A NEW PITCH GENERATION MODEL BASED ON INTERNAL DEPENDENCE OF PITCH
CONTOUR FOR MANADRIN TTS SYSTEM

Jian Yu     Wanzhi Zhang    Jianhua Tao

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
{jyu, wzzhang, jhtao}@nlpr.ia.ac.cn

ABSTRACT

The paper presents a new pitch generation model based on
internal dependence of pitch contour. This model pays more
attention to the impact of adjacent syllables’ pitch contours on
the current one. A new definition of concatenation cost is
presented to measure the naturalness of pitch contours between
every two adjacent syllables. Based on this definition, the model
concentrates on how to remove unnatural pitch contours across
concatenated points which are always the most unstable parts in
current TTS systems. This model can generate natural, fluent
pitch contours and was proved to be able to catch the essential
nature of pitch contour.

1. INTRODUCTION

During the past few years, with the development of the
technology in speech processing, the Text-to-Speech system has
made rapid progress. The synthesis quality has been highly
improved, but the production of a natural prosody still remains a
difficult and challenging problem. Many automatic prediction
methods have already been used for this topic, including
decision trees, neural networks, and HMMs [3][4][5][6]. They
resulted in much better synthetic speech than the traditional
rule-based approach. However, there are still some points need
improvement in these methods. In the synthesized speech by
current TTS systems, the most unstable part is the concatenated
point. In this place, the behavior of pitch contour is more likely
to be unnatural. Current pitch models seldom concentrate on
this point, which consequently leads to unnatural outputs.

To resolve this problem, a new definition of concatenation
cost is presented to measure the naturalness of pitch contours
between every two adjacent syllables. This definition is based
on the internal dependence of pitch contour, making full use of
the impact of adjacent syllables’ pitch contours on the current
one. The cost value is got by statistical model trained on natural
pitch contours, which make the result more convincing. By
minimizing the concatenation cost of the whole sentence, there
will be no unnatural pitch contours between every two adjacent
syllables. In addition, another notion of overall cost is presented
to depict a syllable’s pitch register. Minimizing the overall cost
of the whole sentence can make the output pitch contour similar
with the natural sentence on overall trend.

Because Chinese is a tonal language and syllable is
normally assigned as the basic prosody unit in processing, it is
much suitable to use prosody templates selection method [9].
Based on the definition of the cost, viterbi search can select the
best template series which has the minimal cost value, just like
the searching procedure in unit selection [7][8]. Through this
method, the output pitch contour can be natural in both local
area and overall trend.

The structure of this paper is organized as follows: section
two and section three introduces the extraction and the
pre-selection of prosody templates respectively. In section four,
the definition of the cost is introduced and analyzed in detail,
which is the key issue of this paper. Based on this definition, the
best template series is selected through viterbi search. Section
two makes an evaluation and discussion to show why this
model performs well. Section six is conclusion, listing the
deficiency of this model and the future work.
3. THE PRE-SELECTION OF PROSODY TEMPLATES

In the prosody templates selection, a syllable always has so many candidate templates that selecting the best path from all candidates is a very time-consuming task, so a pre-selection procedure should be done first. The scale for this pre-selection procedure is contextual information difference (CID), which depicts the difference of contextual information between the current candidate template and the synthesized syllable. Suppose that the number of contextual information category is $n$, the formula is as follows:

$$CID = \sum_{i=1}^{n} W_i * D_i$$

Where $D_i$ is the difference of the $i$th contextual information between the current candidate template and the synthesized syllable, and $W_i$ is the weight of the $i$th contextual information.

There are several flaws in this definition of CID: first, both the definition of $D_i$ and the setting of $W_i$ are based on experience; Second, linear relationship among different contextual information is not satisfied. Therefore, the candidate template whose contextual information is most similar with the synthesized syllable is not the most appropriate template. However, because this procedure is just a pre-selection procedure, we don’t expect its results to be very precise. It is satisfying as long as one or more appropriate templates could be included in pre-selection results, which can be achieved by properly setting the number of pre-selection results even the unreasonable definition of CID.

