A Stochastic Polynomial Tone Model for Continuous Mandarin Speech

Yang CAO, Taiyi HUANG, Bo XU, Chengrong LI

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
P.O Box 2728, Beijing 100080, P.R China
{cy, huang, xubo, lich}@nlpr.ia.ac.cn

Abstract

In this paper, a stochastic polynomial tone model is presented for tone modeling in continuous mandarin speech. This model, the pitch contour is described by a stochastic trajectory. The mean trajectory is represented by a polynomial function of normalized time while the variance is time varying. After that, an effective training and recognition algorithm is developed respectively. Also the problem of missing observation is discussed. Decision tree is employed to cluster the tone pattern variations, which are represented by proposed model. Many possible factors other than tone of neighboring syllables were taken into consideration when the decision tree was constructed. The experiments result shows that the tone recognition speed can increase more than 10 times while the recognition error rates decreased by 16% compared with traditional HMM tone model.

1. Introduction

Mandarin is a tonal language, each character is pronounced as a monosyllable with a tone association. Tonal information is very important since the same syllable with different tones has distinct meaning. Therefore tone classification is an indispensable part for mandarin Chinese speech recognition. Basically, there are only five lexical tones in Chinese isolated syllables and the tone of a monosyllable is mainly characterized by the shape of its fundamental frequency (F0) contour. Tone Classification in isolated syllables is an easy task, since tone pattern is stable and simple under this condition. However, in the case of continuous mandarin speech, tone classification is a very hard task because under this condition tone patterns are subject to various modifications. There are two problems associated with tone classification in continuous speech: one is how to build reasonable mathematical model for tone pattern; the other is how to characterize tone pattern variations in continuous speech in the model.

When study the F0 contour in continuous mandarin speech, we find that three-order polynomial is precise enough to characterize the F0 contour. Considering the stochastic property of speech signal and the variations among different speakers, we utilize statistical method to model the tone pattern. In this model, F0 contour is described as a stochastic trajectory. The mean trajectory is parameterized by a polynomial function of normalized time. A time-warping mapping is applied to associate each variance with one of the variance regions. An efficient algorithm is developed to estimate the model parameters. We also discuss the recognition algorithm. Unlike the traditional tone recognition model (HMM or NN) in mandarin speech recognition, our model can handle the problem of missing observations naturally, which is failing to extract F0 value at voiced segment. Another advantage of this model is that it can be used as a pitch curve generation model, so it is easy to adopt this model to speech synthesis.

Most current analysis of tone pattern variations derives from the qualitative observation result. This method is not appropriate for tone recognition and pitch curve generation. In order to overcome this shortcoming, we investigate the tone pattern variation in continuous mandarin speech by stochastic cluster method. Since decision tree is a data-driven method that is easy to incorporate expert knowledge, we chose it as our cluster method. While constructing the decision tree, besides neighboring tone, many other possible factors were considered such as syllable position in the word, Consonant/Vowel type of the syllable, which were not utilized in conventional analysis. After the tree was established, 28 different tone patterns and their corresponding model parameters were acquired. From the result we found that many factors other than tone of neighboring syllable affected tone pattern variation in continuous mandarin speech.

In section 2, we will discuss the stochastic polynomial tone model. The tone pattern clustering via decision tree is presented in section 3. Experimental results are presented in section 4 and the conclusion in section 5.

2. Stochastic Tone Model

2.1 Model definition

For tone pattern \( \alpha \), we assume its pitch sequence is \( F = \{ f_1, f_2, \ldots, f_L \} \), where \( f_i \) represents F0 value at time frame \( i \). In stochastic polynomial tone model, this sequence is represented by an R-th order trajectory model as follows:

\[
    f_i = \mu_i + \delta(i) = \sum_{l=0}^{R} b_{l} \varphi(i/l) + \varphi(i) \tag{1}
\]

\* The work described in the paper is funded by the National Key Fundamental Research Program (the 973 Program) under No.  G1999030504
\[ \varphi(i) \] is a Gaussian white noise with variance \( V_{\alpha,i} \). \( b \) means the coefficient of polynomial, and \( \mu_i \) represents mean at time frame \( i \). The key of this model is how to specify the variance \( V_{\alpha,i} \). In our approach, a deterministic monotone nondecreasing time-warping mapping \( T_\alpha : i \rightarrow m \) \( m \in \{1\cdots M\} \) is applied to associate each variance \( V_{\alpha,i} \) with one of the \( M \) fixed variance, in other words: \( V_{\alpha,i} = V_{\alpha,i_\alpha(i)} \).

