PITCH CONTOUR MODEL FOR CHINESE TEXT-TO-SPEECH
USING CART AND STATISTICAL MODEL

Minghui Dong, Kim-Teng Lua
{dongming, luakt}@comp.nus.edu.sg
School of Computing, National University of Singapore
S15, 1 Science Drive 2, Singapore 117543

ABSTRACT

This paper describes an approach to generating prosody parameters for Mandarin Chinese text-to-speech system. The Chinese fundamental frequency contour is decomposed into two parts, a global intonation contour and a syllable level tone contour. The global intonation contour is converted to pitch target labels in corpus. It is predicted by first predicting pitch target labels using statistical model and classification tree, and then the labels are converted into real pitch values. The local syllable level tone contour is classified into a definite number of contour types using clustering approach. The prediction of local syllable pitch contour is done by classification tree approach. Experiment shows that this approach achieves an accurate prediction result.

1. INTRODUCTION

A text to speech system (TTS) is a system that transforms text sentence into speech utterance. Application of TTS is desirable for an automatic system that responds human inquiries. The quality of speech is determined by many factors. One of the most important factors is the prosody that is being used to generate the speech.

Prosody is a collection of speech parameters that describes the variations of duration, pitch and loudness of speech units. Prosody plays a very important role in natural speech. To make the speech as natural as that of a human, prosody of the generated speech is needed to be as similar as possible to that of a human speech. Generation of good prosody needs the improvements in two aspects. First, good understanding of text is essential for prosody. The text analysis generates some form of linguistic representation that consists of phonetic, lexicon and syntax information. Second, a good model is needed to convert the linguistic representation into prosody parameters that make the speech more natural.

There were many approaches used for generating Chinese prosody. Early systems generate prosody using some simple rules [1]. Currently, most of the prosody model for Chinese and other languages use statistical approaches, such as neural networks[2], CART[3][4], probabilistic approach[3]. In this research, our task is to generate pitch contour for Chinese speech from Chinese text. And we adopted CART and Markov chain in our model.

Chinese prosody is different from that of English in many ways. (1) Chinese is a syllable-based language. Each Chinese character is pronounced as a syllable. There are about 400 basic syllables in Chinese. (2) Chinese is a tonal language, in which there are 4 typical tones and one neutral tone. Same basic syllable with different tones usually carries different meanings. Considering the tone differences, there are around 1300 meaningful tonal syllables in Chinese. (3) Examining the prosodic structure of Chinese, a Chinese utterance can be cut into small prosodic units, which are prosodic words, minor prosodic phrase, and intonation phrase. (4) Theories on Chinese prosody found that a Chinese utterance has a global intonation and each syllable has a tone identity[6].

Based on the above properties of Chinese speech, Chinese pitch contour can be divided into two parts, one is a local part that represents tone, the other is a global part that reflects the general intonation of sentence. The final pitch contour of an utterance is the sum of the two components.

2. CORPUS

As we use corpus-based approach to generate prosody, a speech corpus with linguistic features and its prosody features should be available.

The text script of the corpus in this research was initially from Chinese newspapers. To have a good coverage of different types of text, we selected different types of articles. 40 articles are selected for recording of the speech. A male speaker read each the article in a neutral manner with a normal speed in laboratory environment. The speech was recorded with a digital recorder with a 16-bit precision and a 16k Hz sampling rate. Totally around 5000 utterances and around 42000 syllables are collected.

The script of the corpus is processed to meet the requirement of training. The text is first normalized. The symbols and numbers are all converted into Chinese text. All non-Chinese symbols and numbers in the raw Chinese text are converted into pure Chinese text. As there is no space between words in Chinese like western languages, the Chinese text is...
segmented into word sequence by a word segmentation process. Next the part-of-speech tagging program tags each word with a part-of-speech category. The Chinese characters are also converted into a sequence of pinyin strings. The above processing was done under the assistance of some programs. The word segmentation and POS tagging errors are corrected manually by human.

With the above information available, the rhythm structure of Chinese sentence is determined. In this research, Chinese sentence is decomposed into the following prosodic units. Syllable is the smallest unit. Syllables are combined together to form words. Words are further combined to form prosodic words. Prosodic words form longer unit, prosodic phrase. And one or more prosodic phrases form sentence. The detection of prosodic words and prosodic phrase was done by programs and some obvious errors were corrected manually.

For the speech part, we segment the speech data using speech recognition techniques. First we translate all the text transcripts into phonetic transcripts. We use 14th order CEP and 14th order delta-CEP as our speech recognition parameters. A frame length of 10 ms is used. An HMM model is defined for each Chinese syllable. There are 408 models build for 408 toneless syllables in Chinese in our research. We train the HMM models using all the speech data in our corpus. The training processing is actually a processing of segmentation. After some iterations of training, the boundaries of syllables are very accurate. We record the corresponding start and end frame number of each HMM, and convert the frame numbers into time. By this way the start point and end point of each syllable are obtained. The segmentation result is then manually checked and obvious mistakes are corrected. Then the pitch contour is calculated based on the segmented speech.

