AUTOMATIC STRESS PREDICTION OF CHINESE SPEECH SYNTHESIS

Jian-Hua Tao (1) Sheng Zhao (2) Lian-Hong Cai (3)

Department of Computer Science and Technology, Tsinghua University, Beijing

{ (1) jhtao, (2) clh-dcs}@tsinghua.edu.cn (3) sz00@mails.tsinghua.edu.cn

ABSTRACT

The stress was proved to be the essential links between linguistics and acoustics, and behaves as an important parameter for prosody processing and unit selection in speech synthesis system. In the paper, some acoustical measurements are carried out on F0, duration, silence in order to disclose the relationship between stress and acoustical parameters. The normalized compared acoustic parameters are induced to facilitate the stress detecting from the speech. Furthermore, a rule-learning approach is proposed to predict stress in unrestricted Chinese text. In order to improve the accuracy rate of prediction rules, the most effective linguistic features related to stress are selected according to several experiments. The method is proved to be very successful and has been integrated into our speech synthesis system. We get 86% accurate rate of stress prediction. Further listening tests also show that the expressive force of synthesized speech is improved a lot compared to the systems based on traditional method.

1. INTRODUCTION

During the last several years, there has been a rapid progress in Chinese speech synthesis. Now, the method of unit selection and concatenation, accompanying with large corpus, is used widely in the systems design. Nevertheless, the stress was still proved to be the essential links between linguistics and acoustics, and behaves as an important parameter for prosody processing and unit selection. However, it is a real hard work for us to handle the stress, such as how to detect the stresses in the corpus with high consistency and how to predict the stresses from the linguistic information.

The first question exists in the phase of corpus design and labeling. Normally, stress is not a very well defined term in literature. A common definition of prominence is that it refers to those words or syllables that are perceived as standing out from their environment. Perceived syllable stress was interpreted as a gradual parameter by Fant & Kruckenberg. Subjects rated the perceived stress of syllables on a 30-point scale. The authors investigated a small corpus of Swedish and found linear relationships between perceived prominence and acoustic and articulatory parameters. They also investigated the consistency of their labellers and obtained high correlations; this was confirmed by de Pijper & Sanderman for boundary prominence. Grover et al. showed that the reliability of word prominence ratings is higher for a 10-point scale than for a 4-point scale. As we know, Chinese is a tonal language and syllable is normally assigned as the basic prosodic element in processing. Each syllable has a tone, and has a relatively steady F0 contour. It is difficult to determine the stresses within the influence of various syllabic tone patterns. In the paper, some acoustical measurements are carried out on F0, duration, and silence, in order to disclose the relationship between stress and acoustical parameters, and to make stress labeling in high precise and high consistency. Results show that stress is influenced not only by pitch range and duration of the syllables, but also by the neighboring silence and neighboring stresses.

To get the relationship between linguistic processing and acoustic processing, some data-driven methods have been introduced in English, such as Classification and Regression Tree (CART), Hidden Markov Model (HMM), neural network auto associators. Whereas, rule based stress prediction is still the popular method for most of the current Chinese speech synthesis systems. As a result, the naturalness and flexibility of the system are limited in a certain extent. In the paper, rule learning approach is proposed to predict stress in unrestricted Chinese text. In order to improve the accuracy rate of prediction rules, the most effective linguistic features related to stress are selected according to several experiments. The method is proved to be very successful and has been integrated into our speech synthesis system. We get 86% accurate rate of stress prediction. Compared to other methods, it can be though as a very high performance. Further listening tests also show that the expressive force of synthesized speech is improved a lot compared to the systems based on traditional rule based method.

The paper is organized as following. In Section 2, the acoustic features of stresses are analyzed to make automatic stress tagging for the corpus. Section 3 analyzes the mapping the patterns between syntactic information and stresses. In Section 4, a rule-learning algorithm is described in detail, which is used to predict the stress automatically from the linguistic information. The results are analyzed in Section 4. Section 5 presents the conclusion and the view of future work.

2. STRESS DETECTING

As we know, tone is the most important and basic prosody feature in Chinese. There are four lexical tones exist, and each tone contains relatively constant pitch patterns. The pitch movement of stressed syllables in Chinese is complicated that it cannot be described as one line intonation model. Pitch range of syllables can be described as top-line and bottom-line correlates to the stressed components.

