**The Pause Duration Prediction for Mandarin Text-to-Speech System**

Jian Yu(1) Jianhua Tao(2)  
National Laboratory of Pattern Recognition (NLPR)  
Institute of Automation, Chinese Academy of Sciences  
{jyu(1), jhtao(2)}@nlpr.ia.ac.cn

**Abstract**—In this paper, we enter into detailed analysis on how the pause duration under different prosodic boundaries are affected by various contextual factors in natural speech. To get the correlation between them, the paper calculates the mean pause duration under different prosodic boundaries. The contextual factors investigated in this paper contains both linguistic features, such as boundary types, syllable tones of boundary sides, initial and final types etc, and acoustic features, such as pitch gap across the boundary. The paper makes experiments and discussion which reveals the influence of these factors on pause duration. Based on that, the paper creates a pause duration prediction model for mandarin speech synthesis system. The model was proved to be able to generate high quality prosody output with the listening test.

I. INTRODUCTION

In text-to-speech system, it’s very important to predict the prosody information for producing natural sounding speeches. The pause duration model, which is one of the important parts of prosody model, is essential in improving prosodic quality. The synthesized speech will not be natural and even unacceptable sometimes, if we only use constant pause duration for each prosody boundary. Actually, the pause duration is related to various linguistic features and other prosodic features.

Although the research on pause model has not been paid as much attention as pitch model and duration model before, there is still some work having been done in the last several years. The rule-based method [1] is one of the typical methods, which used linguistic expertise to infer some pause generation rules based on observations on large corpus. This approach is simple and convenient, but it’s quite time consuming to get lots of trivial rules. And the results, which influence the prosody generation, were not so good. Later on, someone tried to use the training model such as ANN for pause duration prediction [2]. It generated better results than traditional rule-based methods, but with ANN model, we had to prepare a very large corpus for training, and the results were still limited in most cases. It is also hard to get the relationship between pause duration and other features with only ANN outputs. Similar work has also been done by some others [6].

Unlike previous works, in this paper, we make the detailed analysis on the relationship between pause duration and various contextual features. These results help us to understand the nature of pause generation and predict pause duration more precisely. Then, a decision tree is used to automatically collect the rules for pause duration prediction with a limited training corpus. The model has successfully been integrated into the CASIA TTS system, and was proved to be able to generate high prosody outputs.

This paper is organized as follows. The speech corpus that our research is based on is introduced in section II. Section III elaborates on how our experiments are carried out for both text information and prosodic information. The results of experiments are listed in detail and analyzed from the points of phonetics and phonology. Section IV introduces the CART model based pause duration prediction, and the experiments of the model outputs. The final discussion and conclusion are arranged in the last part.

II. SPEECH CORPUS

The corpus used in our work contains 5,000 sentences (about 80,000 syllables), which are recorded from a professional female speaker. It is carefully designed to cover all of the Mandarin syllables, tone combinations, and as much as contextual variations. The corpus was manually labeled with prosody boundaries, word segmentation, POS tagging, pitch tagging, and acoustic syllabic boundaries. The prosodic boundaries are classified into four layers. They are,

* B0: syllable boundary.
* B1: prosodic word boundary, a group of syllables that are uttered closely.
* B2: prosodic phrase boundary, a group of prosodic words that has a perceptive rhythm break at the end.
* B3: sentence boundary, the utterance for a whole speech.

Sentence boundaries always contain a long silence which is out of the research in the paper. The statistical distribution of the pause in other boundaries is listed in table 1. It contains mean pause duration, standard deviation and appearing probabilities of pauses.

<table>
<thead>
<tr>
<th>Boundary</th>
<th>Mean</th>
<th>Deviation</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>B0</td>
<td>38 ms</td>
<td>19 ms</td>
<td>47.3%</td>
</tr>
<tr>
<td>B1</td>
<td>51 ms</td>
<td>25 ms</td>
<td>63.6%</td>
</tr>
<tr>
<td>B2</td>
<td>122 ms</td>
<td>68 ms</td>
<td>97.2%</td>
</tr>
</tbody>
</table>
As we see, the longer pause durations are normally related to the higher prosody boundaries. The results are used by most of the previous TTS systems, but they just used some simple rules. On the other hand we can not say the higher prosody boundaries always make the longer pause duration, since there is still a big deviation of the pause in each boundary. Such as in B2, the standard deviation of the pause is 68ms while the mean pause duration is just 122ms. Therefore, more features are necessary for predicting pause duration precisely.

