HMM-based Speech Synthesis with a Flexible Mandarin Speech Stress Adaptation Model

Ya Li, Shifeng Pan, Jianhua Tao
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
{yli, sfpan, jhtao}@nlpr.ia.ac.cn

Abstract—Expressive speech synthesis has recently received much attention. Stress is one key issue which may improve the expressiveness of the synthetic speech. However, rare work was done in Mandarin stress prediction and expression. This paper presents a HMM-based expressive speech synthesis system which supports Mandarin stress synthesis. Mandarin stress was automatically predicted with textual features only using a Maximum Entropy Model. The linear adaptation model was extracted from a large corpus by analyzing their stress related acoustic features. The advantage of the proposed model is it can be easily modified to build a system with another speaking style or emotion. Experiments show that the proposed stress adaptation system can convey stress effectively and generate high expressive speech. The overall performance of the synthetic speech is also improved.

Keywords prosody; Mandarin stress; HMM-based speech synthesis; expressive synthesis;

I. INTRODUCTION

The study on the naturalness and intelligibility of synthetic speech has made a great progress in recent decades with the help of corpus-based technique and PSOLA algorithm. The main disadvantage of concatenation system is that the speaking style can hardly be changed, which limits its expressivity. Stress is the perceptual prominence within words or utterances, and it is an important aspect of prosody. Stress constitutes the ups and downs in an utterance and makes it sounds more expressive. (In English, the “pitch accent” or “accent” is often used with similar sense to the “stress” in Mandarin.) However, most of the research focused on the analysis of stress and a few works have been done in stress synthesis. Y., Shao, J. Han, Y. Zhao and T. Liu [1] utilize an artificial neural network model to predict the Chinese sentential stress. They report that using both acoustic and textual features can get a better performance than that using any one alone. However their study does not involve stress realization. W. Zhu [2] builds an accent detector and a stress-supported unit-selection speech synthesis system. Nevertheless, the expressiveness of Zhu’s system still relies on the audio corpus they used. Although unit selection methods produce high quality speech, they lack flexibility. Secondly, the stress level was manually tagged in Zhu’s speech synthesis system.

The growing demand for expressive speech synthesis forces us to seek an alternative technique. HMM-based speech synthesis system can overcome this drawback by easily modifying the prosodic parameters of the synthetic speech. Much research has already successfully been done upon expressive speech synthesis with HMM-based TTS.

J. Yamafishi, T. Masuko, and T. Kobayashi [3] report their recent progress in generating Japanese expressive speech synthesis. The idea of their work is that prosodic features and spectral features should be controlled properly to model expressive synthetic speech. They use speaking style interpolation and adaptation for HMM-based speech synthesis. In the style adaptation, Maximum Likelihood linear Regression model is adopted. K. Yu, F. Mairesse, and S. Young [4] utilize two new decision tree models and extract word-level emphasis patterns from natural English speech, and then embed the emphasis model in a HMM-based speech synthesis framework. They argue that due to the weakness of emphasis cues, directly using emphasis context features and the traditional adaptation method does not work well. L. Badino, J. S. Andersson, J. Yamagishi, and R. A. J. Clark [5] automatically detect contrastive word pairs with textual features only and use enhanced context dependent labels to synthesize the emphasis. They point out that their methods can convey contrast information effectively. The drawback lies in that the realization of emphasis turned out to be occasionally strong and therefore less contextually appropriate.

Among the previous works only a few works were focused on Mandarin stress synthesis and none of them can generate expressive speech with stress information from raw text automatically. In addition, the focus/emphasis model was embedded in the HMM-based speech synthesis training process, and can hardly be changed to other style once the system was built. Compared to the existing works, the contributions of this paper are:

1) We use Maximum Entropy model to predict Mandarin stress stress within prosodic word for the first time and only textual features are utilized here.

2) A flexible linear stress adaptation module based on the prosodic feature analysis from a large corpus is proposed and integrated in a HMM-based speech synthesis system. No modification is needed in the HMM model training process, in which it is possible that the system can be easily extend to another speaking style or emotion.