4. THE BEST TEMPLATE SERIES SELECTION BASED ON THE COST DEFINITION

After the pre-selection of prosody templates, viterbi search algorithm will be used to select the best template series. The most important issue in this procedure is how to reasonably define the cost function, which is also the key point of this paper. As stated before, our cost definition consists of two parts: concatenation cost and overall cost.

4.1 Concatenation Cost

A reasonable definition of the concatenation cost should be applied to the measurement of the naturalness of pitch contours between adjacent syllables. The simplest definition is the difference between the previous syllable’s $F0_i$ and the current syllable’s $F0_0$, which is based on the assumption that the pitch contour is always continuous on the whole sentence. Because the larynx can not change its state instantaneously, this assumption should be tenable in natural speech. However, in the situation that there exists a certain length of silence or several voiceless initials, the whole curve will be separated into a few isolated parts. Therefore, the simplest definition is tenable only when the current syllable’s initial is nasal, lateral or zero-initial, just as figure 2(a) shows. If the current syllable’s initial is voiceless such as ‘h’ or ‘f’, this definition is obviously irrational, as figure 2(b) and figure 2(c) shows, in these cases, there are always pitch jumps across the syllable boundary. However, even in that situation, we also noticed that there is some relationship between the current syllable’s $F0_0$ and the previous syllable’s pitch contour. It seems that the pitch contour is virtually connected across the silence and voiceless initial. For example, in figure 2(c), the pitch end of the previous syllable tends to stretch out across the span of the silence and the initial, reaching the pitch head of the current syllable. Similarly, the following syllable’s pitch contour also has some impacts on the later portion of the current syllable, as recent research shows [1]: the current syllable’s pitch contour will be raised if followed by a syllable with low pitch register. All in all, because of the existence of the internal dependence of pitch contour, the current syllable’s pitch contour is greatly dependent on adjacent syllables’ pitch contours and other prosodic features such as pause length and voiceless initial length.

Based on these facts, our definition of the concatenation cost makes full use of these prosodic features, especially adjacent syllables’ pitch contours. Template parameters involved in the concatenation cost include $F0_i, F0_0, F0_0$, and $F0_0$, which can be considered as boundary features of a syllable’s pitch contour. According to the above analysis, when we predict these four template parameters, the prosodic features listed in Table 1 should be included, which will greatly improve the predicted results. Classification and Regression Tree (CART) is used as the predicting model. Table 2 show the predicted results of these template parameters with these prosodic features included and not. From this table we can see that including prosodic features in the tree’s feature set has greatly improved its performance in terms of both RMSE and correlation, which is one of the most important reasons for the success of this model.

<table>
<thead>
<tr>
<th>features in predicting $F0_i$ and $F0_0$</th>
<th>features in predicting $F0_i$ and $F0_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequently used text information</td>
<td>frequently used text information</td>
</tr>
<tr>
<td>previous syllable’s $F0_i$ and $F0_0$</td>
<td>following syllable’s $F0_i$ and $F0_0$</td>
</tr>
<tr>
<td>length of pause before current syllable</td>
<td>length of pause after current syllable</td>
</tr>
<tr>
<td>length of current syllable’s initial</td>
<td>length of following syllable’s initial</td>
</tr>
</tbody>
</table>

Table 1: features used in predicting $F0_i, F0_0, F0_0$, and $F0_0$.
Prosodic features included | Prosodic features excluded
---|---
RMSE | RMSE | Correlation | Correlation
\(F_0_s\) | 24.7hz | 0.92 | 32.8hz | 0.84
\(F_0_e\) | 23.4hz | 0.91 | 35.2hz | 0.81
\(F_0_{\text{dd}}\) | 0.36hz/ms | 0.75 | 0.45hz/ms | 0.61
\(F_0_{\text{df}}\) | 0.34hz/ms | 0.78 | 0.49hz/ms | 0.63