III. EXPERIMENTS AND DISCUSSION

When predicting prosodic information in the past, researchers always only made use of the results of text analysis, neglecting the prosodic information itself [4] [5]. Therefore, the various factors investigated in this paper include not only text information, such as boundary type, initial category and final category, etc, but also prosodic information itself, such as pitch gap across the pause.

A. Influenced by inner-syllabic feature

Considering the great influence of initial and final category on syllable duration, we may suspect that the initial category of the following syllable and final category of the previous syllable have some influence on pause duration. Fig. 1 and Fig. 2 show the statistic results based on our corpus.

These figures demonstrate the influence of the previous syllable’s final category and the following syllable’s initial category, showing three remarkable points: (1) under non-phrase boundary, the following syllable’s initial category has great influence on pause duration. Fig. 1 displays that the pause duration before stops and affricates is much longer than that before fricatives, nasals, and zero initials. This phenomenon may be caused by different initials’ articulatory manners. For example, some part of vocal track is closed before stops are pronounced, which leads to the appearing of pause. (2) Compared with initial category, the influence of the previous syllable’s final category under non-phrase boundary is weaker, but we can still observe that the pause duration after nasal finals gets a little shorter under non-phrase boundary. (3) Under phrase boundary, neither the previous syllable’s final category nor the following syllable’s initial category has influence on pause duration. One reason might be that the pause duration under phrase boundary is already long enough that articulators can complete any action happened in vocal track, such as closing action before pronouncing stops, so the influence of initial category and final category can not be revealed.

B. Influenced by tone combination

Tone identities of the previous and following syllables may also have some influence on pause duration. The means of pause durations under different prosodic environments engendered by different tones and boundary categories are listed in table 2 and table 3. The influence of neutral tone is not included in these tables due to the sparsity and imbalance of neutral tone. For example, neutral tones hardly occur in the first syllable of phrase, so the means of pauses durations under this environment is meaningless. The difference in the means of pause durations under phrase boundary can be neglected on account of the long pause duration under this environment. That is to say the tone identity has little influence on pause duration under phrase boundary. However, under non-phrase boundary, the situation is different: Table 2 shows that the pause duration after syllables with tone3 is much longer than that in the other situations; Table 3 tells us that the pause duration may be lengthened under non-phrase boundary when the next syllable’s tone is tone 1 or tone 4.

![Fig. 1. The influence of the following syllable’s initial category on pause duration](image1)

![Fig. 2. The influence of the previous syllable’s final category on pause duration](image2)

| TABLE2 THE INFLUENCE OF THE PREVIOUS SYLLABLES TONE ON PAUSE DURATION |
|-----------------|-----------------|-----------------|-----------------|
| The tone of previous syllable | Syllable boundary | Word boundary | Phrase boundary |
| Tone1           | 14.5 ms         | 28.6ms         | 119.4ms         |
| Tone2           | 16.0 ms         | 27.9ms         | 107.8ms         |
| Tone3           | 30.4 ms         | 41.3ms         | 120.3ms         |
| Tone4           | 16.2 ms         | 31.1ms         | 118.5ms         |
TABLE 3
THE INFLUENCE OF THE FOLLOWING SYLLABLE'S TONE ON PAUSE DURATION

<table>
<thead>
<tr>
<th>The tone of following syllable</th>
<th>Syllable boundary</th>
<th>Word boundary</th>
<th>Phrase boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone1</td>
<td>22.4 ms</td>
<td>36.4ms</td>
<td>112.1ms</td>
</tr>
<tr>
<td>Tone2</td>
<td>13.6 ms</td>
<td>28.1ms</td>
<td>110.9ms</td>
</tr>
<tr>
<td>Tone3</td>
<td>16.0 ms</td>
<td>27.6ms</td>
<td>117.2ms</td>
</tr>
<tr>
<td>Tone4</td>
<td>18.3 ms</td>
<td>35.9ms</td>
<td>120.8ms</td>
</tr>
</tbody>
</table>

C. Correlation with more context features

Prosodic structure is another important factor that may have great influence on pause duration. Here prosodic structure includes position in word, phrase, and sentence.