The rest of the paper is structured as follows. Section II provides the text-based Mandarin stress prediction model. Section III introduces the stress adaptation model. Section IV presents the experiments and discussions. Section V concludes this work and introduces future work.
II. TEXT-BASED STRESS PREDICTION MODEL

Stress is an important factor in expressive speech synthesis. Since Mandarin is a tonal language, stress is highly related with rhythm, intonation and tone. Automatic stress prediction is difficult but useful to a number of speech processing tasks. We use Maximum Entropy model to predict stress in this work.

A. Textual Feature templates

The text feature templates used in Maximum Entropy Model including Part-of-Speech of a word (POS), the word length (L), the PINYIN script (PY), the tone identity (T), the syllable’s prosodic boundary (B), and the distance from the current syllable to the previous word (DPW) and the distance to next word (DNW). These are the typical textual features used in prosody prediction [6]. The sliding window method was adopted in feature extraction. The features of previous, current and next syllables are considered. We also use a wrapper method [7] in selecting the most effective features to achieve higher precision. All the features selected are combined features, which are:

- B0&B1,
- B_1&T_1,
- DNW0&T0,
- DNW0&T0&T1,
- PY0&T0,
- B_1&DNW0&T0&T1&T_1,
- PY0&T1

The symbol “&” indicates the feature template is a combined feature template. The number after the feature templates indicates the offset. For instance, “B0&B1” means the current syllable prosodic boundary and the next syllable prosodic boundary. The POS and word length related features are not selected which is opposite to our expectation and may deserve further study.

B. Corpus

The audio corpus used in this work contains 6000 sentences (about 73000 syllables), which are read by a professional female speaker. All the utterances are segmentally labeled according to the audio data by research assistants. During the stress labeling, three assistants were trained with a subset corpus several times first. A small percentage of disagreement is acceptable due to the frequent perception confliction among tone, intonation and stress. The aim of training is to keep consistency of each annotator with their own during the whole annotating process and among annotators as much as possible. Three levels of stress are adopted here, namely stressed, regular and unstressed syllable according to their prominence degrees within a prosodic word. To reduce the impact of the surrounding syllables on the perception of the current syllable, we segmented the utterance into prosodic words and stored them according to their tone patterns separately. The average stress degree was used as the ground truth in the text-based stress prediction model training and testing process. In the ground truth data, the percentages for unstressed, regular and stressed syllables are 6%, 58%, 36%, respectively. The whole corpus is divided into the training and test sets according to a ratio of 9:1 randomly. In the training process, the training corpus is divided into 10 parts and a 10-fold cross validation is conducted.

III. STRESS ADAPTATION MODEL

The previous works [4, 5] have proved that the traditional decision tree based clustering may not be effective to capture the emphasis/stress context features. Our preliminary work also shows that the stress information has little power in the tree splitting of HMM model clustering. Therefore, we proposed a flexible stress adaptation model. The framework of this method can be extended to other speaking style or emotion with only a little work.

A. Linear Stress Transformation Model

Stress is highly related with tone, rhythm level and intonation [8]. Tonal coarticulation is also an important factor in Mandarin. Therefore, we adopted the tone pattern category from [9] which categorized tone patterns as “compatible” and “conflicting”. The compatible context is an environment that the adjacent syllables have similar pitch value. While in a conflicting context, the adjacent syllables have very different pitch values. Tone 3 was eliminated due to its strong coarticulation in continuous speech. Fig. 1 shows the tone patterns of disyllables and trisyllables. (Quadrisyllables are rare in the corpus.) The prosodic boundaries are categorized as syllable boundary, prosodic word boundary, prosodic phrase boundary and intonation phrase boundary. Through this method, we defined 120 syllable context categories, i.e., 5 (tone categories) × 4(prosodic boundary categories) × 3(stress levels) × 2(tone pattern categories). The normalized pitch, including the mean pitch, starting pitch and ending pitch as well as normalized duration of all the syllables were statistically summed and then averaged according to their context categories. The normalization methods were listed as (1) and (2) which can express the syllable’s “prominence degree” within prosodic words. Through this method, a linear stress transformation matrix was obtained. By observing the transformation model, we can see that when a syllable is stressed, the normalized duration and pitch are greater than 1 in most cases. On the contrary, when a syllable is unstressed, the normalized parameters are less than 1. This indicates a stress syllable has longer duration and higher pitch, which confirmed with [1, 2]. The detailed information can be found in [8].

