ABSTRACT

To improve the expressiveness of speech synthesis, the paper proposes a model to simulate the prosody features of exclamatory speech with modal tags. While compared with reading speech, we found that the major difference between the reading speech and the kind of exclamatory speech is caused by the strong stresses on some modal words and their heavy impacts on adjacent speech units. Then, a CART-based prosody transformation model is introduced to automatically generate the prosody features of exclamatory speech by using reading speech as the baseline. Final perception and comparison experiments have proven the high quality of the model in the simulation of the kind of exclamatory speech.

Index Terms—Prosody Conversion, Exclamatory Speech, Stress, Speech Synthesis

1. INTRODUCTION

During the past several years, speech synthesis technologies have achieved great improvement, but they still fall short of expressiveness in their ability to convey information via different speaking styles. To make synthesized speech more expressive, some methods have been tried, such as rule-based [1], unit selection [2, 3], neural network [4], CART model [5], etc. Among them, the unit selection method with a large expressive corpus has been most popularly used [2, 3]. Unfortunately, this method is very costly and not easy to be used for generating stable synthesized speech, because it is normally difficult to collect a large expressive corpus with detailed labeled information. Recently, a prosody conversion method [6] has been tried for expressive speech generation. The method only requires a very small amount of training corpus and can be used for both speech synthesis [6] and voice conversion [5]. However, most of current investigation about expressive speech still focuses on emotional speech. Some spontaneous styles like exclamatory speech have seldom been touched.

As the exclamatory speech has been widely used in speech dialog systems, in this paper, we are trying to create a prosody model to automatically simulate the features of exclamatory speech. Most of the exclamatory speech happens in some special situations while speakers are trying to express some strong emotions, like “strong anger”, “strong happiness”, or “strong doubt”, etc. This kind of exclamatory sentence normally contains some modal words including some adverbs, adjectives, verbs, like “tai4”, “hen3”, “duo1me5”, “hen4bu4de5”, etc. Through the analysis on parallel corpus, containing reading and exclamatory corpus, we find that the major difference between them is mainly caused by the strong stresses of these modal words and their heavy impacts on adjacent speech units. Although the stress has been in research for many years [7, 8], most of them are still limited in word or phrase level [9], and there are much fewer results for the stress impact analysis on the adjacent speech units within a continuous speech. The paper analyzes the impact of stress and finds that subsequent speech units generally get more impact of stress while compared with former speech units. We then use these results to construct a new group of context information for prosody model. Finally, a CART-based prosody transformation model is introduced to automatically generate the prosody features for exclamatory speech with strong stress styles by using reading speech as the baseline. Experiments show that the synthesized speech can generate exclamatory prosody features with very high quality.

The paper is organized as follows: In Section 2, prosody characteristic of exclamatory speech is analyzed based on a parallel speech corpus. In Section 3, a CART-based method is introduced to generate exclamatory prosody by learning the difference of reading speech and exclamatory speech with a new group of context information. In Section 4, the paper makes some experiments and proves the high quality of the exclamatory prosody model proposed in the paper. The final conclusion is presented in Section 5.

2. PROSODIC FEATURES IN EXCLAMATORY SPEECH

To fully understand prosodic features in exclamatory speech with modal words, our method focuses on the prosodic difference between exclamatory and reading speech, which can be obtained from the comparison between exclamatory corpus and reading corpus.

2.1. Parallel Exclamatory Speech Corpus
The corpus used in the paper contains both exclamatory speech and reading speech, which are based on the same transcript selected from TV serials. The transcript consists of 250 sentences. A female professional speaker was asked to read these sentences using both reading style and exclamatory style. After the recording, all the utterances were segmentally and prosodically annotated with break index and stress index information. In this kind of exclamatory speech, the stress is mainly represented by modal words. This paper mainly focuses on the stress of modal words. Neither word stress nor sentence stress is annotated, and only these stresses of modal words are annotated.

2.2. Impacts Analysis of Strong Stresses in Exclamatory Speech

Fig. 1 An example of exclamatory and reading speech, “zhe4 ye3 tai4 hu2 du4 le5” (In English, it does mean that this is too stupid.)

