Text-based Unstressed Syllable Prediction in Mandarin

Ya LI [1], Jianhua TAO[2], Meng ZHANG[3], Shifeng PAN[4], Xiaoying XU[5]

[1,2,3,4] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
[5] Beijing Normal University, 100875, Beijing, China
{yli[1], jhtao[2], mzhang[3], sfpan[4]}@nlpr.ia.ac.cn xuxiaoying2000[5]@bnu.edu.cn

Abstract

Recently, an increasing attention has been paid to Mandarin word stress which is important for improving the naturalness of speech synthesis. Most of the research on Mandarin speech synthesis focuses on three stress levels: stressed, regular and unstressed. This paper emphasizes the unstressed syllable prediction because the unstressed syllable is also important to the intelligibility of the synthetic speech. Similar as the prosodic structure, it is not easy to detect stress from text analysis due to the complicated context information. A method based on Classification and Regression Tree (CART) model has been proposed to predict the unstressed syllables with the high accuracy of 85%. The method has been finally applied into the TTS system. The experiment shows that the MOS score of synthetic speech has been improved by 0.35; the pitch contour of the new synthesized speech is also closer to natural speech.

Index Terms: Text-to-Speech, stress, unstressed syllable, prosody

1. Introduction

Recently, an increasing attention has been paid to stress processing to improve the naturalness and expressiveness of speech synthesis. Some statistical methods have been successfully applied in this area (e. g. [1, 2, 3, 4, 5]). For instance, Wightman and Ostendorf [1] used decision trees with acoustic and lexical information to classify pitch accent (In English, the "pitch accent" or "accent" are often used with similar sense to the "stress" in Mandarin), obtaining accuracy of approximately 84%. Hirschberg [2] found that though the detailed syntactic, semantic and discourse level information can enhance the prediction of stress, it is indeed possible to model accent with fair success. They extract textual features automatically and get an overall prediction precision of about 80%-98%. Kim and Lee [3] used a conditional Maximum Entropy classifier to predict the pitch accent with six C-ToBI pitch accent types in Mandarin. They achieved 70.1% of accuracy with linguistic features only. Shao et. al. [4] compared the acoustic features among syllables that have different stress levels and proposed an artificial neural network model to predict the Chinese sentential stress. They reported that using both acoustic and textual features can get a better performance than that using any one alone.

Stress is a super-segmental feature. As a tonal language, it is not easy to find an exact definition of the stress in Mandarin speech due to the frequent perception confliction among tone, intonation and stress. To simplify the work, most of researchers limited their works for classifying the stress into three levels: stressed, regular and unstressed according to the syllable's prominence degree in a word or in a sentence [4, 6], leaving the characteristics of the three levels of stress not very clear.

Among the three stress types, the unstressed part not only influences the naturalness of the synthetic speech, but also has a special impact on the word sense, the part-of-speech (POS) or the syntactic function of the word. For instance, when the second syllable of "dong1 xi1 (east and west)" is unstressed as "dong1 xi0 (thing)" ("0" indicate the current syllable is unstressed), the whole word sense is changed. Nevertheless, not all the unstressed syllables would change the word sense, only part of them does. For instance, the words "xin1 xian1" and "xin1 xian0" both mean something is fresh. We can also easily find many corresponding examples in English, such as PROject and proJECT (The capital letters indicate the lexical stress. Precisely, this difference is caused by the placement of stress because English is a stress language). However, the regular and stressed syllables do not have this function.

With this reason, the paper focuses on the unstressed syllable processing within prosodic word for Mandarin speech synthesis and taking the regular and stressed syllables as one category which is not stressed syllable. Several works [4, 6] have proved that the stress is various among different context information, and when a syllable is stressed, probably, the duration is longer and the pitch range is wider than the unstressed syllable. Intensity is not a critical cue in identifying stress. To solve this problem in Mandarin, the paper tries to analyze the acoustic impact of unstressed syllables on the prosodic structure and proposes a CART-based model to predict the locations of unstressed syllables. Five kinds of text features were used in this model, including the initials, finals, tone and position information of a syllable, as well as the POS identity of the word. The experiments show that the accuracy of the unstressed syllable prediction is about 85%. The method has been finally applied into the Wiston TTS system [7]. The Mean Opinion Score (MOS) of synthetic speech is improved by 0.35, and the pitch contour of the new synthesized speech is closer to natural speech than the previous speech synthesized by the system without the stress prediction model.

