more expressive and without too much physical effort. Additionally, people sometimes use unstressed syllable to express different meanings. Considering the importance of unstressed syllable in the naturalness and intelligibly of speech, we proposed a novel two-level Mandarin stress modeling method, 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.

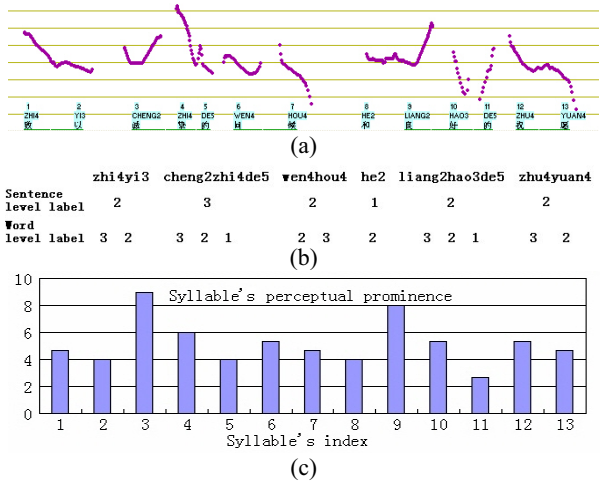

Figure 1: *Hierarchical stress annotation for utterance of "zhi3 yi3 cheng2 zhi4 de5 wen4 hou4 he2 liang2 hao3 de5 zhu4 yuan4" (Meaning: show sincere greetings and best wishes) (a) the pitch contour of the utterance, (b) two-stage stress label result by one annotator, (c) average perceptual prominence.*

Figure 1 demonstrates the proposed hierarchical stress modeling. First, a prominence degree is assigned to each prosodic word. Annotators could see the utterance pitch contour as shown in Figure 1(a) and play the whole utterance for several times. Second, all the prosodic words in the whole corpus are segmented and saved in separate files according to the word tone pattern. As a result, the words' orders are different from their original orders in utterance and the impact of adjacent words on the perception of the current word can be reduced. Afterwards, annotators are asked to assign a prominence degree to each syllable in a word. To simplify the labeling work, three discrete levels of prominence are adopted to describe the perceptual difference as far as possible, which are marked as 3 (stressed), 2 (regular) and 1 (unstressed). Figure 1(b) shows the two-stage labeling result by one annotator. We can combine the two-stage labeling results together to get a more specific prominent degree of each syllable in sentence level. Figure 1 (c) demonstrates one possible combination with the follow equation, averaged by three annotators,

$$PD_{syl}^{sen} = PD_{pw}^{sen} \times PD_{syl}^{pw} \quad (1)$$

where PD_{syl}^{pw} and PD_{syl}^{sen} are the prominence degree of each syllable in the prosodic word and the whole sentence, respectively. PD_{pw}^{sen} is the prominence degree of the prosodic word in the whole sentence. PD_{syl}^{sen} is the sentence stress studied in this work. Other weighted summing approach could also be applied under this framework.

A speech synthesis corpus which contains 6000 sentences with stress annotation is finally constructed with the hierarchical stress labeling approach. Three professional assistants are asked to label stress. After a period of training, they could achieve agreement on most of the annotations and keep consistency of their own during the labeling. The annotation consistency among them is about 72.5%.

3. Stress prediction from text

Automatic stress prediction from textual features has been widely studied due to its critical role in TTS. Many statistical machine learning techniques are introduced to this task.

3.1. Multiple textual features

We choose Maximum Entropy (ME) model to predict stress in this work which has been successfully applied to many prosody information predictions. ME model seeks the probability distribution with the maximum entropy subject to certain constraints. Such constraints force the model to match its feature expectations with those observed in the training data. Therefore, selecting the most effective and distinctive features can greatly improve the performance of ME model. Hirschberg [8] indicates that though the detailed syntactic, semantic and discourse level information can enhance the prediction of pitch accent, it is indeed possible to get a fair success with some automatically extracted features. Hence, only the shallow grammatical information which could be easily and reliably acquired from raw text is considered in this paper. The atomic text feature templates used in this work include: (The term *Word* means lexical word in this paper)

- PINYIN transcript (PY) and tone identity (T).
- Syllable's prosodic boundary (B).
- Part-of-Speech and the length of a word (P, L).
- Syllable description and the word, prosodic word descriptions (C, W, PW).
- Prosodic word length (PL).
- Normalized index of the current syllable in the prosodic word (RPW).
- Index of the syllable in current word and current prosodic word (IW, IPW).
- Distance from current syllable to the previous and the next word (DPW, DNW).
- Distance from current syllable to previous and next prosodic phrase (DPP, DNP).
- Distance from current syllable to the beginning and the end of the utterance (DB, DE).
- Distance from current word to the beginning and the end of the utterance (DBW, DEW).
- Distance from current prosodic word to the beginning and the end of the utterance (DBPW, DEPW).
- The stress ratio of the current syllable (SRC).
- The stress ratio of the current prosodic word (SRW) [5].