So the likelihood of a pitch sequence \( F \), given that it is generated by a tone pattern \( \alpha \), can be expressed as:
\[
P(F | \alpha) = \prod_{i=0}^{L} p(f_i) = \prod_{i=0}^{L} N(f_i - \mu_i, V_{\alpha,i_\alpha(i)})
\]

Where \( N(\cdot) \) is the normal distribution function.

### 2.2 Training algorithm

Assuming there are \( N \) pitch sequences \( F_1,F_2\cdots F_N \) in the training data for tone pattern \( \alpha \), the model parameters \( \theta_\alpha = \{B_\alpha, V_\alpha \} \) can be estimated via the Maximum-likelihood rules. That is, we wish to find model parameter:
\[
\theta_\alpha^* = \arg \max_\theta \prod_{i=1}^{N} P(F_i, F_{i+1} \cdots F_N | \theta)
\]

(2)

Expand equation (3), we can get:
\[
\sum_{m=1}^{M} \sum_{i=1}^{L} \sum_{j=1}^{N} b_{\alpha,m,i} \delta_{m,T_{\alpha,i}}(t/L_i)^{m-1} = \sum_{m=1}^{M} \sum_{i=1}^{L} \sum_{j=1}^{N} f_{ij} \delta_{m,T_{\alpha,i}}(t/L_i)^{m-1}
\]

(3)

\[
\sum_{m=1}^{M} \sum_{i=1}^{L} \sum_{j=1}^{N} f_{ij} \delta_{m,T_{\alpha,i}}(t/L_i)^{m-1} = k = 0\cdots R
\]

(4)

and
\[
V_{\alpha,m} = \left\{ \sum_{i=1}^{L} \sum_{j=1}^{N} \delta_{m,T_{\alpha,i}}(f_{ij} - \sum_{r=0}^{K} b_{\alpha,r}(i/L_i)^r)^2 \right\} / \left( \sum_{i=1}^{L} \sum_{j=1}^{N} \delta_{m,T_{\alpha,i}} \right)
\]

(5)

\( \delta \) represents Kronecker delta.

By solving system of equations (4) and (5), we can obtain the required parameter \( B_\alpha \) and \( V_\alpha \). However (4) and (5) are nonlinear equations, it is very difficult to find analytical expression. So we develop the following iteration algorithm to estimate the parameters efficiently:

1. Let \( V_{\alpha,m}^{(0)} = V_{\alpha,1} \), \( m = 2 \cdots M \), \( i = 1 \);
2. By replacing \( V_{\alpha,m} \) in equation (4) with \( V_{\alpha,m}^{(i-1)} \), we get \( b_{\alpha,m}^{(i)} \);
3. Replace \( b_{\alpha,m} \) in equation (5) with \( b_{\alpha,m}^{(i)} \), get \( V_{\alpha,m}^{(i)} \);
4. Calculate probability of training data with respect to \( b_{\alpha,m}^{(i)} \), \( V_{\alpha,m}^{(i)} \) and denotes it as \( P^{(i)} \);
5. If \( \log(P^{(i)}) - \log(P^{(i-1)}) < \varepsilon \), go to step 7;
6. \( i = i + 1 \), go to step 2;
7. \( b_{\alpha,m} = b_{\alpha,m}^{(i)} \), \( V_{\alpha,m} = V_{\alpha,m}^{(i)} \), \( r = 0 \cdots R, m = 1 \cdots M \), Termination.

### 2.3 Recognition algorithm

Suppose there are total \( k \) tone patterns \( C_j \), \( i=1\cdots K \), each pattern is characterized by parameter sets \( \{B_j, V_j \} \). If an observing pitch sequence \( F = \{f_0, f_1\cdots f_L \} \) is given, according to Bayesian decision rules, the classifying result can be written as:
\[ C(F) = C_j \text{ if } P(F | C_j)P(C_j) = \max_j P(F | C_j)P(C_j) \]

Then we can define:
\[ g(F | C_j) = \log P(C_j) + \log P(F | C_j) \]

By omitting the constant for all patterns, \( g \) can be rewritten as:
\[ g(F | C_j) = 2 \log P(C_j) - \sum_{m=1}^{M} \sum_{i=0}^{L} \delta_{m,T_{\alpha,i}} \left[ \log(V_{\alpha,m}) + \left( f_i - \sum_{r=0}^{K} b_{\alpha,r}(i/L_i)^r \right)^2 / V_{\alpha,m} \right] \]

(6)

The decision rule now becomes:
\[ C(F) = C_j \text{ if } g(F | C_j) = \max_j g(F | C_j) \]
2.4 Case of missing observation

In continuous Mandarin speech, there are some weak voiced parts that often appear to be unvoiced, such as the center part of tone 3. So there often exist some unexpected zero points in the estimation sequences of pitch. In practical speech recognition system, this problem may be more serious. As the recognition result cannot be completely error free, some unvoiced parts may be taken as voiced. In traditional HMM or NN tone model, due to constraint of the model, the pitch sequence must be continuous. For HMM or NN system, this problem is usually tackled by extracting F0 with lower threshold followed by smoothing process, which might result in wrong F0 contour. While in stochastic tone model, missing observation is tolerant.

Assuming \( f_i = 0 \) in the pitch sequence \( F \), since it provides no information, we can define \( p(f_i | \alpha) = 1 \) for any tone pattern \( \alpha \). The formula (6) can be changed to:

\[
g(F | C_i) = 2 \log(P(C_i)) - \sum_{m=1}^{N_i} \sum_{i=1}^{l_i} U(f_i) \frac{s_{m,T_i(i)}}{\log(V_{i,m})} \left[ f_i - \sum_{r=0}^{K} b_{i,r}(i/L)^r \right]^2 / V_{i,m}
\]

Where \( U(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \)

The training algorithm is fixed correspondingly.

So in our model, only reliable F0 is utilized in training and recognition process, thus the affect of missing data is eliminated intrinsically, which ensure the robustness and precise of the model.

3. Tone Pattern Clustering

For each tone, we construct a decision tree that specifies its acoustic environment. After the tree is established, from the leaf node, we can obtain the result of tone pattern variations of continuous speech. The following four aspects must be considered before the tree construction: input data, the question set, evaluation function and stopping criteria.

3.1 Input Data

In our approach, the features are normalized pitch sequences. Then the input data for the decision tree are the features of every sample, and its answer to all questions.

3.2 The question set

We take many possible factors into consideration and designed the questions about the following items (details can be found in [1]):

1. Tone of neighboring syllables, 12 questions.
2. Syllable position in the word, 4 questions.
3. Consonant/Vowel type of the syllable, 6 questions about Consonant type, 14 about Vowel type, total 20 questions.

3.3 Evaluation function

Let \( F_n \) represents the set of labeled data associated with node \( n \), \( Y_n \) represents the stochastic polynomial model training from \( F_n \).

Let \( P(F_n | Y_n) = \sum_{i=1}^{N_n} s(F_n) Y_n \), (here the priori probability is eliminated), \( P(F_n | Y_n) \) denotes the logarithm probability assigned to \( F_n \). So \( P(F_n | Y_n) \) is a measure of how well the model \( F_n \) fits the data at node \( n \); if the data in \( Y_n \) are similar to each other, \( P(F_n | Y_n) \) will be large. Now if a question \( q \) splits the data at node \( n \) into two subsets \( F_l \) and \( F_r \), \( Y_l \) and \( Y_r \) are the corresponding stochastic polynomial models, the outcome of this split is:

\[
m(n, q) = P(F_l | Y_l) + P(F_r | Y_r) - P(F_{n-1} | Y_{n-1})
\]

Since \( m(n, q) \) is a measure of the improvement in purity as a result of this split by question \( q \), it is a very good choice for evaluation function.

3.4 Stopping criteria

Stopping criteria is very simple in our system, if the outcome of the best question split at a node \( n \) is less than threshold, or number of the data at a node falls below threshold, we stop the split of this node and designate it as a leaf node. The thresholds are selected empirically.

Based on the above specification, the tree is constructed. In the tree, each leaf node corresponds to one tone pattern, questions associated with it specify the context of this pattern. The corresponding model parameters are acquired in building process.