The rest of this paper is organized as follows. Section 2 provides a brief analysis to the influence of the unstressed syllable on the prosodic structure. Section 3 introduces the method of unstressed syllables prediction with a CART-based model. Section 4 presents the experimental results and evaluations. Section 5 discusses the conclusion and future research.

2. Unstressed syllable

The terms "unstressed syllable" and "neutral-tone syllable" are often wrongly treated as interchangeable. To clarify, the unstressed syllable in this paper is slightly different from the neutral tone syllable. From the linguistic point of view, there are many distinctions between them, including the definitions.

However, it is generally accepted that both of them are weak syllables. Yuenren Chao argues that there are static weakened syllables and provisional weakened syllables. Most auxiliary words and suffixes, which have a small semantic value or with a purely grammatical function, are steadily weakened in speech [8]. In some research, the former is defined as a neutral tone syllable and the latter is defined as an unstressed syllable [9]. In this paper, Pinyin transcript with tag "5" indicate the syllable is a neutral tone syllable, while Pinyin with tag "0" indicate the syllable is unstressed. The neutral-tone syllables in Mandarin do have some relatively stable patterns in acoustic realization [10]. Figure 1 shows four pitch contours of the neutral tone syllable "le5" from four different natural utterances. Although the contexts of these samples are different, the pitch contours of these syllables seem alike. Therefore, synthesizing them with high naturalness is not a problem in a large corpus-based concatenation TTS system.

However, it is not the case for the unstressed syllable, whose acoustic realization varies as the context changes. The following example demonstrates the influence of an unstressed syllable in word sense and their prosodic feature.

In Chinese, the words "mai3 mai4 (buy and sell)" and "mai3 mai0 (business)" have the same corresponding Chinese characters, but they pronounced differently. Figure 2 shows four pitch contours of the "mai3 mai4" and "mai3 mai0" in natural speech, among which two "mai0" are unstressed, and the other two labeled as "mai4" are not. The four "mai3" are not unstressed. We can see that the acoustic features can change a lot when a syllable is unstressed. The pitch is dropping, the pitch range becomes narrower and duration becomes shorter. Moreover, the neighboring syllables are also weakened due to the unstressed syllable, but not that much. The unstressed syllables have more than one clearly distinguishable type of pitch contours, which are influenced by the context, e.g., tone, intonation, rhythm level etc., and should be synthesized accordingly in a TTS system.

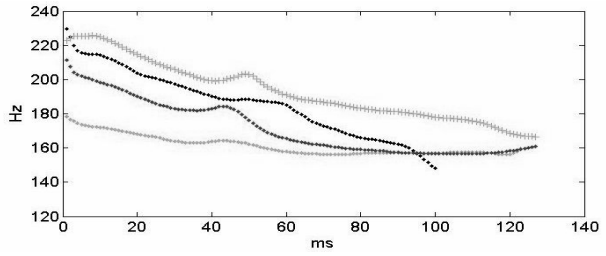

Figure 1: *The pitch contours of "le5" in natural speech.*

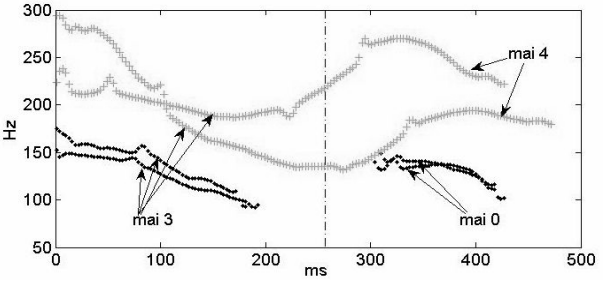

Figure 2: *The pitch contours of the "mai3 mai4" and "mai3 mai0"in natural speech. (the vertical dash line represents the rough boundary between syllables. Note that the syllable boundaries are not exactly the same in these four words.)*

Considering the importance of unstressed syllables in the naturalness and intelligibility of the synthetic speech, the paper will emphasize on the prediction of unstressed syllables and adopt a one-versus-others classification scheme, which is the unstressed syllable versus regular and stressed syllables. A CART model with textual features only is built in this work.