Fig. 3 shows the influence of position in sentence on pause duration under different boundaries. In Fig. 3, under all boundaries pause duration becomes longer as the position in sentence approaches the end. This phenomenon can be interpreted in such way, speakers become rather tired as the syllables of a sentence are pronounced one by one, therefore in the posterior of the sentence the pause duration becomes longer for speakers to release pressure.

However, this phenomenon does not occur in Fig. 4 that shows the influence of position in phrase. No matter under syllable boundary or word boundary, we cannot see any explicit relationship between pause duration and position in phrase. But under phrase boundary, where position in phrase equals the length of phrase, the pause duration is lengthened as the length of phrase increases.

Fig. 5 shows the influence of position in word. Under syllable boundary, the change of pause duration, as a function of position in word, is ruleless and stochastic. However, under word boundary, where the position in word equals the length of word, the pause duration and the length of word have some relationship that is similar to linear increase, just as what Fig. 5 shows. This phenomenon is similar to the change of pause duration as the function of phrase length under phrase boundary, which is showed in Fig. 4. One reason for these phenomena could be that when speakers pronounce a large number of syllables successively, he needs comparative long pause to relax.

D. Influenced by pitch gap between two syllables

As an important part of prosodic information, pause has close connection with other prosodic information, like pitch gap across the pause. But in conventional prosody models, the prosodic information itself is always neglected. Therefore, in this section we try to elucidate the relationship between pause duration and pitch gap under different prosodic environments, offering reference for constructing a well-performed prosody model.

It is known that there is some specific connection between pause duration and pitch gap across the pause. According to statistic results, we plot several curves to display how the pause duration changes as a function of pitch gap.

From Fig. 6, we can see that there is some explicit relationship between pause duration and pitch gap under word boundary and syllable boundary. When pitch gap approaches zero, the pause duration is almost minimum, and when the abstract value of pitch gap becomes larger, the pause duration also becomes longer. But this phenomenon is not obvious under phrase boundary, for there are several peaks in this curve. Then we will minutely study the relationship between these two variables under phrase boundary.

Because of the complexity of pitch contour in mandarin, pitch may rise or decline after pause. The correlation between pause duration and pitch rise, and that between pause duration and pitch decline are calculated respectively, as showed in table 4.

The correlation between pause duration and pitch decline is rather little, only –0.03. So when pitch declines after pause,
there is no explicit relationship between pause duration and pitch decline. Just as in Fig. 6, when pitch gap is negative, the curve of pause duration under phrase boundary is similar to the white noise. Meanwhile, the correlation between pause duration and pitch rise is 0.19, showing that there is some specific relationship between these two variables. However, in Fig. 6, when the pitch gap is positive, there is more than one peak in the curve of pause duration under phrase boundary, which shows that the relationship between these two variables is not simple. Given the complexity of pitch contours of various tones, we study this relationship respectively according to various tone combinations.

Previous research proves that, there is one-to-one correspondence between pitch targets and tones [3]. So the four normal tones can be represented by four basic targets: “high” (tone1), “rise” (tone2), “low” (tone3), and “fall” (tone4). Among these rise can be seen as “low”-to-“high”, and “fall” can be seen as “high”-to-“low”. Then the correlation between pause duration and pitch rise under phrase boundary is calculated according to this sorting method.