3. PITCH CONTOUR MODEL

The generation of prosody from text is in fact to find a mapping from the linguistic information to prosodic parameter information. The mapping is learned from the training data, the speech corpus. The prosody generation model in this research works in the following way. (1) Text analysis (2)Prosodic break determination (3)Prosody parameter generation part. In this paper, we use a model for the prosody to do prosody parameter generation. Particularly, we are interested in the pitch contour of prosody. The pitch contour is decomposed into two parts, therefore there are following steps in generating pitch contour. (1) Global pitch contour prediction (2) Syllable level local pitch contour prediction (3) Reconstruction of whole pitch contour.

3.1 A Superpositional Model

In this research, we use a superpositional model for Chinese pitch contour. Pitch contour is decomposed into two parts. The pitch contour is considered as the sum of global intonation contour and syllable level tone contour.

- **Global intonation contour**: Global intonation contour means the global change of pitch values over the syllables in a sentence. It controls the whole intonation of an utterance. The global contour is determined by the grammatical function and pragmatic function of each word and phrase in the sentence.

- **Local syllable tone contour**: Syllable F0 contour means the local change of pitch values in a syllable. It controls the tone identity of a syllable and therefore determines the meaning of words. Syllable contour is usually determined by tones of surrounding syllables, whether the syllable is stressed, etc.

Suppose the F0 contour for the voice part of the ith syllable is \( f_i(t) \) and \( s_i \) and \( e_i \) are end time and start time of the voiced part of the syllable in a sentence of \( n \) syllables. Then the mean pitch value of the ith syllable is \( p_i = \left( \frac{1}{e_i - s_i} \int_{s_i}^{e_i} f_i(t) dt \right) \). We call \( p_i \) pitch level of the syllable (PL). The global intonation contour of a sentence is defined as \( \{ p_1, p_2, \ldots, p_n \} \). And the syllable tone contour of syllable \( i \) is \( s_i = f_i(t) - p_i \). In this research, tone contour is expressed using a vector. We get \( n \) samples in the pitch contour evenly to form an \( n \) dimensional vector. This gives a uniform representation of all syllable pitch contour. As this syllable pitch contour is in fact a representation of the tone of this vector, we name it tone vector (TV).

The two parts, \( p_i \) and \( s_i(t) \), are separately predicted and then combined together to reconstruct the whole pitch contour of the sentence. Pitch levels are quantified into pitch target labels and tone vectors are clustered into some classes in this work.

3.2 Labelling of Pitch Level

Labelling of pitch level is to convert the continuous values into discrete values. This captures the global pitch change of an utterance. The discrete labels are easy to predict in our model. The labelling of pitch levels in this work is a quantization of the pitch mean of each syllable. Suppose \( m \) is the number of levels to label, \( f_{\text{max}} \) and \( f_{\text{min}} \) are maximal and minimal pitch level values in the sentence respectively. The labelling is done as following:

\[
L_i = \left\lfloor \frac{f_i - f_{\text{min}}}{(f_{\text{max}} - f_{\text{min}})(m - 1)} + 0.5 \right\rfloor
\]

Using this scheme, all pitch mean of syllable are converted into discrete labels ranging from 0 to \( m \). The labels will be predicted to represent the intonation pitch contour.
3.3 Prediction of Pitch Label

The probabilistic approach to prediction of prosodic labels uses a stochastic model $P(a_i^n | Y_i^n)$ that represents the conditional dependence of the sequence of the labels the sequence of feature vectors $Y_i^n = \{Y_1, Y_2, ..., Y_n\}$. $a_i$ is the label of the syllable $i$, and $Y_i$ is a vector of features that are relevant to the label. Using the chain rule [3]:

$$P(a_i^n | Y_i^n) = p(a_i | Y_i^n) \prod_{i=2}^n p(a_i | a_{i-1}, Y_i^n)$$

Under the first-order Markov assumption, it is assumed that $a_i$ is only dependent on $a_{i-1}$ and $Y_i$, which gives:

$$P(a_i^n | Y_i^n) = p(a_i | Y_i^n) \prod_{i=2}^n p(a_i | a_{i-1}, Y_i)$$

To calculate the $p(a_i | a_{i-1}, Y_i)$, CART approach [5] is applied. $a_{i-1}$ and $Y_i$ are used as input features of the tree and $p(a_i | a_{i-1}, Y_i)$ is the output value of the tree. Normally when using a decision tree, terminal nodes assign the most likely classification. In this research, each node is associated with a discrete distribution, which represents the conditional probabilities for each label type. That is, we have the values of $p(a_i = l | a_{i-1}, Y_i)$ from the tree, where $l$ is a label type.

The feature vector $Y_i$ includes 21 features, which consist of the following information:

- Syllable information: Initial type, final type and tone of the syllable.
- Information of surrounding syllables: initial type, final type and tone of the previous and next syllable.
- Word information: the length of the word, the location of the syllable in the word, part-of-speech of the word.
- Information of surrounding words: length of the previous word and next word, part-of-speech of previous word and next word.
- Position in prosodic word: Prosodic word means the smallest prosodic trunk in speech. The features of prosodic word include length of the word, location of the syllable in the prosodic word.
- Position in intonational phrase. The relative location of the syllable in the intonational phrase. We use 3 values to indicate the first third, second third and final third of the intonational phrase.

