Modeling Incompletion Phenomenon in Mandarin Dialog Prosody

Jian Yu(1), Lixing Huang(2), Jianhua Tao(3), Xia Wang(4)

(1)(2)(3) National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
(4) Nokia Research Centre, China

jyu@nlpr.ia.ac.cn, lxhuang@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn, xia.s.wang@nokia.com

Abstract

The paper proposes a prosody generation method for dialog speech synthesis in Mandarin. The method is an extension of a prosody model for read speech and also takes the essential characteristic of dialog speech into account. Besides the faster speaking rate and narrower pitch range in dialog speech, our method concentrates on the more underlying and essential characteristic: the incompletion of pitch contour within a syllable and its impacts on adjacent syllables. To simulate this phenomenon, a CART-based method is constructed to predict whether a syllable is incomplete or not. Based on that, a prosody generation model which focuses on the prosody constraint between adjacent syllables is constructed, and this method can simulate the influence of incomplete syllable on adjacent syllables. Experiments show that the synthesized results based on this prosody model sound much more natural and colloquial.

Index Terms: speech synthesis, prosody, dialog speech

1. Introduction

As the development of corpus-based speech synthesis system, the synthesized result has achieved a high overall quality, but still lacks naturalness. To make the synthesized speech sound more natural and colloquial, some effects that are observable in spontaneous speech should be integrated. As we know, there are much more linguistic, para-linguistic, and even non-linguistic phenomena in spontaneous speech compared with read speech. Researchers across the world have shown their interests in the simulation of these phenomena: Werner used a pronunciation sequence model to simulate the prosodic variation in spontaneous speech [1]; Kiriyama tried to control the prosodic focus in reply speech generation for a spoken dialog system [2]; Campbell used phrase-sized segments for the concatenative synthesis of conversational speech [3]; Sundaram proposed a data driven approach for modeling both speaker-independent and speaker-dependent spontaneous features at lexical and acoustic levels [4]. Adell analyzed the changing pattern of pitch and duration in repetitions and filled pauses [5]. Carlson investigated acoustic correlates to perceived hesitation in spontaneous speech [6].

Due to the large variability in the features of spontaneous speech, the goal of this paper is only restricted to the simulation of the dialog speech without any para-linguistic or non-linguistic features. The most obvious characteristic in that kind of dialog speech is the narrower pitch range and faster speaking rate compared with read speech. However, only these two points can not truly reflect the essential prosody characteristic of dialog speech. In this paper, we argue that the most essential difference between dialog speech and read speech is the incompletion of pitch contour within a syllable and its impacts on adjacent syllables. Dialog speech tends to be quicker and less strict, so in some situations the syllable cannot complete the whole articulatory process before it ends, which makes the pitch contour be only part of corresponding pitch contour in read speech. That’s what we called incompletion. The incompletion not only makes the current syllable’s shape different from the corresponding one in read speech, but also influences adjacent syllables. To simulate the incompletion phenomenon, a CART-based method is constructed to predict whether a syllable is incomplete or not in dialog speech. Based on that, a prosody generation model which focuses on the prosody constraint between adjacent syllables is constructed, and this method can simulate the influence of incomplete syllable on adjacent syllables. Because our method catches the essential prosody characteristic of dialog speech, the synthesized result has achieved a high overall quality, but still lacks naturalness. To make the synthesized speech sound much more natural and colloquial.

The rest of this paper is organized as follows: section two illustrates the prosody characteristic of dialog speech, especially for the incompletion phenomenon. Section three introduces how to model this phenomenon in the prosody generation method, including CART-based syllable type prediction and pitch generation method based on prosody constraint. Section four is experiment which shows the performance of the proposed method. Finally, section five is conclusion and future work.

2. The dialog speech’s prosody characteristic

To extend current read-style prosody model for modeling dialog-style prosody, our approach focuses on the essential prosodic difference between dialog speech and read speech, which can be obtained from the comparison between parallel dialog corpus and read corpus.

