A New Method of Building Decision Tree Based On Target Information

Yi-Jian Wu  Yu Hu  Wu Xiaoru  Ren-Hua Wang

iFly Speech Laboratory
University of Science & Technology of China
jasonwu@mail.ustc.edu.cn

ABSTRACT

The algorithm of unit pre-selection based decision tree has been adopted in the corpus-based TTS system, but the effect of the decision tree, which is built by conventional methods, is not satisfactory. This paper proposes a new method of building a decision tree based on a mass of target information obtained in the synthesizing course. By synthesizing a large quantity of text and logging the information of target units, we used these target information as training data to classify the corpus unit and build the decision tree. The splitting criterion is to ensure that the best units of target units correctly classified, which guarantee that the synthesis quality would not be reduced. And the decision tree built by this method has an advantage that it has equal units among the different classes. By evaluating the synthesis quality in objective and subjective measure, this method is better than the conventional ones.

1. INTRODUCTION

The present waveform concatenation based TTS system [2], which depends on a large corpus, normally synthesizes text as follows. Firstly, input text is converted to a sequence of phonetic and high-level prosodic transcriptions. Then by these transcriptions the most matched units are selected from the corpus. Finally pitch and duration modification algorithm such as PSOLA, is applied to the final units before concatenating them. Yet they often suffer from significant quality decrease in timbre after modification. So in order to improve the effect of waveform concatenation based TTS system, the corpus should cover more phonetic and prosodic variations from which we can select a more suitable unit. As the corpus became larger, e.g. duration from thirty minutes to two hours or even more, the number of units for each target also become bigger, e.g. from dozens to hundreds even thousands. Although the effect is improved, there comes the problem of expensive computation of selecting appropriate unit from corpus, which would decrease the efficiency of synthesis.

There are two steps in unit selection. First, the target cost of every unit in the corpus is calculated, which is based on the target features (e.g. phonetic, prosodic), and some similar units are pre-selected. Second, the concatenation (or join) cost is calculated and the units are determined with a searching algorithm (e.g. viterbi). While the corpus is large, we found that

unit pre-selection from corpus is the most time consuming phase in unit selecting. So in order to improve the efficiency of unit selecting, we should search for an efficient method to pre-select unit. The present way is to adopt the decision tree to solve this problem.

2. CONVENTIONAL METHOD

Decision tree is widely applied in the pattern recognition to classify data for its efficiency. Its application in the speech synthesis includes the prediction of prosodic features and unit selection and so on. Here, we adopted the decision tree to pre-select unit from corpus.

There are two conventional methods to build the decision tree, which is employed in unit pre-selection [1][3]. The first one works as follows. It clusters the unit based on the measurement of distance between units, such as cepstrum, LPC, F0. Thus every unit in the corpus has its value of attributes and the index of class, which should be regarded as training data. Finally, the decision tree is built with the existing algorithm (e.g. C45). Another one is CART. Firstly, the root was constructed which contained all the units. Then from the root, the nodes are split up-to-down based on the specified attributes, until the condition of termination is reached. The splitting criterion is the greatest reduction of model cost, which is commonly defined as the weighted sum of square regression error of some features (e.g. phonetic, prosodic).

The decision trees built by these two methods have both been utilized in some speech synthesis systems, but the results are not satisfactory. The conventional methods are to train and build the tree by analyzing the parameter information of the existing units in corpus, which only ensured the accurate classification of units in the corpus. But because of the limitation of its size, the corpus cannot cover all phonetic and prosodic variations, and the decision effect of those non-covered units is not guaranteed. That is to say, those methods are not robust enough. Further more, there is no uniform unit in any two classes by the conventional decision tree. This does not accord with the fact, because the different phonetic and prosodic variation has not absolute border to each other, though the influence of different attribute is not the same. As far as the decision tree is concerned, it should be permitted that the uniform units appear in similar classes. Based on these ideas, we propose a new method of building the decision tree.
3. GENERATION OF DECISION TREE

The process of building a new decision tree is as follows: First we use the synthesis, in which unit pre-selection is based on calculating the target cost instead of decision tree, to synthesize a large quantity of text and log the target information (TI). Second, the logged TI is used as training data to classify the corpus unit and build the decision tree. Comparing to the decision tree built by conventional methods, it has the following advantages:

- Excellent robustness and coherence.
- The selected set of units should promise good quality synthesis.
- The different unit sets could have overlapping units.