Table 2: comparison of the predicting results of \(F_0_s, F_0_e, F_0_{\text{dd}}\) and \(F_0_{\text{df}}\)

Worth particular mentioning is that the predicted value by this method can be considered as the expected value by adjacent syllables, so the difference between these predicted values and real values can be used to measure the naturalness of pitch contours between adjacent syllables. As equation (2) demonstrates, the value of the concatenation cost is defined as the weighted sum of differences between predicted values and real values of these four template parameters, noted by \(DF_0_s, DF_0_e, DF_0_{\text{dd}}\) and \(DF_0_{\text{df}}\) respectively. Figure 3 schematically illustrated the definition of the concatenation cost. This definition is perfectly reasonable because it makes full use of the internal dependence of pitch contour.

\[
\text{concatenation\_cost} = w_s \cdot DF_0_s + w_e \cdot DF_0_e + w_{\text{dd}} \cdot DF_0_{\text{dd}} + w_{\text{df}} \cdot DF_0_{\text{df}} \tag{2}
\]

From Table 2, we can also see that the predicted results of \(F_0\) derivative are a little poorer than those of \(F_0\) value, which might be resulted from the method of pitch derivative calculation. The calculation procedure is as follows: first, the pitch contour is smoothed by moving average method, which is used to calculate the raw pitch derivative. Then the raw pitch derivative also needs to be smoothed to get the final pitch derivative value. How to improve the calculating method for \(F_0\) derivative is one of unresolved problems in our pitch model.

In summary, the concatenation cost depicts the naturalness of pitch contours between every two adjacent syllables. By minimizing the concatenation cost, this model can make sure that there are no unnatural pitch contours across all syllable boundaries.

### 4.2 Overall Cost

Parameters involved in the overall cost include \(F_0_s, F_0_e, F_0_{\text{dd}}\) and \(F_0_{\text{df}}\), which denote the pitch register and the pitch range. Like the concatenation cost, the value of overall cost is also defined as the difference between predicted value and candidate templates’ real value of these three parameters, noted as \(DF_0_s, DF_0_e\) and \(DF_0_{\text{df}}\).

\[
\text{overall\_cost} = w_s \cdot DF_0_s + w_e \cdot DF_0_e + w_{\text{df}} \cdot DF_0_{\text{df}} \tag{3}
\]

These three parameters are predicted by traditional method which predicts prosodic information only from text information. CART is also used as the predicting model. Table 3 lists the predicted results by CART. From this table, we can see that all the three parameters can be predicted precisely.

The overall cost is presented to depict the overall trend of the pitch contour. For example, pitch declination in naturally read discourses can be realized by all the three parameters descend as the sentence approaches the end. Minimizing the overall cost can make the output pitch contour similar with the natural one on overall trend.

\[
\begin{align*}
F_0_s & \quad \text{RMSE} \quad \text{Correlation} \\
& \quad 26.6hz \quad 0.88 \\
F_0_e & \quad 32.1hz \quad 0.83 \\
F_0_{\text{df}} & \quad 25.8hz \quad 0.89
\end{align*}
\]

Table 3: The predicted results for \(F_0_s, F_0_e\) and \(F_0_{\text{df}}\) by CART

### 4.3 Best Template Series Selection

All in all, our cost definition is comprised by two parts: concatenation cost and overall cost. The formula is as follows:

\[
\text{COST} = \text{concatenation\_cost} + \text{overall\_cost}
\]

\[
\begin{align*}
&= w_s \cdot DF_0_s + w_e \cdot DF_0_e + w_{\text{dd}} \cdot DF_0_{\text{dd}} + w_{\text{df}} \cdot DF_0_{\text{df}} \\
&\quad + w_s \cdot DF_0_s + w_e \cdot DF_0_e + w_{\text{df}} \cdot DF_0_{\text{df}} \tag{4}
\end{align*}
\]

As seen in equation (4), there are so many prosodic parameters included in our definition that it’s hard to assign appropriate weight for each parameter. For simplicity, the weights are assigned according to experience. For example, a larger weight for starting \(F_0\) value should be assigned than that for ending \(F_0\) value because previous syllable’s pitch contour has more impact on current one than next syllable does.