---

![Figure 1. Tone patterns of disyllables and trisyllables.](image)
\[ \text{duration}_{\text{norm}} = \frac{\text{duration}}{\text{duration}_{\text{Word}}} \quad (1) \]

\[ \text{pitch}_{\text{norm}} = \frac{\text{pitch}}{\text{pitch}_{\text{Word}}} \quad (2) \]

**B. Stress Adaptation**

As mentioned above, the HMM-based speech synthesis has the impressive control power of spectrum, pitch as well as duration. Consequently, most of the expressive speech synthesis studies were based on HMM model. To enhance the flexibility of the system, we use stress adaptation method which can be extend to further applications with a little amount of work. Therefore the HMM-based parameter generation is the same as [10] in our work. Fig. 2 illustrates the flowchart of the proposed expressive speech synthesis with stress adaptation module. After text analysis, features, namely word segmentation, POS and rhythm level etc. are passed to the HMM-based static parameter generation module, except the stress information. The outputs are the sequences of pitch, duration and spectrum parameters. Simultaneously, the linear transformation ratio of each syllable is obtained by searching the large stress adaptation matrix according to its stress level. Afterwards, we can get the transformed static parameters, i.e., pitch and duration. Then the dynamic feature parameters are generated from HMM model. Through this method, the pitch and duration are accurately controlled and that the dynamic parameters are smooth. Finally, the expressive speech is produced by a synthesizer.

The advantage of the proposed system is that the stress adaptation module is independent from the HMM model training, making it easier to build a new system with other speaking style or emotion. Similar claim can be found in [3].

**IV. EXPERIMENT AND DISCUSSION**

In this section, the text-based stress prediction model and the overall performance of synthetic speech are evaluated.

**A. Text-based Stress Prediction Results**

In evaluation of the text-based stress prediction, we just examine the overall precision, which is the percentage of the amount of the correctly predicted samples in all the test samples. We did not calculate the precision, recall and F-score of each stress category, because one criterion is easier to conduct in feature selection using wrapper method. The baseline model used all the single feature templates mentioned above, including POS, the word length, the PINYIN script, the tone identity, the syllable’s prosodic boundary, and the distance from current syllable to previous word and the distance to next word. The sliding window length is three syllables. The baseline model got the precision of 63%. By feature selection, a small amount of effective features were selected as listed in section II. The final precision is improved to 72.5%. Although this is not a very promising result, it is pretty close to the consistency among the three annotators, which is about 75%. This also implies that stress perception is a very complex issue and further work should be done in this field.

![Figure 2. The proposed expressive speech synthesis system with stress.](image)

**B. Evaluation of Synthetic Speech**

High precision of text-based stress prediction only denotes the prediction results perfectly match the human-annotated labels. We should evaluate the overall performance of the expressive synthetic speech to see if any improvement was achieved. Two kinds of synthetic speeches, one is original HMM-based synthetic speech without stress, and the other is the proposed HMM-based expressive synthetic speech with stress were generated. We performed a fully evaluation among these two speeches and natural speech. Moreover, to show the flexibility and generalization of the proposed method, the test natural speech is from a female speaker. The two synthetic speeches are based on a male speaker’s parameters.