3.1. Attributes selection

It is of significant importance to select the decision attributes for its aptness whether good or bad would immediately affect the decision effect. In our corpus-based Chinese synthesis with contextual-dependent unit selection [5], the syllable is the basic concatenative unit. Referring experiential knowledge, we’ve tried all kinds of attribute combination, and selected the following 8 attributes as the final decision attributes.

- Position in phrase (PIP): position of current syllable in its carrying prosodic phrase. It takes 7 values.
- Position in word (PIW): position of current syllable in its carrying prosodic word. It takes 4 values.
- Left (right) boundary (LB, RB): category of left (right) boundary type of current syllable. It takes 5 values.
- Left (right) tone (LT, RT): category of the tone of left-neighbored (right-neighbored) syllable. It takes 6 values.
- Left phonetic context (LC): category of the final of the left-neighbored syllable. It takes 4 values.
- Right phonetic context (RC): category of the initial of the right-neighbored syllable. It takes 4 values.

These attributes can be derived directly from text. Among them, PIP, PIW, LB, RB, LT and RT are factors mainly contributing to prosodic variations, and, LC and RC are factors reflecting the co-articulation effects between syllables. So all the combinations of them are representation of all possible phonetic and prosodic variation.

3.2. Preparation of training

To get plentiful TI (i.e. training data), which should cover as many phonetic and prosodic variations as possible, we synthesized a large quantity of text with the TTS system, which pre-selects unit by calculating target cost, and logging the TI. A text corpus of four-year People’s Daily, five-year GuangMing’s Daily and some other kinds of newspaper is selected, which contained about 150 million Chinese characters. The logged target information (TI) includes:

- Target attribute (TA): the values of the target unit of the afore-mentioned 8 attributes.
- Pre-selected units (PU): the units pre-selected from the corpus and the corresponding target costs.
- Best unit (BU): the final concatenative unit derived by the optimal searching algorithm.

Figure 1: (a) is the composition of TI (target information). (b) There may be same units in different TIC.

The structure is shown in Figure 1(a). In addition, The TI’s candidates (TIC) are defined as the candidate of the target, which includes pre-selected units and best unit. As the different target units could have the same pre-selected unit or best unit, there should be same units in different TIC, which shown in Figure 1(b). All the logged TI is regarded as the training data of building the decision tree. By analyzing these TI, we found that if the candidate number of the syllable in corpus is big, there will also be a great deal of the corresponding TI in TI corpus, and vice versa. So the training data of the high-frequency syllable is plentiful, and the generated decision tree is also accurate, which is just our expectation. At the same time, the training data of the low-frequency syllable is not enough to generate an accurate decision tree. Due to the shortage of the candidate in corpus, the pre-selection of the low-frequency syllable could adopt the algorithm of calculating target cost instead of the decision tree, which would hardly affect the efficiency of unit selection.

3.3. Classification and generation

As the selection of decision attributes and preparation of training data (i.e. TI) has been accomplished, the classification and generation of the decision started. The decision tree over here is binary tree.

First, all TI belongs to the root node. All the corpus units, adhering to the TIC, compose of a unit set. From root, the current node is split as illustrated in Figure 2 by traversal of all the attributes and the values, and the split score is calculated. According to the split score, a most ideal attribute and corresponding value is selected as the final split attribute. In the splitting course, all the TI in the node is divided into two child-nodes by their values of the final split attribute, and the TIC is also split accordingly, as a TIC belongs to a TI.

<table>
<thead>
<tr>
<th>TI</th>
<th>TIC1</th>
<th>TIC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PU1</td>
<td>PU2</td>
<td>BU1</td>
</tr>
<tr>
<td>u11, u12, ..., u1k</td>
<td>u21, u22, ..., u2k</td>
<td></td>
</tr>
</tbody>
</table>
Because the corpus units adhere to the TIC, the unit set is split to two sets via the split of the TI and TIC in the node. Comparing to the conventional methods, the corpus units are split by the attribute’s value of the target unit instead of the corpus unit. And in unit pre-selection by decision tree, the unit set is also selected by the value of the target unit. This shows that the course of building the decision tree in the new method accord with the course of decision-making. Furthermore, although the TIC in the two child-nodes of current node is unique, there should be duplicate units in the different TIC, which result in repetition in two unit sets. As the aforementioned, this classification accord with the actual case.

Figure 2: As current node is split to two nodes, the unit set is also split to two sets (unit set 1 and unit set 2). Especially, there may be same units in the different split sets.