We also combine the above atomic feature templates to get more sophisticated feature templates and the sliding window method was adopted in feature extraction. For instance, feature template "B_1&B0&B1" means the previous, current and the next syllable prosodic boundaries and the number after the feature templates indicates the window offset. For the upper level stress prediction, we set the window width to 5 to capture the long distance relation. In lower level stress prediction, the window width is set to 3. We also use a wrapper method for selecting the effective feature templates to achieve better performance. The stop criterion is that the improvement of average correct rate is less than 0.1%.

3.2. Corpus

The final corpus used for stress prediction model training is presented in this subsection. As introduced above, the whole corpus contains 6000 sentences (about 34500 prosodic words and 72900 syllables) which are pronounced by a professional female speaker. Syllable boundary and three rhythm levels, namely, prosodic word, prosodic phrase and intonational phrase boundaries are manually labeled, too. The stress distribution is listed in Table 1. In this table, sentence level stress is counted by prosodic word and word level stress is counted by syllable. The combination result is calculated by Eq. (1) and then categorized into three levels, using the following equation,

$$Stress = \begin{cases} 3, & if\ PD_{syl}^{sen} \geq 6 \\ 2, & if\ 4 \leq PD_{syl}^{sen} < 6 \\ 1, & others \end{cases} \quad (2)$$

The combination data is finally used in sentence stress prediction and the word level stress model only utilizes the word level stress annotated data. The ratio of training set and testing set is 9:1. In this work, two ME prediction models are binary, namely, stressed and non-stressed classification in sentence level stress prediction, while unstressed and non-unstressed classification in word level stress prediction.

Table 1. *Stress distribution in the corpus.*

	3	2	1
Sentence level (by word)	24%	56%	20%
Word level (by syllable)	48%	39%	13%
Combine (by syllable)	48%	30%	22%

4. Experiments and discussion

The first experiments are conducted to select the most effective feature templates for word and sentence stress predictions. The two baseline systems utilize all the atomic feature templates mentioned above. The average correct rates of two models before and after template selection are given in Table 2. Although it is possible to combine the templates with prior knowledge to get more sophisticated feature templates, automatically selecting can achieve better performance with fewer templates. The original atomic templates are more than 80, and after two independent template selections only 10 templates are select for each stress prediction model. Table 3 shows the detailed selected templates.

The second experiments are performed to evaluate the performances of concerned stress types in each prediction model, and the results are shown in Table 4. For sentence level stress prediction, the F-score of stressed syllable is 73.3%, and for word level stress prediction, the F-score of unstressed syllable is 78.7%. In [13], we also use CART model to predict unstressed syllable, and the precision, recall and F-score are 86.3%, 56.3%, and 68.1% respectively. It implies that ME model performs better than CART in text-based stress prediction. As for the comparison between our results with the previous Mandarin stress prediction studies, no direct comparison could be presented because all other

related works only perform a single layer sentence stress prediction, not to mention the difference in corpus. In [5], Ni et al. construct several sentence stress prediction models and the best overall correct rate which using Boosting method with CART is above 80%. However, the corpus they used ASCCD is a discourse corpus and there is no detailed information of stressed syllables prediction result in [5].

Table 2. *Average correct rate of two stress prediction models before and after feature selection (FS).*

Experiment	Before FS	After FS
Word level	92.7%	94.1%
Sentence level	70.1%	75.9%

Table 3. *Selected feature templates for word-level unstressed syllable and sentence-level stressed syllable prediction.*

	Word level	Sentence level
1	B0&DEW	B0&SRW0
2	B1&PW0	B1&T0&T_1
3	DB&SRW0	C0&RPW0
4	DNP0&DNP1	DBPW&SRC0
5	DNP0&PL0	DNP_2&PW0
6	DNP_1&PL0	DPP_1
7	DNP_1&P_1	PL2&PW0
8	PL1&SRC0	T0&T_1
9	T0	P1&PY1
10	T0&T1	P_1&T1

Table 4. *Stress predictions performance.*

Experiment	Precision	Recall	F-score
Word level	94.2%	67.6%	78.7%
Sentence level	73.1%	73.4%	73.3%

The two level stress labels are helpful to describe the long term and short term prosodic characteristics of each syllable as far as possible. In addition, the experimental performances demonstrate the feasibility of the proposed hierarchical stress modeling. With the relatively ideal stress assignation, the next step for stress generation in a TTS system could be carried out. Fortunately, the latest pitch modeling research which utilizes the similar hierarchical approach for generating natural speech [14-16] has already shown its effectiveness. The future work will take the advantage of the proposed hierarchical stress modeling (serve as the front-end of a TTS system) as well as the hierarchical prosodic parameter generation [14-16] (serve as the back end of a TTS system) to get more natural and expressive speech.