In the paper, we try to describe a method on how to detect stress from the speech in the large Chinese speech database. Firstly, to be able to compare with normal behaviors, the acoustic parameters, duration, top F0 and bottom F0 of stressed syllables, are divided by their statistic average values (normal behaviors).

\[ R_{dyi} = \frac{\text{Duration of Syllable } i}{\text{Average Duration of all Syllables}} \]  

(1)

\[ P_{dyi} = \frac{\text{Top F0 of Syllable } i}{\text{Mean value of Top F0 of all Syllables}} \]  

(2)

\[ \text{Syllables of Duration Average} \]

Syllables of Duration Mean

\[ \text{Syllables of FTop of value Mean} \]

**Supported by 863 program (2001AA114072)**
Here, $i$ means syllable $i$. $R_{dur,i}$, $P_{top,i}$, and $P_{bottom,i}$ represent the compared acoustic parameters, their distributions are shown in Figure 1(a),1(c),1(e).

\begin{equation}
\text{Bottom F0 of Syllable } i
\end{equation}

\begin{equation}
\text{Mean value of Bottom F0 of all Syllables}
\end{equation}

\begin{equation}
\text{Top F0 of Syllable } i
\end{equation}

\begin{equation}
\text{Mean value of Top F0 of Syllable } i
\end{equation}

Normalized compared acoustic parameters are defined as,

\begin{equation}
R'_{dur,i} = \frac{1}{N} \sum_{i=1}^{N} R_{dur,i}
\end{equation}

\begin{equation}
P'_{top,i} = \frac{1}{N} \sum_{i=1}^{N} P_{top,i}
\end{equation}

\begin{equation}
P'_{bottom,i} = \frac{1}{N} \sum_{i=1}^{N} P_{bottom,i}
\end{equation}

Then, we can get the new distribution shown in figure 1(b),1(d),1(f). It can be seen that $R'_{dur,i}$, $P'_{top,i}$ and $P'_{bottom,i}$ of most stressed syllables are more than 1. The distribution of them shows that normalized compared acoustic parameters represent the stressed syllables much better. To get the more relationship between stressed syllable with other information, silence and adjacent weakened syllables are also taken into account. The compared duration of silence, and statistical number of weakened syllables closed to stressed syllables are got by the following formulas. The results are shown in figure 2.

\begin{equation}
S_{sil} = \text{Length of Silence before Stressed Syllables}
\end{equation}

\begin{equation}
S_{sil} = \frac{\text{Average Length of Silence}}{\text{Total Number of Stressed Syllables}}
\end{equation}

\begin{equation}
N_{sil} = \frac{\text{Number of Adjacent Previous Weakened Syllables}}{\text{Total Number of Stressed Syllables}}
\end{equation}

\begin{equation}
N_{sil} = \frac{\text{Number of Adjacent Succeeding Weakened Syllables}}{\text{Total Number of Stressed Syllables}}
\end{equation}

From figure 1 and figure 2, some results can be got,

A. The pitch movement of syllable stress is realized by shifting up of the pitch with relatively constant pitch contours and enlarging the duration, which confirm the Wu’s view [2].
B. Stressed syllables are also influenced by the neighboring silence. Normally, succeeding silence of stressed syllable is shortened. But there is no enough facts support that stressed syllables are influenced by the preceding silence.

C. Experiments also show that the stress may also be influenced by the previous syllables. Normally, the syllables are weakened if they appear before a stressed syllable. There is no enough phenomenon support that the stress syllables are also influenced by the following syllables.

Based on above knowledge, a criteria of stress determining can be defined by,

\[ A_n = \alpha \cdot R_{\text{dur},n} + \beta \cdot P_{\text{top},n} + \gamma \cdot P_{\text{bot},n} + \eta \cdot S_{\text{Next}} + \delta \cdot A_{n-1} + C \]

(14)

Where, \( A_n \) means stress degree of the syllable \( n \) in the sentence. \( \alpha, \beta, \gamma, \eta, \delta \) are the coefficients (between 0 and 1). \( C \) is the constant. In our database, we use \( \alpha = 0.7, \beta = 0.65, \gamma = 0.45, \eta = 0.3, \delta = 0.5 \). However, the coefficients may be various to different conditions. Furthermore, more accurate and efficient coefficients can also be generated by a training method.

3. SYNTAX TO STRESS MAPPING

A linguistic theory of syntax-stress mapping must consider the underlying syntactic structure in terms of its hierarchical organization, especially if the syntax of a given language allows different directions of branching as it is the case in an Object-Verb-language (OV) such as Chinese. In Chinese, the syntactic OV-parameter means that in structures with verb-final word order, i.e., in most subordinate clauses, the verb takes its argument from its syntactic structure, the ‘normal’ default accentuation has to be adopted. In A1 the second verb ‘think’ is accented.