3. Unstressed syllable prediction

3.1. Corpus

The audio corpus used in this work contains 6000 sentences which are read by a professional female speaker. Three professional research assistants are asked to do the data annotation by both listening to the utterances and reading the text transcriptions. Before the annotation, they were trained several times. At the training step, they were required to discuss criteria for annotation so that they could achieve agreement on most of the annotations and keep consistency of their own during the labeling. A sample sentence is as follows.

xu3 duo1$kao4#jie4 kuan3|jin4 xing2#mai3 mai0 de5#ren2, tu1 ran2 jian4|bian4 de5#shen1 wu2#yi4 wen2|shen4 zhi4#po4 chan3$
(Many people who run business relying on loan became impecunious or even bankruptcy.)

In this sample, the first line is the Pinyin script. Tag "1"-"4" are four tones in standard Mandarin, tag "5" is the neutral tone and tag "0" is the unstressed syllable. The amount of unstressed syllables is 4688, which accounts for about 4% of the total syllables. The prosodic words, prosodic phrases and intonational phrases were separated by a "#", "|" and "$", respectively.

The whole corpus is divided into the training and test sets according to a 9:1 ratio randomly. In the training process, the training data is divided into 10 parts and a 10-fold cross validation is conducted.

3.2. Multiple linguistic features

Prosody is usually related to syntactic structure, but due to the lack of sophisticated syntactic parser, only the shallow grammatical information which could be acquired easily and reliably is considered in this work.

POS is the most commonly used feature in prosodic structure prediction. The syllable position in a word is also important because the last syllable in a word is more likely to be unstressed. Also the initials, finals and tone of a syllable contain many phonetic features and should not be neglected in this study. These features are automatically extracted by the front-end of Wiston. An accuracy of 96% for word segmentation and 91% for POS tagging were achieved [11].

Table 1. *Text-based feature templates.*

Feature template	# of classes
$initial_{-n}, initial_{-n+1}, ..., initial_{n-1}, initial_n$	23
$final_{-n}, final_{-n+1}, ..., final_{n-1}, final_n$	43
$tone_{-n}, tone_{-n+1}, ..., tone_{n-1}, tone_n$	5
$POS_{-n}, POS_{-n+1}, ..., POS_{n-1}, POS_n$	44
position in a word	4

Table 1 shows the textual feature templates used in this model. A sliding window method is adopted here, and n is the half-width of the window. The second column indicates the number of classes for each type of feature templates used in this model. The 23 initials are "b", "f", "m", "ch", etc.. The 43 finals are "a", "ao", "ai", "ang", "uang", etc.. The five tones also include the neutral tone apart from the four standard tones in Mandarin. The POS set used in this work is the PKU-POS set which is defined by Peking University. The four values of the position in a word is "0 (monosyllabic word)", "1 (the first syllable of the multisyllabic word)", "2 (the middle syllable of the multisyllabic word)" and "3 (the last syllable of the multisyllabic word)".

4. Result and discussion

4.1. Prediction results

In the prediction process, we examined the effect of two different window sizes, [-1, 1] and [-2, 2] to predict whether a syllable is unstressed. The performance measure in this test was simply defined as:

$$precision = \frac{c}{M}, \quad recall = \frac{c}{N} \tag{1}$$

where,

c is the number of correctly predicted unstressed syllables, N is the number of unstressed syllables in test corpus, and M is the number of predicted unstressed syllables.

Table 2 shows the prediction results. The window size [-1 1] performs slightly better than the other. It indicates the features of previous and next syllables may influence whether the current syllable is unstressed greatly. When the scope of predictors increases, noise data are added and the performance decreases. From the CART result, the POS_{-1}, position in word and $initial_0$ are the top three important predictors in this task. This is consistent with our prior knowledge and therefore useful to choose the efficient predictors to make the model compact and precise. The result shown in Table 2 also shows that precision is higher than recall. As explained before, some of the unstressed syllables do not have the syntactic function in a word; therefore, it is no need for people to weaken this syllable strictly to differentiate the meaning of the sentence. People may utter these words under different moods. These kind of unstressed syllables are highly related to the intonation and emotion of a sentence and it is hard to predict with shallow morphology information only. However, the other unstressed syllables play a grammatical role in a word. Considering this, the precision is of great importance in this task.