2.1. The corpus introduction

The corpus used in this paper includes both dialog speech and read speech, which are based on the same transcript selected from real conversation. A female speaker is asked to read these sentences using both read style and dialog style. In the recording of dialog speech, another person is asked to make conversation with the speaker to help her produce natural dialog speech. Through the analysis on this kind of corpus, the essential characteristic of dialog speech can be obtained.
2.2. Faster speaking rate and narrower pitch range

Compared with read speech, the most obvious characteristics of dialog speech are the faster speaking rate and the narrower pitch range. However, are those characteristics enough to reflect the difference between dialog speech and read speech? A simple linear regression method can be used to answer that question. Suppose the relationship can be represented as follows:

\[
\text{Pitch}_{\text{Dialog}} = a_1 \times \text{Pitch}_{\text{Read}} + b_1; \\
\text{Duration}_{\text{Dialog}} = a_2 \times \text{Duration}_{\text{Read}} + b_2.
\]

The parameters \((a_1, b_1, a_2, b_2)\), which describe the relationship between dialog speech and read speech, can be obtained based on linear regression. Table 1 shows the performance of that simple model. From this table we can see that the performance of pitch transformation is not quite good, which means only those characteristics can not truly reflect the essential difference between dialog speech and read speech.

<table>
<thead>
<tr>
<th>Table 1: The predicted result of linear regression model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Value</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>24.8 (hz)</td>
</tr>
<tr>
<td>0.76</td>
</tr>
</tbody>
</table>

2.3. The description of incompletion phenomenon

The meaning of incompletion phenomenon is that, the pitch contour within a syllable in dialog speech is only part of the corresponding one in read speech in some cases. The reason for the phenomenon is that some syllables in dialog speech cannot complete the whole articulatory process because of the faster speaking rate and less strict speaking style. This phenomenon affects not only current syllable, but also adjacent syllables.

Fig 1 shows an example of the incomplete syllable and its impact on adjacent syllables. The incomplete syllable in dialog speech is the second syllable with tone 4. In the read speech, the pitch contour of that syllable rises first and then falls. In the dialog speech, the contour also starts with rising trend, but ends before it gets to the inflexion point because of the faster speaking rate and the shorter syllable length, which make the contour only show the rising part. In addition, this phenomenon also leads to some difference in the next syllable with tone 2. As we know, the essential trend of tone 2 is rise. In the read speech, the third syllable starts with falling trend because of the influence from the previous one, so the pitch contour falls first and then rises. However, in the dialog speech, because the pitch contour already starts with rising trend, it immediately rises without falling first.

Figure 1. An example of the incompletion phenomenon

(a) Read speech

lun4: rise first and then fall
yi2: fall first and then rise

(b) Dialog speech

lun4: rise
yi2: rise

Figure 2. An overview of the dialog prosody generation model

3. The prosody generation model

Considering the existence of the incompletion phenomenon, we try to simulate that phenomenon by extending current read-style prosody model. Fig 2 shows an overview of our prosody generation model for dialog style, which will be introduced step by step in this section.

3.1. The automatic label of incomplete syllable

As mentioned above, some syllables in dialog speech end before their pitch contours get to the inflexion point, which changes the pitch contour from original parabolic shape to linear shape. This observation is used as the guideline for the label of incomplete syllable. The incomplete syllables are defined as those who have linear pitch contours in dialog speech, while have parabolic or more complicated contours in read speech.

The Levenberg-Marquardt (LM) Algorithm is used here to determine the shape of pitch contours: linear or more complicated. The procedure is as follows: suppose the pitch contour within a syllable can be represented as a linear equation: \( p = a \times i + b \). According to LM algorithm, the value of \((a, b)\) and the fitting error can be obtained. The error is represented as follows:

\[
\text{err} = \frac{1}{N} \sum_{i=1}^{N} \left( p_i - (a \times i + b) \right)^2
\]

Where \( p_i \) is the real pitch value, \( a, b \) are parameters of the model, \( N \) is the number of points on the pitch contour. This fitting error is used to determine the pitch contour shape. If it is smaller than a threshold, the pitch contour can be modeled as linear shape; otherwise, it can not be modeled well as linear shape.