In Table 5, “LL”, “LH”, “HH”, and “HL” represent all kinds of tone combinations. For example, “LH” represents that the previous syllable’s ending pitch is low, which denotes the tone of this syllable is tone 3 or tone 4, and the following syllable’s starting pitch is high, which denotes the tone of this syllable is tone 1 or tone 4. From the statistic data in table 5, we can see that, when the previous syllable’s ending pitch is low, namely, the tone of previous syllable is tone3 or tone4, there is close relationship between pause duration and pitch rise, the correlations is 0.42, 0.44 respectively. While the previous syllable’s ending pitch is high, this relation is very weak. Fig. 7 shows the curve of pause duration under different tone combinations. When previous syllable’s ending pitch is low, just like Fig. 7(a) and Fig. 7(b), there is almost linear relationship between pause duration and pitch rise, But in Fig. 7(c) and Fig. 7(d), when previous syllable’s ending pitch is high, there is no obvious relationship between pause duration and pitch rise.

IV. CART-BASED PAUSE DURATION PREDICTION

Our final goal is not only to analyze the influence of various factors on pause duration, but rather to precisely predict pause duration in our TTS system. The classification and regression tree (CART) is an effective method to solve this prediction problem. Based on the knowledge of various factors’ influences on pause duration, a precisely cart-based pause model can be constructed.

One more thing worth mention, we do not predict pause duration directly, otherwise we predict the logarithm of pause duration, so the goal of CART is to minimize the mean standard error of the logarithm of pause duration. This method can improve the objective perception of pause prediction. For example, when actual pause duration is 200ms, if the predicting error is 20ms, listener can not feel some unnatural. But if actual pause duration is 20ms and the predicting error is also 20 ms, then the discomfort is very large. Using the logarithm of pause duration as predicting target can resolve this problem to some extent.
We construct two trees (Tree A and Tree B) in which Tree A only uses text information to predict pause duration, while Tree B also uses prosody information besides text information when predicting pause duration. Through the comparison of these two trees, we can get the value of the prosody information in predicting pause duration. Table 6 and Table 7 respectively list the predicting features and results of these two trees.

From these two tables, we can see that adding pitch gap as one predicting feature can largely improve the precision of pause duration prediction. This result also validates our analysis in section III and section IV. For the application of Tree B in TTS system, the pause duration model should be put in the back of pitch model, which can generate precise pitch contour, so the pitch gap between two syllables can be got. This method has been used in our TTS system, which can generate high quality prosody output.

V. CONCLUSIONS

This paper systematically and thoroughly studies various factors that have influence on pause duration in mandarin. The factors include not only the information from the results of text analysis, but also the prosodic information itself, such as pitch gap across the pause. Experiments designed in this paper explicitly reveal the relationship between pause duration and these factors: (1) Under non-phrase boundary, text information that has influence on pause duration includes initial category of the following syllable, final category of the previous syllable, tone identities of the previous and following syllables, and position in sentence. Moreover, word length also has some influence on pause duration under word boundary. (2) Under phrase boundary, only a little text information has influence on pause, such as position in phrase and phrase length. So it’s necessary to make use of the prosodic information for predicting pause duration precisely. (3) Pause duration also has close connection with pitch gap across the pause. Under non-phrase boundary, pause duration is in proportion to the abstract value of pitch gap. Under phrase boundary, however, their relationship becomes more complex. When pitch declines after pauses, the correlation between pause duration and pitch decline is little, only –0.03, which shows there is no explicit relation between these two variables. When pitch rises after pauses, the correlation between pause duration and pitch rise is 0.19, which demonstrates these two variables has some specific connection. And Experiments show this relationship varies according to different tone environments: When previous syllable’s ending pitch is low, which denotes the tone of this syllable is tone 3 or tone 4, there is almost linear relationship between pause duration and pitch rise, but when previous syllable’s ending pitch is high, which denotes the tone of this syllable is tone 1 or tone 2, there is no obvious relationship between pause duration and pitch rise.

Revealing these relationships is not our final goal, and we make use of these results to construct a more precise pause model in text-to-speech system in the section IV. We construct two cart-based pause models to predict pause duration, one use only text information and another also includes prosody information itself. From the comparison between these two models, we can find that prosody information is very useful in predicting pause duration.

REFERENCES