Table 2. *Performance in each window size.*

Window size	[-1, 1]	[-2, 2]
Precision	86.3%	85.6%
Recall	56.3%	55.6%

4.2. Test in Wiston TTS system

In order to evaluate how much improvement of naturalness and intelligibility of the TTS system with the unstressed syllable prediction model can achieve, we applied this model to Wiston and then evaluated the synthesized speech. The original Wiston system is the same as the system reported in [7], while the new Wiston system is not only embedded with the unstressed syllable prediction model, but also with a new audio corpus from which the candidate syllables are selected. In the reconstruction of the new audio corpus, whether the syllable is unstressed or not is taken into consideration. In the following, objective evaluation and listening test are carried out.

In the objective evaluation, we compared the average distance of pitch contour between the speech synthesized by original Wiston system and the new Wiston with natural speech. Ten parallel utterances were generated with two Wiston systems, and their pitch contours were extracted by STRAIGHT [12]. To reduce the impact of the difference in syllable duration, the average distance of pitch contours were computed through scaled pitch contours between the synthesized speech and natural speech, which were read by the same speaker. The distance is defined as follows:

$$Dis = \frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{N_k}|syn\_pitch(j') - pitch(i)| / N_k \tag{2}$$

$$j' = (j - s_k) \times \frac{syn\_duration_k}{duration_k} + s_k' \tag{3}$$

where,

$pitch(i)$ is the i[th] pitch point of the current syllable in the natural pitch contour, and $syn\_pitch(j')$ is the j' [th] pitch point of the current syllable in the synthetic pitch contour. $duration_k$ and $syn\_duration_k$ are the duration of the k[th] syllable in the natural speech and synthetic speech respectively. s_k and s'_k are the start pitch point index of the k[th] syllable in the natural speech and synthetic speech respectively. j' is the scaled pitch point index of the current syllable. N_k is the total number of the pitch points of k[th] syllable in the utterance. M is the total syllable number of the utterance.

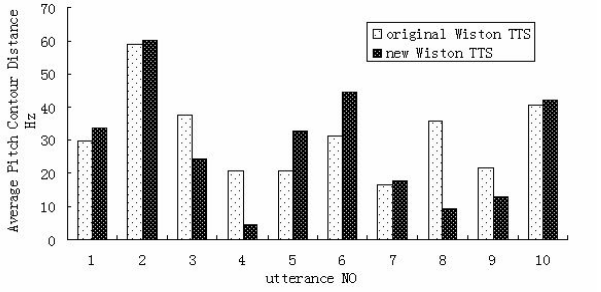

Figure 3: *The average pitch contour distance between two synthesized speech and the natural speech.*

Table 3. *MOS score of two synthesized speech.*

	Original Wiston	New Wiston
MOS score	3.86	4.21

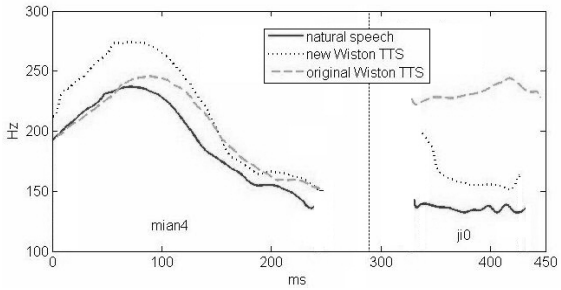

Figure 4: *Three pitch contours of the word "mian4 ji0". (the vertical dash line represents the rough boundary between syllables. Note that the syllable boundaries are not exactly the same in these words.)*

The average pitch contour distance declined 4 Hz from 32Hz to 28 Hz with the unstressed syllable prediction model. Figure 3 shows the each average pitch distance of the ten utterances. Some utterances synthesized by the original Wiston system seem slightly better, which may be due to the data sparseness problem in corpus-based system. In unit selection part, the contextual feature templates of current unit have been enriched in the new Wiston system, therefore the amount of candidate units of the syllables before and behind the syllable that is unstressed becomes less. Moreover, the number of unstressed syllable samples is much less than the syllables which are not unstressed. Thus it is harder to select an appropriate unit in the new Wiston, and the overall average pitch distance may become larger. We expect that this disadvantage can be overcome by increasing the speech corpus size. Figure 4 shows the pitch contours of "mian4 ji0". We can see that the pitch contour of synthesized speech in new Wiston system is closer to natural speech.