Fig. 3 shows pitch contours of two synthetic speeches as well as the natural speech. The ups and downs in the synthetic speech with stress are much more visible and the pitch range is wider. On the contrary, the pitch contour of synthetic speech without stress is somewhat flat, which makes the utterance lack expressiveness. However, we should note that although the proposed stress adaptation method can convey prominence effectively, the linear transformed pitch and duration may affect the naturalness potentially. Therefore, we also compare the synthetic speech with natural speech. It can be observed from Fig. 3 that while the last four syllables’ pitch contour of the synthetic speech with stress is not very close to natural speech, the pitch contour of the first four syllables is similar to the natural speech. This indicates, firstly, that the expressiveness of the proposed system is as much as natural speech; secondly, the synthetic speech is not exaggerative which may downgrade the naturalness. Regarding the pitch contour of last four syllables, the principal reason is there are two tone 3 syllables in the utterance. As already noted, tone coarticulation effect is serious in tone 3 syllable. The surrounding syllables’ pitch contour would be affected, too.
A listening test was conducted to measure the ability of the proposed system for conveying stress as well as the overall performance. For each system, ten utterances were generated with and without stress. Nine speech experts and three students who are not very familiar with speech processing were asked to rate these sentences. All of them are native speakers. The natural speech was scored at the same time. MOS score, which is from 1.0 to 5.0, is designed to judge the overall speech performance, emphasizing the naturalness, intelligibility and expressiveness. A preference test was also conducted on the ten pair-wised synthetic speeches to mainly focus on the expressiveness. As illustrated in Table I, the MOS score is improved of about 0.2. The preference test result demonstrated in Fig. 4 shows the speech with stress is preferred. These results indicate that the proposed expressive speech synthetic system achieves better overall performance, both in naturalness and expressiveness. However, the difference in MOS score between the two synthetic speeches are not as much as that in preference test, the reason is stress prediction error may result in unnaturalness in the final synthetic speech. Nevertheless, the improvement in MOS score implies this defect is not a very big problem from users’ perspective.

V. CONCLUSIONS AND FUTURE WORKS

The naturalness and intelligibility of synthetic speech have been studied for many years and have already achieved a remarkable progress. Now it is the expressiveness that attracts great interests of researchers and users. Stress is one key factor that can improve the expressiveness of synthetic speech. Concatenation speech synthesis system can not convey stress successfully because its expressiveness is highly dependent on the audio corpus it used. HMM-based speech synthesis system could control the speech parameter accurately, which is useful for expressive speech study. However, stress related context splitting process in the traditional HMM-based speech synthesis system. Therefore, we proposed a stress adaptation model in HMM-based synthesis system to generate expressive speech. Also, we use a Maximum Entropy model to predict the stress information from textual features automatically. A wrapper feature selection is conducted to make the model more precise and compact. The aim of stress adaptation model is to modify the speech features according to their stress levels. The adjusting parameters are obtained by analyzing prosodic features of syllables from a stress-labeled audio corpus. Experiments show that the mean opinion score is improved from 3.7 to 3.9, and more than 75% synthetic speeches with stress are chosen in the preference test. These indicate the proposed system can convey stress information successfully. The naturalness of the synthetic speech is improved, too.

The future work will focus on the followings. Firstly, since a high expressive speech synthesis system will expose and enlarge the shortcomings at the same time, further study will be conducted on text-based stress prediction to improve its performance. Secondly, we will introduce prosodic features in stress prediction for corpus labeling. Because a limitation of this work is that the linear stress adaptation model used here is obtained from a human-labeled corpus, which is very time-consuming. Automatically labeling the audio corpus is very helpful for extending the proposed system to other expressive speaking style.

REFERENCES


Figure 3. Two synthetic speeches and a natural speech. (The two synthetic speeches are male, the natural speech is female. Because female speech is higher than male speech in pitch, the female utterance is decreased 100Hz to have more or less the same pitch value as male speech, in which the contrast among three utterances can be demonstrated clearly.)

Figure 4. Preference comparison between two synthetic speeches.

<table>
<thead>
<tr>
<th>TABLE I. MOS OF TWO SYNTHETIC SPEECHES AND NATURAL SPEECH</th>
<th>Speech synthesis system</th>
<th>Mean Opinion Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic speech without stress</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td>Synthetic speech with stress</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>Natural speech</td>
<td>4.8</td>
<td></td>
</tr>
</tbody>
</table>