Based on this cost definition, Viterbi search algorithm can select the best template series whose cost is minimized, which is similar with the search procedure in unit selection. In summary, minimizing the concatenation cost can make sure that there are no unnatural pitch contours between every two adjacent syllables. And minimizing the overall cost can make the output pitch contour similar with the natural pitch contour on overall trend. So under various constrains, this model outputs very natural pitch contours.

### 5. EVALUATION AND DISCUSSION

In our evaluation experiment, our new pitch model is compared with other two primitive models. Among these, Model I excludes Viterbi search module, in which the template with minimal CID is selected as the output template. Model II uses...
the simplest definition of concatenation cost introduced in section 4. Evaluation task is done by two ways: one subjective test with MOS which shows human’s perception feelings and one objective test with correlation and RMSE between real and synthesized pitch contours, in which correlation indicates the similarity in shape and RMSE indicates the characteristic divergence.

The corpus this paper used contains 6000 sentences, in which 5000 sentences are used for training, and others are used for open test, among these sentences, only 200 are used in MOS scoring for simplicity. Table 4 shows the comparison results, which reveals that the new model generates much more natural pitch contours than the others.

Take the sentence “qi2 mo2 tuo1” as an example to give a further discussion about the performance of these models. Figure 4(a) outlines the real pitch contour taken from speech corpus. Figure 4(b) shows the same sentence’s pitch contour generated by Model 1. Because the definition of CID is not perfectly reasonable and the internal dependence of pitch contour is not considered, F0 gaps at syllable boundaries are not well modeled, bringing an unnatural result. Figure 4(c) show the output of module II. Based on the simplest cost definition, Viterbi search will make all syllables’ starting F0 values similar with the ending F0 values of previous syllables. So in the case of a syllable with an unvoiced initial, the pitch jump which should appear in natural speech is absent. Figure 4(d) shows the output of the new pitch model. The expected starting F0 value of third syllable is predicted with the help of the second syllable’s pitch contour, which makes the result more convincing. With this method, the pitch jump happening in natural speech is well simulated. From this example, we can conclude that the definition of the concatenation cost plays an important role in pitch generation.

6. CONCLUSION

This paper presents a new pitch generation model for mandarin TTS system based on the internal dependence of pitch contour. By concentrating on this internal dependence, the model can make sure that there is no unnatural pitch contours between every two adjacent syllables. At the same time, minimizing the overall cost can make the output pitch contour be natural on overall trend. Under the effects of these two factors, this model can generate natural and fluent pitch contours.

However, this model also has some shortcomings. Firstly, the predicted results for pitch derivative are not very well, and the correlation is only around 0.75. This phenomenon leads to that the weight of pitch derivative in the definition of cost should not be very large, so in fact, the information of pitch derivative is not fully utilized. Secondly, in the definition of cost, the weights of all prosodic parameters are assigned based on experience, which also leads to inefficiency. How to design an algorithm to assign weight automatically is another part to be improved.

One more thing need attention, the prosody templates selection module has some similarities with the unit selection module. So if duration and spectral information can be included in the definition of the cost, this pitch model can evolve to be a new unit selection model, which is also a guideline for our future research topic.

7 REFERENCES

[8] Renhua Wang, Zhongke Ma, Wci Li, and Donglai Zhu, ” F0 Prediction Model of Speech Synthesis Based on Template and Statistical Method " ICSLP2000, Beijing, CHINA, 2000

<table>
<thead>
<tr>
<th>Objective evaluation</th>
<th>Subjective evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
</tr>
<tr>
<td>Model I</td>
<td>53hz</td>
</tr>
<tr>
<td>Model II</td>
<td>40hz</td>
</tr>
<tr>
<td>New Model</td>
<td>24hz</td>
</tr>
</tbody>
</table>

Table 4: comparison results for these three models

Figure 4: an example which shows the performance of these models