We manually label some syllables to indicate whether their pitch contours can be modeled by linear function or not. A reasonable threshold can be calculated based on the manual labels. After getting the threshold, all syllables both in read speech and dialog speech are labeled automatically. As mentioned before, the syllable which can be modeled well as linear function in dialog speech but can not in read speech is labeled as incomplete syllable. All the labeling procedure is done offline.

(a) Read speech

lun4: rise first and then fall
yi2: fall first and then rise

(b) Dialog speech

lun4: rise
yi2: rise

Figure 1. An example of the incompletion phenomenon

The meaning of the transcription: talking about a question.
3.2. CART-based syllable type prediction

With labeled results, the incomplete syllables are picked up from the corpus and the other syllables are considered as normal syllables. The classification and regression tree (CART), which can make use of both continual feature and discrete feature, is built to predict whether a syllable is incomplete or not. The input features of CART include: Left tone context, right tone context, position in word, position in phrase, position in sentence, left phonetic context, right phonetic context, word length, phrase length, and so on. All of them are directly extracted from synthesized text. The predicted result will guide the generation of pitch contours further. Table 2 shows the syllable type prediction result by CART.

By looking through the generated tree’s structure, we can find that the incomplete syllable appears very often in the middle of prosodic word. This observation is consistent with our priori knowledge. The syllable in the middle of prosodic word is always spoken in a relatively fast and loose way. According to the STML theory [7], the strength of that kind of syllable is weak. All these make them more likely to be incomplete.

Table 2. The syllable type prediction

<table>
<thead>
<tr>
<th>Real</th>
<th>Predicted</th>
<th>Normal</th>
<th>Incomplete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>85.3%</td>
<td>14.7%</td>
<td></td>
</tr>
<tr>
<td>Incomplete</td>
<td>21.2%</td>
<td>78.8%</td>
<td></td>
</tr>
</tbody>
</table>

3.3. Current read-style prosody model

The proposed dialog-style prosody model is an extension of a prosody model for read style[8, 9], whose basic principle is that there is strong mutual constraint between adjacent syllables. In [9], the existence of the mutual constraint in read speech has been strictly proved. Because of the faster speaking rate, that constraint is even stronger in dialog speech.

The essential idea of that model is the definition of concatenation cost, as illustrated in Fig 3. The concatenation cost is used to measure the naturalness of pitch contours between adjacent syllables. CART-based models are constructed to predict current syllable’s prosody parameters based on not only context features but also adjacent syllable’s candidate prosody parameter values. The predicted values by that method can be considered as the expected values by adjacent syllables, thus the differences between the predicted values and the real values reflect the naturalness of pitch contours between these two adjacent syllables. Therefore, the value of concatenation cost is defined as the weighted sum of differences between predicted values and real values of the four boundary prosodic parameters: starting value, ending value, starting derivative value and ending derivative value, noted as \( F_{0_s}, F_{0_e}, F_{0_d}, \) and \( F_{0_{de}} \) respectively. More detailed information about this definition of concatenation cost can be found in reference[8, 9]. In the prosody generation for dialog speech, this concatenation cost can be used to simulate the influence of the incomplete syllable on adjacent syllables.

3.4. The prosody parameter prediction

Based on the CART-based syllable type prediction result, two kinds of function are used to depict the pitch contour within a syllable. For the normal syllable a two-order polynomial function is adopted, and for the incomplete syllable a linear function is adopted to limit the shape of the pitch contour. More complicated functions such as three-order polynomial function are not involved because the pitch contour with more than one inflexion point is so infrequent in dialog speech. The equations are as follows:

Normal syllable: \( at^2 + bt + c \)
Incomplete syllable: \( mt + n \)

Where \( t \) represents time from the start of current syllable, and \((a, b, c, m, n)\) are coefficients need to be determined.

While predicting the coefficients, to make good use of the definition of concatenation cost, the coefficients \((a, b, c, m, n)\) are not predicted directly. Instead, meaningful prosody parameters mentioned in the concatenation cost are predicted first and then the coefficients are calculated based on these parameters. For the pitch contour with parabolic style, the four boundary features, including \( F_{0_s}, F_{0_e}, F_{0_d}, \) and \( F_{0_{de}} \), are all involved. While for the pitch contour with linear style, only \( F_{0_s} \) and \( F_{0_e} \) are involved, as Fig.4 shows.

Based on the statistic on the large corpus, a number of possible candidate values for each prosodic parameter mentioned above are preset according to contextual information, as Fig.4 shows. For every possible combination of parameter values in the concatenation point, a concatenation cost value can be calculated. By minimizing the whole sentence’s concatenation cost, all prosody parameter values can be obtained.
there are four equations but only three unknown coefficients, so no exact solution can be got. The Levenberg-Marquardt (LM) algorithm is adopted here to get an approximate solution. With all of the coefficient values, the whole pitch contour can be generated. Finally, some pitch-jitters can be added to avoid over-smoothness.

4. Experiment

4.1. The perception of dialog style

Currently, the paper focuses on how to simulate the prosody characteristic of dialog speech, while remains the spectral part untouched. In this section, a perception experiment is done to show which feature is more important for the perception of dialog style: prosodic feature or spectral feature?

In this experiment, four kinds of speech are compared. Among these, the first two kinds are the recorded read speech and dialog speech introduced in section 2.1, and the other two kinds are synthesized results. One is synthesized based on spectrum with dialog prosody and style with read style, and another is based on spectrum with read style and prosody with dialog style. Fifteen persons, all of them under-graduate university students, are asked to provide a rating between 0 and 5, in which 0 means the sentence sounds like read style, and 5 means the sentence sounds like dialog style. Table 3 shows the rating result.

Table 3: Rating result for different kinds of speech

<table>
<thead>
<tr>
<th></th>
<th>spectrum</th>
<th>prosody</th>
<th>Rating result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kind 1</td>
<td>Dialog</td>
<td>Dialog</td>
<td>4.5</td>
</tr>
<tr>
<td>Kind 2</td>
<td>Read</td>
<td>Read</td>
<td>0.9</td>
</tr>
<tr>
<td>Kind 3</td>
<td>Dialog</td>
<td>Read</td>
<td>1.6</td>
</tr>
<tr>
<td>Kind 4</td>
<td>Read</td>
<td>Dialog</td>
<td>3.7</td>
</tr>
</tbody>
</table>

From this table we can see that, the synthesized results with dialog prosody and read spectrum sound much more like dialog style than that with read prosody and dialog spectrum, which proves that the prosody feature is more important than the spectral feature in the simulation of dialog style.

4.2. The evaluation for the presented model

A subjective evaluation experiment is done to test the performance of the proposed dialog prosody generation model. In this experiment, the proposed model is compared with a simple transformation model. In the simple model, the pitch contour with read style is generated first, and then the linear transformation method introduced in section 2.2 is adopted to transform read style to dialog style. In both systems, the duration is transformed as linear relationship and the spectrum information comes from a same HMM-based engine. The same fifteen persons are asked to rate these sentences between 0 and 5 to show which sentence sounds more natural and colloquial. Table 4 shows the rating result. From this table, we can see that the presented method can generate much more natural and colloquial pitch contours.

Table 4: 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 LR-based method</td>
<td>2.8</td>
</tr>
</tbody>
</table>

Besides this subjective evaluation, another objective evaluation experiment is done. The two methods are compared in terms of RMSE and shape similarity. Among these, the shape similarity is measured by the correlation of pitch contour within a syllable between real dialog pitch contour and synthesized one, which reflects the prediction result of incomplete syllable position. Table 5 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.

Table 5: 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>19.5</td>
<td>0.81</td>
</tr>
<tr>
<td>The LR-based method</td>
<td>27.6</td>
<td>0.70</td>
</tr>
</tbody>
</table>

5. Conclusion

This paper proposes a prosody generation method for dialog speech without any para-lingual and non-lingual phenomenon in Mandarin. The essential idea of this method is the simulation of the incomplete syllable and its influence on adjacent syllables. The experiment results prove the reasonability of that method. The final goal of our work is to construct a truly spontaneous speech synthesis system. The simulation of the dialog style in this paper is just a beginning, which builds a strong foundation for future work. In the next step, kinds of para-linguistic and non-linguistic effects will be integrated with that dialog speech synthesis system, which can finally make TTS system generate truly spontaneous speech.

6. Reference