For reasons of explanatory adequacy, we make use of theories which consider the information structure. In Jacobs (1993), for both the so-called ‘normally intonated’ sentences, e.g., widely focused sentences, and sentences containing narrowly focused constituents, stress positions are predictable by terms of integration. Stress positions in terms of their relative prominence (e.g., the weight of accents distributed over a syntactic structure) is thus calculable if the position is fixed.

3.1. Influenced by syntactic structure

In a first step of syntax-stress mapping, the syntactic structure and the position have to be determined. There are 191 patterns concerned for Chinese totally in the paper. Let us assume a basic syntactic structure reflecting the superficially linear VO-order (A1):

\[ \text{A1:} \quad \text{}\quad \text{\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \ quartef
The corpus preparing
Speech database used in this evaluation is the continuous male speech database of phoneme balanced 3000 Chinese sentences chosen from the four years’ news paper of PEOPLE DAILY. There are 19806 Chinese characters in the corpus, which constitute 13375 words. All of the sentences were labeled with stress in three-level speech database of phoneme balanced 3000 Chinese sentences.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule based</td>
<td>0.72</td>
</tr>
<tr>
<td>CART</td>
<td>0.83</td>
</tr>
<tr>
<td>TRBL</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table 3: Prediction precision

From table 3, we can find the rule based method is superior to the pure rule based method. In contrast, decision trees are aimed at classifying independent vectors, though questions about local context can be incorporated by making Markov assumptions and using dynamic programming and the most likely sequence. For this reason, TRBL tends to be less sensitive to data scarcity, and is better able to learn parameters associated with independent factors.

5. RESULTS AND ANALYSIS

5.1. The corpus preparing
Speech database used in this evaluation is the continuous male speech database of phoneme balanced 3000 Chinese sentences chosen from the four years’ news paper of PEOPLE DAILY. There are 19806 Chinese characters in the corpus, which constitute 13375 words. All of the sentences were labeled with stress in three-level with automatic labeling method based on formula (14). And, two experienced annotators, guided by a senior phonetist, checked all of the results by listening. The labeling results of them achieve a high consistency rate of 93.1%.

5.2. Evaluation Parameters
Stress prediction was evaluated with subjective or objective measure. The subjective measure is generally performed by perceptive tests, which are undoubtedly convincing but time-consuming to conduct on large corpus. In this paper, only the objective measure is adopted. As a classification task, stress prediction should be evaluated with consideration on all the stresses. The trained classifiers are applied on a test corpus to predict the label of each stress.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule based</td>
<td>0.72</td>
</tr>
<tr>
<td>CART</td>
<td>0.83</td>
</tr>
<tr>
<td>TRBL</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table 3: Prediction precision

From table 3, we can find the rule based method is superior to the pure rule based method. In contrast, decision trees are aimed at classifying independent vectors, though questions about local context can be incorporated by making Markov assumptions and using dynamic programming and the most likely sequence. For this reason, TRBL tends to be less sensitive to data scarcity, and is better able to learn parameters associated with independent factors.

6. CONCLUSIONS

The paper compared the acoustic parameters of stressed syllables with the corresponding normal status, and draw some efficient views for automatic stress labeling of large speech database. Facilitated by the manual checking, a high labeling consistency can be acquired. The paper also introduced a new approach to symbolic prosodic label prediction based on transformational rule-based learning. Experiments on stress prediction with a news corpus show that TRBL gives a small improvement over simple decision tree predictors, despite a more simplified approach to set membership rule design. In addition, the experiments showed that stress prediction benefits from phrase structure, but not vice versa. The use of average absolute distance is proposed as a new metric for design and evaluation of stress prediction, which is motivated by the graded acoustic cues observed for different stresses. On the other hand our results are based on cross validation tests, which give a better estimation of the performance when the classifiers are running on noisy inputs. Thanks Prof. Zhou Tongchun for his effort of labeling most the training corpus.

7. REFERENCES


Table 1, Examples of rule templates

<table>
<thead>
<tr>
<th>Rule template</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS_0 POS_1 =&gt; STRESS_1</td>
<td></td>
</tr>
<tr>
<td>POS_-1 POS_0 POS_1 =&gt; STRESS_1</td>
<td></td>
</tr>
<tr>
<td>STRESS_0 POS_-1 POS_0 POS_1 =&gt; STRESS_1</td>
<td></td>
</tr>
<tr>
<td>POS_0 POS_1 WLEN_0 WLEN_1 =&gt; STRESS_1</td>
<td></td>
</tr>
</tbody>
</table>