Finally, in the listening test, seven speech experts were asked to rate these sentences with MOS score, which is from 1.0 to 5.0. As shown in Table 3, the MOS score is improved by 0.35, which indicates the proposed model offers better overall performance than the original system.

5. Conclusion and future works

Automatic detection and prediction of Mandarin word stress is very useful and important to many spoken language processing tasks such as Text-To-Speech. Although study on stressed syllables seems straightforward and instinctive, unstressed syllables are critical to the intelligibility of synthetic speech, and they also augment the "prominence degree" among syllables, making the speech sounds more natural and expressive. In addition, the unstressed syllable is much complex in both the grammatical function and their acoustic realization, which makes the unstressed syllable prediction a hard work but very useful in speech synthesis. To solve this problem, this paper investigates the impact of the unstressed syllable on the prosodic structure and a CART-based model which only utilizes textual features is proposed for predicting the unstressed syllables automatically. The prediction precision is above 85%. We also evaluate the performance of the TTS system with the proposed model. MOS score is improved from 3.86 to 4.21, and the average pitch contour distance between new synthesized speech and natural speech also decreases.

The future work will focus on the followings. Firstly, we will try more effective text features and acoustic features to improve the prediction model and enhance the acoustic module of the Wiston system to synthesize the stressed and unstressed syllables. Our other preliminary work confirmed with [4] in that the acoustic features, e.g., pitch and duration have strong relations with stress. This will help us to enhance the stress prediction model which would be used in the automatic corpus labeling. Secondly, we will integrate the stress prediction model into the HMM-based speech synthesis system. There are two reasons; one is the data sparseness problem in corpus-based system. Although we can obtain a text-based stress prediction model with very high precision, the performance of the synthetic speech still relies on the acoustic realization of the acoustic module. To solve this problem, a very much larger audio corpus is needed which is not very easy to collect in a short period of time. So we will try to modify the pitch and duration of a syllable according to their stress level. This is easier to construct and would be a possible way to deal with the over-smoothing in HMM-based speech synthesis, which results in a low expressive synthetic speech. Thirdly, we will extend it from two-class stress to three-class stress model to further improve the expressiveness of the synthetic speech in the condition that a good performance of the unstressed syllable prediction has been achieved.

6. Acknowledgements

The work was supported by the National Science Foundation of China (No. 60873160), 863 Program (No. 2009AA01Z320) and China-Singapore Institute of Digital Media (CSIDM). The authors would also like to thank Dr. Sim Khe Chai for his very helpful comments and suggestions.

7. References

[1] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns", IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, 1994, pp. 469–481.

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

[3] B. Kim and G.G. Lee, "C-TOBI-Based Pitch Accent Prediction Using Maximum-Entropy Model", Lecture notes in computer science, 3982, 2006, pp. 21–30.

[4] Y., Shao, J. Han, Y. Zhao and T. Liu, "Study on automatic prediction of sentential stress for Chinese Putonghua Text-to-Speech system with natural style", Chinese Journal of Acoustic, Vol. 26 No.1, 2007, pp. 49-92.

[5] P. Zervas, N. Fakotakis and G. Kokkinakis, "Pitch accent prediction from ToBI annotated corpora based on Bayesian learning", 2004, Lecture notes in computer science, SKP04, pp. 545-552.

[6] J. Xu, M. Chu, L. He, and S. Lu, "The influence of Chinese sentence stress on pitch and duration", Chinese Journal of Acoustics, Vol. 25, No.4, 2000, pp.335-339.

[7] J. Tao, J. Yu, L. Huang, F. Liu, H. Jia and M. Zhang, "The WISTON Text to Speech System for Blizzard 2008", The Blizzard Challenge 2008 workshop, Oct.2008.

[8] Y. Chao, Linguistic essays by Yuenren Chao, The commercial press, Beijing, 2006.

[9] J. Lu and J. Wang, "On defining 'qingsheng'", Contemporary Linguistics, 7(2), 2005, pp. 107-112. (In Chinese)

[10] J. Cao, "On neutral-tone syllables in Mandarin Chinese", Canadian Acoustics, No.3, 1992.

[11] F. Liu, H. Jia and J. Tao, "A Maximum Entropy Based Hierarchical Model for Automatic Prosodic Boundary Labeling in Mandarin", The 6th International Symposium on Chinese Spoken Language Processing, ISCSLP2008, Kunming, pp.257-260.

[12] http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTtrial