ABSTRACT

In this paper we introduce a classification and regression tree (CART) based method to improve the efficiency of our corpus-based Mandarin Chinese synthesis system and at the same time maintain the quality of the synthesized speech. CART is one kind of the popular decision tree, through which, the candidates of the same tonal syllable in the corpus are pre-classified in three different ways, taking into account their experiential rule distance, segmental features or prosodic features separately. The difference of these methods exists in the different measurement of the distance between any two candidates, while the distance is used to construct the decision tree. The implementation and comparison of these three kinds of unit pre-selection methods mentioned above and their results are presented. And finally we come to the conclusion that prosodic characteristics of syllables are more important than segmental characteristics in Mandarin Chinese synthesis.

1. INTRODUCTION

In our corpus based Mandarin Chinese synthesis system with tonal syllable as the synthesis unit [1][2], for each target tonal syllable, the object costs of every candidate in the corpus are calculated first, then some of the best candidates are selected. By calculating the link costs of the joint syllables, a Viterbi algorithm is applied to search the optimal path among the selected candidates and exporting the most appropriate candidate unit. The object cost mentioned above includes two parts: experiential rule distance and pitch distance. The details are shown in Fig.1.

With the increase of the size of corpus for synthesis, the efficiency of the synthesis system will be greatly decreased through calculating the object cost between target units and all of the candidates one by one, especially for such target syllable with thousands of candidates. So we pre-classify the candidates by a CART. The optimal class of the pre-classified candidates will first be decided according to the attributes of the target unit by the corresponding decision tree. Then processing with the candidates in the optimal class, such as calculation of their object costs and link costs, will carry on, and the candidates of other classes will be discarded. Because of the low resource cost of making a decision by CART and the distinct decrease of the number of the candidates for object cost calculation, the efficiency of the synthesis system will be greatly improved.

2. UNIT PRE-SELECTION BASED ON RULE DISTANCE

2.1 Details of the Object Cost Calculation

The object cost introduced in section 1 includes two parts: rule distance and pitch distance.

\[
ObjectCost = OC_{RULE} + OC_{PITCH} \tag{1}
\]

In Eq. (1), the rule distance \(OC_{RULE}\) integrates the difference of segmental features and prosodic features between the candidate and the target syllables, and evaluates the substitution ability of the candidate. It is defined as follows:
\[ OC\_RULE = \sum_{i=1}^{N} d_i(a_i, b_i) \]  

For calculating the rule distance between a target syllable and a candidate, the attribute values \((a_1, a_2, \ldots, a_N)\) of the target syllable are obtained beforehand through text analysis. These values include the syllable position information and the prosodic environment information. These attribute values are compared with the corresponding values \((b_1, b_2, \ldots, b_N)\) of the candidate. For the difference between \(a_i\) and \(b_i\), an experiential distance \(d_i\) is given. Integrating such distance on each attribute, we get the rule distance between the target and the candidate. Likewise, we can calculate the rule distance between any two candidates in the same way.

The pitch distance is the difference between the pitch contour of the candidate and the predicted pitch contour of the target syllable, which is a complement of the rule distance because there is no pitch modification when concatenating the optimal candidate to output in our system. The pitch prediction is calculated the rule distance on each attribute, we get the final object cost of each candidate syllable.

2.2 Introduction of CART

In recent years, all kinds of decision trees are widely used in speech synthesis for data mining from the corpus [3][6][7], and classification and regression tree (CART) is one of them. In fact, it is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes [4]. CART can be used for either classification or prediction, and here, we primarily make use of the classification ability of CART.

The application of CART includes two parts: training and deciding. The wagon program in the Edinburgh Speech Tools [5] is adopted to train the CART. For training, the input parameters include a distance matrix, which is composed of the distance between every two patterns for training, and an attribute list for test, the output is a constructed tree with similar patterns in the same leaf node. For deciding, the input parameters are the attribute values of the target syllable and the output is a decided class with similar patterns.

2.3 Unit Pre-Selection Based On Rule Distance

Our corpus is composed of 83267 syllables in 6204 sentences, covering all the 1441 tonal syllables in Mandarin Chinese, recorded by a professional female broadcaster. The total size of the corpus is about 800MB. Because the training of CART needs enough patterns and it is of no worth to apply unit pre-selection for the tonal syllable with sparse candidates, only the tonal syllables with more than 50 patterns are selected for CART construction. Here the number of proper tonal syllable is 461, including 65708 patterns in our corpus with a proportion of about 80%. The discussion about the coverage of the selected high-frequency tonal syllables during synthesis will be given in section 5.1.

For each tonal syllable, the rule distance between every two candidates are calculated at first to form the distance matrix according to (2), then a CART is trained by the following attributes of a syllable (the values in the parentheses are the possible values for each attribute):

- Syllable number in rhythm word (1,2,3,4,5,6,7)
- Position in rhythm word (1,2,3,4,5,6,7)
- Previous border category and next border category: (syllable, step, rhythm word, phase or sentence)
- Previous syllable tone and next syllable tone (1,2,3,4, light or none)
- Previous final category (mono, complex, nasal or none)
- Initial tonal category (unaspirated stop, aspirated stop, unaspirated affricate, aspirated affricate, nasal, liquid or none)

When synthesizing a sentence, for each target syllable with more than 50 candidates, a decision is made by the corresponding trained CART to firstly select the best-fitted class. Then the object cost of each candidate in the selected class is calculated following the way introduced in section 2.1. All the candidates in other class are discarded.

3. UNIT PRE-SELECTION BASED ON SEGMENTAL FEATURES

3.1 Brief Introduction

The rule distance Eq.(2) integrates the distance on segmental features and prosodic features between two syllables. But many experiential distances are needed in the calculation of Eq.(2). It is very difficult and seems impossible to ensure that all these distances are most proper, and any little change of the corpus composition or reading style may cause them to be modified manually. So it is necessary to find an unit pre-selection method with no dependency on such expert experience. We resolve this problem from the view of syllable’s segmental features or prosodic features. That is what we will discuss in section 3 and 4.

3.2 Selection of Proper Acoustic Parameter

In order to pre-classify the candidates by their segmental features, we should first find a proper acoustic parameter to measure the segmental difference between two syllables. The acoustic parameter can be MFCC, LSF, formant and so on [6][7][8]. By experiment, we find that choosing LSF for such difference measurement is more suitable with our Mandarin Chinese synthesis system.

3.3 Implementation

For the same reason, only the tonal syllables with more than 50 patterns are selected for CART construction. For each tonal syllable, the following steps are performed:

- Extract the LSF of every candidate frame by frame. The order of LSF is 16 and the frame shift is 10ms.
- Calculate the LSF distance between every two candidates by DTW to create the distance matrix. The original LSF parameters are first weighted according to order for the distance calculation.
- Train the CART with decision attributes.
The decision attributes used here are almost the same as those described in section 2.3, only added with some more detailed syllable position information such as the word position and word number in sentence. During synthesis, the candidates are pre-selected using the constructed decision trees according to the target syllables. It can be considered that each candidate in the decided class has similar acoustic characteristics with the target syllable, the distance between the candidate and the centre of the decided class can be treated as the object cost on segmental features. The pitch prediction in section 2.1 is used to calculate the object cost on prosodic features. The final object cost is given by adding the object cost on segmental features and prosodic features together. The candidates out of the decided class are also discarded. Fig.2 is the flow chart.

![Figure 2](image)

In Fig.2, it is shown that no experiential rules or distances are used during the whole object cost calculation procedure. So it is also a feasible approach for totally automatic speech synthesis with no any dependency on experiential rules.

### 4. UNIT PRE-SELECTION BASED ON PROSODIC FEATURES

#### 4.1 The Measurement of Prosodic Features

The syllable’s prosodic features include pitch, duration, energy and so on. Here, the pitch and duration are taken into account to measure the difference on prosodic features between two syllables, as Eq.(3).

\[
\text{Dist}(A, B) = \left( \sum_{i=1}^{N} (PA_i - PB_i)^2 \right)^{1/2} + \alpha \cdot |DA - DB|
\]  

(3)

In Eq.(3), PA and PB are pitch data of syllable A and syllable B, which are normalized to N points according to the their duration. DA and DB are duration of the two syllables. \(\alpha\) is a weight to adjust the proportion between pitch and duration. Because duration can be modified flexibly during concatenating, \(\alpha\) is commonly rather small to emphasize the difference on pitch values.

#### 4.2 Implementation

The procession of training decision tree here is in almost the same way discussed in section 3.3. The LSF distance in section 3.3 is replaced with the prosody distance described in Eq.(3).

![Figure 3](image)

Fig.3 is the flow chart of object cost calculation. The candidates in the pre-selected class which is decided by the trained CART may have similar prosodic features and meet the demand of target syllable. Likewise, the distance between the candidate and the centre of the decided class is calculated as the object cost on prosodic features. A rule-based distance is still necessary to evaluate the cost on segmental features of these candidates in the decided class, but the rule distance used here is much more simple than that used in section 2.1. Only the attributes that influence the segmental features strongly are taken into account, such as the syllable position in word, and other attributes having more significant influence on prosodic, such as tonal environment, are not considered.

### 5. EXPERIMENT AND CONCLUSION

#### 5.1 Coverage of High-Frequency Tonal Syllables

In above discussion, only the tonal syllables with more than 50 patterns are selected for CART construction. So, the effect of the application of unit pre-selection in practical speech synthesis system depends greatly on the coverage of these high-frequency syllables. A test is conducted to evaluate such coverage. A news text with 2091 syllables is synthesized using unit pre-selection for experiment. The number of syllables using the decision tree for unit pre-selection is 1833. The coverage is about 88%, a satisfactory result.

#### 5.2 The Efficiency Improvement By Unit Pre-Selection

Once the constructed decision tree is loaded into memory, it can be used repetitively. For each target syllable, the application of decision tree is in fact a procedure of searching path in the memory. So, the system resource cost is very low during unit
pre-selection compared with the object cost calculation of each candidate. In our Mandarin Chinese synthesis system, the synthesis speed can be improved by 3-4 times by applying unit pre-selection. Especially, the superiority is distinct when the size of the corpus grows to GB level.

5.3 The Quality of Synthesized Speech Using Unit Pre-Selection

In order to test the effect of unit pre-selection on the quality of synthesized speech, 10 sentences with 228 syllables are synthesized in the following 4 different ways to create 4 groups of test samples.

Group A: Original synthesis system without unit pre-selection

Group B: Synthesis system with unit pre-selection based on rule distance

Group C: Synthesis system with unit pre-selection based on segmental features

Group D: Synthesis system with unit pre-selection based on prosodic features

10-experienced listeners are invited to evaluate the quality of synthesized sentences in the four groups on a scale of 1(bad) to 5(good). The result is presented in Fig.4.

![Figure 4](image)

It is shown in Fig.4 that the quality of synthesized speech is maintained with little degradation. Comparing the three unit pre-selection method, the unit pre-selection based on prosodic features gives the best result that is very close to the original synthesis system and even better in some cases, while the unit pre-selection based on segmental features gives the lowest score. Because the selectivity on prosodic features is much limited after applying the unit pre-selection based on prosodic features, we can conclude that prosodic features are more important than segmental features in Mandarin Chinese synthesis and keeping the similarity on prosodic features is a rule we should follow.

6. CONCLUSION

In this paper, we describe an unit pre-selection method by CART for our Mandarin Chinese synthesis system to improve its efficiency. In practical test, the coverage of the syllables using unit pre-selection is more than 85% and the efficiency of our synthesis system is improved by 3-4 times. The quality of synthesized speech can be preserved with little degradation especially choosing prosodic features for CART construction. We also conclude that prosodic features are more important in Mandarin Chinese synthesis and some attempt on the fully automatic unit selection with no dependency on any experiential rules is taken.

7. ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of China (69975018), and China National Hi-Tech Project under N0.2001AA110333.

8. REFERENCE