As the label $a_i$ is dependent on $a_{i-1}$, determination of $a_i$ is in fact a dynamic process. In this research, Viterbi search approach is used to find the best assignment of labels to syllable sequence. After the labels are determined, each label is converted into a continuous value, which is the average value of the pitch label.

3.4 Clustering of Syllable Tone Contour

We use clustering approach to classify each vector in the corpus into a class. Using this way, all the vectors are converted into a definite number of pitch contour types. Prediction is to specify a tone contour class for a syllable.

For all the TVs in the corpus, the clustering is done using K-mean clustering algorithm. The desired number of cluster we choose is 16. As there are 5 tones in Chinese, using 16 types allows an average of 3 to 4 variants for each tone.

3.5 Prediction of Syllable Tone Contour

The task of tone contour vector prediction is to find out which type of pitch contour is to be selected as the tone pitch contour of a syllable. In this research, the prediction is done using CART approach. The features for the prediction include all the features used in the prediction of the pitch level labels plus one more feature, the pitch label of the syllable.

3.6 Reconstruction of Pitch Contour

The final pitch contour is a combination of the intonation contour and tone contour. The tone contour is first reconstructed using interpolation over the duration of voiced part of the syllable. Then the pitch level is added to each tone contour to get the final pitch contour of the whole utterance.

4. EXPERIMENT

The performance of the approach can be observed by looking at the accuracy of prediction of pitch labels, prediction of tone contour classes, and the error of final pitch contour values. We use 75% of our data as training data, and hold the rest 25% data as testing data.

4.1 Pitch Level Prediction

Intonation contour is expressed by the means of pitch values of a sequence of syllables. The pitch values of the training data are first converted into a sequence of labels. In this research, 5 levels of labels are defined. The following table shows the result of the prediction. Rate1 means the rate of correctly predicted label. From the table we can see that the best value for the exact prediction of labels is 71.5% for label type 5. And the lowest value is only 22.3%. For the labels are not correctly predicted, most of the labels fall into a nearest label group. So we calculated the accuracy considering the second best choice of labels. Rate2 means the correct rate if considering the closest neighbouring labels as correct. We can see, if considering the second best choice as correct labels, the accuracy can be as high as 93.8%. The average
accuracy is 90.0%. This shows the effectiveness of the approach.

<table>
<thead>
<tr>
<th>Actual Label</th>
<th>Predicted Label</th>
<th>Rate1</th>
<th>Rate2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1843 707 62 16</td>
<td>92</td>
<td>67.8% 93.8%</td>
</tr>
<tr>
<td>2</td>
<td>615 1871 398 48</td>
<td>90</td>
<td>61.3% 95.5%</td>
</tr>
<tr>
<td>3</td>
<td>92 656 763 146 205</td>
<td>41.0%</td>
<td>81.1%</td>
</tr>
<tr>
<td>4</td>
<td>31 161 334 270 416</td>
<td>22.3%</td>
<td>84.2%</td>
</tr>
<tr>
<td>5</td>
<td>64 81 148 152 1118</td>
<td>71.5%</td>
<td>81.3%</td>
</tr>
</tbody>
</table>

Table 1. Confusion matrix of label prediction

4.2 Tone Contour Class Prediction

We defined 16 tone classes in this research. Each tone vector has a dimension of 9. We cluster the tone vectors in the corpus using K-Mean clustering algorithm, and use the Euclid distance measure. After 30 rounds of clustering, we get 16 vectors to indicate the centre of each class. The following figure shows the clustering result. From the figure, we can see the diversity of tone contour. For example, rising curves are curves for tone 2 and falling curves are for tone 4.

<table>
<thead>
<tr>
<th>Num of closest clusters</th>
<th>Correct Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35.4%</td>
</tr>
<tr>
<td>2</td>
<td>51.7%</td>
</tr>
<tr>
<td>3</td>
<td>64.6%</td>
</tr>
<tr>
<td>4</td>
<td>77.2%</td>
</tr>
<tr>
<td>5</td>
<td>83.7%</td>
</tr>
<tr>
<td>6</td>
<td>90.0%</td>
</tr>
</tbody>
</table>

Table 2. Accuracy of tone contour prediction

4.3 Pitch Contour

The accuracy of reconstructed pitch contour is calculated by comparing the predicted pitch values with the actual pitch values of sampled pitch point. We achieved an average error of 16.3 Hz, a correlation of 0.78, and an RMSE of 22.4 Hz.

5. CONCLUSION

The prosody model decomposes the pitch contour into two parts, the global intonation of sentence and the local tone contours of syllables. The pitch level of syllable is converted into discrete labels in the prediction of global intonation contour. The tone pitch contour is clustered into classes in this research. The two parts are separately predicted and then combined together. CART approach and Markov assumption is used in this research to build a probabilistic model. The model achieves good performance for pitch contour generation.

6. REFERENCES