As we observe that there is not less than one strong stress in this kind of exclamatory speech. Fig.1 shows pitch contour of parallel exclamatory speech and reading speech. The stressed syllable in exclamatory speech is the third syllable “tai4”, which has wider F0 range and longer segment compared with reading speech. Moreover, the strong stress has some heavy impacts on prosodic features of adjacent syllables. From Fig.1, we find that the following syllables of the stressed syllables usually have larger pitch ranges and shorter durations compared to the reading speech. For instance, the last syllable “le5” of the exclamatory speech has the flatter shape than that in the reading speech. These analysis results can be found among most of our exclamatory speech corpus.

To measure the change level caused by strong stress, we use three parameters (F0 mean, F0 range and duration) to describe prosody difference of two parallel speeches. The statistical result of the difference is calculated by the following equation:

\[ D = \frac{1}{N} \sum_{i=1}^{N} (P_{\text{exclamatory},i} - P_{\text{read},i}) \]

Where P is a vector containing F0 mean, F0 range and duration of each syllable. \(d\) also denotes the vector containing the average differences of the three parameters as well. N is the number of syllables. Based on Equ.1, the strong stress and its six adjacent syllables are analyzed. Experimental results are shown in Fig.2.

In Fig.2, the x-axis means that the distance between the analyzed syllable and the strong stress. Positive value means that the syllable is behind the stress syllable, and vice versa. Zero value is the stressed syllable. All comparison is based on parallel exclamatory and reading speech. From Fig.2 we can see, F0 range of syllables behind strong stress has smaller difference than that in front of strong stress, whereas both F0 mean and duration have larger difference. Beside, the smaller the difference of three parameters becomes, the farther the distance from the strong stress is. All in all, the strong stress has heavier impacts on the syllables that are close to it than those that are far from it. Furthermore, the syllables behind strong stress have heavier impacts than those in front of strong stress.

3. PROSODY MODELING FOR STRONG STRESS

3.1 Denoting F0 Contour for Exclamatory Speech

As the pitch target model [10] has the good ability in Mandarin stress description and has the convenience of parametric alignment in prosody transform, it is used in our paper to describe F0 contour of parallel training corpus. Supposing the syllable boundary is \([0, D]\), the pitch target model is denoted by the following equations [9]

\[ T(t) = at + b \]

\[ y(t) = \beta \exp(-\lambda t) + at + b \]

\[ 0 \leq t \leq D, \lambda \geq 0 \]

A pitch target model of one syllable can be represented by a set of parameters \((a, b, \beta, \lambda)\). \(T(t)\) is the underlying pitch target, and \(y(t)\) is the surface F0 contour. The parameter \(a\)
and $b$ are the slope and intercept of the underlying pitch target respectively. The coefficient $\beta$ is a parameter measuring the distance between the F0 contour and the underlying pitch target at $t=0$. $\lambda$ describes how fast the underlying pitch target is approached. The Levenberg-Marquardt algorithm [11] is used for estimating the set of parameters.

By applying the pitch target model in the parallel corpus, we find the parameter $b$, having the least change level among these parameters, is decrease by 19%. Fig.3 shows movement of pitch target parameters. The change of F0 mean and F0 range is given as well in Table 1.

![Fig.3 Movement of F0 Contour with pitch target parameters](image)

In Fig.3, we can see that the parameter $b$ influences F0 contour most, and the parameter $\lambda$ does least. From Table 1, we can see that the pitch target parameters can represent the change of F0 range and F0 mean precisely. They all imply that pitch target model can easily describe strong stress and its impact on adjacent syllables.

Table 1: Variability of F0 Mean and F0 Range (Unit: Hz)

<table>
<thead>
<tr>
<th>Pitch Target Parameters</th>
<th>Change of F0 Mean(Hz)</th>
<th>Change of F0 Range(Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$b$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>-0.653</td>
<td>200</td>
<td>12.58</td>
</tr>
<tr>
<td>-0.4</td>
<td>247.2</td>
<td>12.58</td>
</tr>
<tr>
<td>-0.653</td>
<td>247.2</td>
<td>30</td>
</tr>
<tr>
<td>-0.653</td>
<td>247.2</td>
<td>12.58</td>
</tr>
</tbody>
</table>

### 3.2. Model Construction

Classification and Regression Trees (CART) has been proved to efficiently simulate the prosody difference of two parallel speeches by integrating the contextual information [5]. It is also used in our work to model the difference between exclamatory speech and reading speech. Most of the context information in [5] is also kept for using in the paper. However, based on the analysis results in Section 2, the features of strong stress and its impact on adjacent syllables are integrated, and they are listed with bold. Then we get a new group of context information:

- Initial Identity (Including current and following syllables' initial types, with 8 categories).
- Final Identity (Including current and previous syllables' final types, with 4 categories).
- Break information (Including the preceding and following break type, with 4 categories).
- Position information (in sentence).
- Stress level of the syllables (it only happen for syllables in modal words).
- Distance from the preceding stressed syllable.
- Distance from the following stressed syllable.