5. Conclusions and future works

The ultimate goal of our work is synthesizing human-like speech with stress. In this paper, we proposed a hierarchical stress modeling method to get a fine-grained stress description. In the model, sentence level stressed syllables are studied as traditional methods do, while in word level, we emphasized unstressed syllable research for the first time due to its importance in both naturalness and expressiveness. According to this architecture, a two-level stress prediction model under Maximal Entropy framework was constructed. Experiments showed that the proposed method could obtain a fine-gained stress structure description reliably.

The future work will focus on Mandarin stress generation which is the third bottleneck for Mandarin stress processing. Our preliminary work on synthesizing speech with stress which utilizes a stress adaptation approach is verified to be effective. The only drawback lies in that sometimes a syllable is too strong compared to surrounding syllables. We can expect that this weakness could be overcome through introducing the hierarchical stress modeling method proposed in this paper. Future research will also cover seeking alternative techniques to generate speech with stress.

6. Acknowledgements

The work was supported by the National Science Foundation of China (No. 60873160, 61011140075 and 90820303) and partly supported by China-Singapore Institute of Digital Media (CSIDM), Beijing Normal University (004-127028) and the Funda-mental Research Funds for the Central University.

7. References

[1] Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J. and Hirschberg J., "ToBI: A standard for labeling English prosody", Proc. ICSLP, 867-892, 1992.

[2] Li A., Chen X., Sun G., Hua W., Yin Z., Zu Y., Zheng F. and Song Z., "The phonetic labeling on read and spontaneous discourse corpora", ICSLP, 724-727, 2000.

[3] Wang,Y., Chu, M. and He L., "Location of sentence stresses within disyllabic words in Mandarin", in the Proceedings of the 15th ICPhS, Barcelona, 1827-1830, 2003.

[4] Shao Y., Han J., Zhao Y. and Liu T., "Study on automatic prediction of sentential stress for Chinese Putonghua Text-to-Speech system with natural style", Chinese Journal of Acoustic, 26(1):49-92, 2007.

[5] Ni C., Liu W. and Xu B., "Mandarin pitch accent prediction using hierarchical model based ensemble machine learning", IEEE Youth Conference on Information, Computing and Telecommunication, YC-ICT '09. Beijing, 327-330, 2009.

[6] Ananthakrishnan S. and Narayanan S. S., "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence" , IEEE Trans on Audio, Speech, and Language Processing, 16(1):216-228, 2008.

[7] Wightman C. and Ostendorf M., "Automatic labeling of prosodic patterns", IEEE Trans. on Speech and Audio Processing, 2(4):469–481, 1994.

[8] Hirschberg J., "Pitch accent in context: Predicting intonational prominence from text", Artificial Intelligence, Vol.63, 305-340, 1995.

[9] Yu K., Mairesse F. and Young S., "Word-level emphasis modeling in HMM-based speech synthesis", ICASSP 2010. Dallas, March, 4238-4241.

[10] Wu Z., "Research on naturalness related prosodic variables in Mandarin speech synthesis", the 5th phonetic conference of China, 291-294, 2001. (in Chinese)

[11] Cao J., "Mandarin stress structure judging from speech synthesis point of view", Report of international symposium for commemorating the 100 birthday of Lv Shuxiang, June, 22-23, 2004, Beijing. (in Chinese)

[12] Timothy B. J., "The psychology of language", Prentice Hall, 2002.

[13] Li Y., Tao J., Zhang M., Pan S. and Xu X., "Text-based unstressed syllable prediction in Mandarin", INTERSPEECH 2010, Makuhari, Japan, 1752-1755.

[14] Qian Y., Liang H. and Soong F. K., "Generating natural f0 trajectory with additive trees," INTERSPEECH 2008, 2126-2129.

[15] Zen H. and Braunschweiler N., "Context-dependent additive log f0 model for HMM-based speech synthesis," INTERSPEECH 2009, 2091-2094.

[16] Lei M., Wu Y., Soong F. K., Ling Z. and Dai L., "A hierarchical f0 modeling method for hmm-based speech synthesis," INTERSPEECH 2010, 2170-2173.