4. Experiment

4.1 Conditions

Two databases were used in the follow experiments. One was for the training, the other for the testing. The training database was based on the Chinese continuous speech database of national “863” project. There are total 8,740 utterances consisting of 110,696 syllables uttered by 16 male speakers in our training database. Each sentence was divided into Consonant/Vowel segments by viterbi alignment using 138 models of vowels and
consonants [4]. In our approach, a cepstrum pitch determination algorithm was applied for pitch detection. Then the pitch contour of each sentential utterance was normalized by averaged pitch value of the current speaker to reduce interspeaker variability. The normalization for testing was from “b63” dictation testing corpus which consist of 240 utterance, totally 3,145 syllables uttered by 6 male speaker.

In the following experiments, the order of stochastic polynomial model is set to 3 and the number of variance 4. The time-warping transform $T_L$ is set as:

$$T_L(i) = \begin{cases} 
1 & i=0,1 \\
2 & i=2\cdot L/2 \\
3 & i=L/2 \ldots L-1 \\
4 & i=L-1, L 
\end{cases}$$

4.2 Results of the decision tree

After the decision tree was constructed, total 28 tone patterns were got. There are 4 variation patterns for tone 1, 6 variation patterns for tone 2, 9 for tone 3, 4 for tone 4 and 5 for tone 5. After thorough analysis of the result, we found many tone variations, which were consistent with research result from linguist. For example, the well-known Chinese Sandhi rules, when Tone 3 precedes another Tone 3, it will be pronounced approximately as Tone 2.

By analyzing the questions associated with leaf nodes, we find that 60% questions comes from tone of neighbor tone. 20% from Consonant/Vowel type of syllable (almost all questions are about unvoiced/voiced initial problem), 20% from syllable position in the word. The result is consistent with our previous work[1].

4.3 Tone recognition experiments

In order to test how exactly the tone model matches tone pattern in continuous speech, we applied tone recognition as follows: first divided the whole sentence into Consonant/Vowel segments using viterbi alignment, then performed a full search of all models in each voiced segments with a tone association. We use three sets of models to do the comparing experiments: one is the HMM tone models composed of 29 models generated by decision tree [1](model set 1), the second is the context-independent stochastic polynomial model (model set 2), the last one composed 28 stochastic polynomial model generated by the decision-tree (model set 3). The result is shown in Table 1.

<table>
<thead>
<tr>
<th>Model set</th>
<th>Tone1</th>
<th>Tone2</th>
<th>Tone3</th>
<th>Tone4</th>
<th>Tone5</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74.3%</td>
<td>68.3%</td>
<td>64.9%</td>
<td>72.2%</td>
<td>63.5%</td>
<td>70.1%</td>
</tr>
<tr>
<td>2</td>
<td>72.9%</td>
<td>60.2%</td>
<td>58.3%</td>
<td>68.5%</td>
<td>45.3%</td>
<td>65.0%</td>
</tr>
<tr>
<td>3</td>
<td>80.1%</td>
<td>71.6%</td>
<td>68.0%</td>
<td>76.7%</td>
<td>67.7%</td>
<td>74.9%</td>
</tr>
</tbody>
</table>

It is obvious that the recognition accuracy is increased significantly when considering context information. This shows that the decision tree is effective. However, the recognition accuracy of tone 3 and tone 5 are still below the average level, this may due to the relatively high variabilities of F0 contour for both tone 3 and 5. Experiments also show that the recognition error rate of our new model is decreased by 16% averagely comparing with the HMM tone model.

The classification error may be related to the following factors:

1. The energy and duration information are not incorporated in the current stochastic polynomial tone model
2. In training process, the normalization base is evaluated by the averaged pitch value of every speaker. But in testing process, we cannot know this value during the recognition process, so each sentence is just normalized by its own means. This will introduce some additional errors.
3. In training and recognizing process, sentence is divided into syllable segments by viterbi alignment, which will introduce some additional errors.

The recognition speed is also examined: the time cost of our new model is about 8s, while HMM model costs about 107s.

5. Conclusions

A stochastic polynomial tone model is proposed in this article. Efficient training and recognition algorithm is developed. Unlike the traditional HMM or NN tone models, this model can handle the problem of missing observations naturally. Decision tree based clustering method is employed to capture tone pattern variations in continuous speech. The experiment result shows that the tone recognition speed can increase more than 10 times while the recognition error rates decreased by 16% compared with the HMM tone model.

In future work, we will try to improve tone model accuracy by incorporating more information, such as energy and duration information and apply this model for tone pattern generation in speech synthesis system.

6. REFERENCES