For prosody conversion, the output parameters are the differences of pitch target parameters $a, b, \beta$ between exclamatory and reading speech.

In the training procedure, source and target pitch contour from parallel corpus is firstly aligned according to syllable boundary, and then pitch target parameters are extracted from syllabic F0 contour, finally both the context information listed above and difference of $a, b, \beta$ are inputted into CART for training. In the conversion procedure, pitch target parameters of reading speech are estimated and the difference of $a, b, \beta$ is predicted by CART according to context information and then the pitch target parameters of exclamatory speech can be obtained by adding the difference into the pitch target parameters of reading speech. In the end, the pitch target parameters of exclamatory speech are used to generate new pitch contour associated with exclamatory characteristics. In the procedure, parameter $\lambda$ is not considered, because it weakly influences F0 contour as described in Section 3.1.

![Fig.4 CART-based exclamatory prosody generating by using reading speech as the baseline](image)

4. **EXPERIMENT AND EVALUATION**

To evaluate our exclamatory prosody model, we try to use it to convert the prosody of reading speech to exclamatory style. Based on reading spectrum and transformed exclamatory prosody feature, STRAIGHT model [12] is adopted to synthesize new speech with exclamatory style.

4.1. **Experiments on subjective evaluation**


A subjective evaluation experiment is then implemented to evaluate the proposed exclamatory prosody conversion model. In this test, the proposed model is compared with a linear regression transformation model (LR-based) and a general transformation model. In LR-based model, the pitch contour with reading style is obtained, the linear conversion method [11] is adopted to transform reading style to exclamatory style. In the general model, the last three features listed in Section 3.2, stress and distance features, are not considered to predict the difference of pitch target parameters. In these systems, the duration is transformed as linear relationship and the spectrum information directly comes from reading speech. Sixteen undergraduates are asked to rate which sentence sounds more exclamatory with a scale 0-5. Table 2 shows the rating result. From this table, we can see that the presented method can generate exclamatory speech with higher quality.

### 4.2. Experiments on objective evaluation

Besides this subjective evaluation, another objective evaluation experiment is done. The three methods are compared in terms of RMSE (Root Mean Squared Error) and shape similarity. Among them, the shape similarity is measured by the correlation of pitch contour within a syllable between aligned exclamatory pitch contour and synthesized one, which reflects the result of our proposed conversion model. Table 3 shows the objective evaluation result. From this table, we can see that both the RMSE and shape similarity are greatly improved by the proposed method. Besides, the presented method can generate pitch contour which is more like exclamatory style.

**Table 2: MOS results for different methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Rating Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>The presented method</td>
<td>3.9</td>
</tr>
<tr>
<td>The General method</td>
<td>2.8</td>
</tr>
<tr>
<td>The LR-based method</td>
<td>2.0</td>
</tr>
</tbody>
</table>

**Table 3: objective evaluation result for different methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>Shape Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>The presented method</td>
<td>18.6</td>
<td>0.82</td>
</tr>
<tr>
<td>The General method</td>
<td>29.1</td>
<td>0.65</td>
</tr>
<tr>
<td>The LR-based method</td>
<td>36.4</td>
<td>0.53</td>
</tr>
</tbody>
</table>

### 5. CONCLUSION

This paper proposed a method to synthesize a specific kind of exclamatory speech. The essential idea of this method is to model the strong stress of modal words and its impacts on adjacent syllables. The experimental result has proven the reasonability of the method.

Although the prosody model proposed in the paper has to use reading speech as the baseline, it can be easily extended to text-to-speech system by two methods: adding the model on the acoustic part which can output neutral synthesized speech, or changing the prosody module with the group of context information mentioned in Section 3.2. However, a large corpus with stress labeling is required for training while we are using the later method.

While in the paper we focus on simulating the strong stress of modal words, the exclamatory speech actually has more styles. Our future work will further concentrate on other kinds of exclamatory speech. We will also try to make more detailed classification of the spontaneous speech in our future work.

### 6. ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China (No. 60873160) and Nokia funding.

### 7. REFERENCES


893