When current node is split by certain attribute, the corresponding factors of calculating the split score include:

- Child score ($S_1'$, $S_2'$): the score due to the number and frequency of pre-selected unit and best unit in the child-node.
- Ratio of units ($R_1$, $R_2$): the ratio of number of units in two child-nodes, which include pre-selection units and best units.
- Square error of distance ($D_{1}''$, $D_{2}''$): the square error of spectrum distance between units in the child-node.
- Equal units ($N_e$): the number of equal units in the two different child-nodes.

The split score is calculated by equation (1):

$$S_{\text{split}} = w_1' \sum_{i=1}^{n_1} S_{1i}' + w_2' R_1 + w_3' R_2 + w_4' \sum_{i=1}^{n_2} D_{2i}'' + w_5' N_e$$  \hspace{1cm} (1)

$w_1'$, $w_2'$, $w_3'$, $w_4'$, $w_5'$ in equation (1) are weights for each factor.

According to $S_{\text{split}}$, the best attribute and value is confirmed and current node is split into two child-nodes by it. Then the child-nodes are split in the same way. The split operation would not finish until the condition of termination is reached. The ending condition of split is achieved when $S_{\text{split}}$ is less than certain threshold, or the child score $S''$ is less than a certain threshold, and the node becomes the leave of the tree.

After the split process has been accomplished, the operation of scissoring should be done to the unit sets. As every TIC has several units and the node may have much TIC, the units in the set (i.e. node) probably are excessive and miscellaneous and some units ought to be cut from the set. By calculating the score of every unit in the set, we delete the unit whose score is less than the threshold from the set. The corresponding factors of calculating the unit score include:

- Pre-selection frequency ($N_c$): the times that the unit regarded as the pre-selection unit.
- Sum of target costs ($N_c$: sum of the target costs in $N_c$ times, $S_t = \sum_{i=1}^{N_c} C_{iow}$).
- Best frequency ($N_b$): the times that the unit regarded as the best unit.
- Unit distance ($D_{uw}$): the distance between the unit and the core of unit set.

The unit score is calculated by equation (2):

$$S_{\text{unit}} = w_{ue} N_e + w_s S_t + w_{ub} N_b + w_d D_{uw}$$ \hspace{1cm} (2)

$w_{ue}$, $w_s$, $w_{ub}$, $w_d$ in equation (2) are weights for each factor. The weights can be adjusted by the different focus. In our decision tree, in order to reserve the unit that was regard as best unit or high-frequency pre-selection unit of TIC, which guarantee the decision effect, we set $w_{ub}$ as the most value and $w_{ue}$ as the secondary value.

The decision tree comes into being after the operation of split and prune of the node is finished. By analyzing the structure of the tree, we can see that the level of attribute is in exact accordance with the transcendent grade. The attribute that has more influence in variation is more close to the root, and vice verse. Furthermore, as the attributes in the low-level have less influence in variation, the border of two sets is more undistinguishable and the duplicate units in two neighbor-sets is more.

4. EVALUATION AND COMPARISON

Finally, we evaluated and compared the effect of three different TTS system which are as follows:

- NT-TTS: the TTS system which pre-selects units without decision tree.
- CT-TTS: the TTS system which pre-selects units by CART tree.
BT-TTS: the TTS system which pre-selects units by BTI (Based Target Information) tree.

The uniform testing texts, which were composed of 200 sentences and contained 3088 Chinese characters, were respectively synthesized by the three TTS system. In synthesizing process, we logged the final units of target units and save the output waveforms. Then the objective and subjective measure were adopted to evaluate and compare the effect of the different three TTS system (Figure 3).

![Fig 3](image.png)

Figure 3: The flow map of evaluation and comparison of three TTS system with objective and subjective measure.

### 4.1. Subjective measure

Each MOS (Mean Opinion Score) of the output waveforms are derived by 5 people who were trained before and have good coherence.

<table>
<thead>
<tr>
<th></th>
<th>NT-TTS</th>
<th>BT-TTS</th>
<th>CT-TTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>People 1</td>
<td>4.68</td>
<td>4.61</td>
<td>4.01</td>
</tr>
<tr>
<td>People 2</td>
<td>4.59</td>
<td>4.57</td>
<td>3.95</td>
</tr>
<tr>
<td>People 3</td>
<td>4.66</td>
<td>4.60</td>
<td>3.98</td>
</tr>
<tr>
<td>People 4</td>
<td>4.60</td>
<td>4.57</td>
<td>4.00</td>
</tr>
<tr>
<td>People 5</td>
<td>4.64</td>
<td>4.62</td>
<td>3.99</td>
</tr>
<tr>
<td>Mean</td>
<td>4.634</td>
<td>4.594</td>
<td>3.99</td>
</tr>
</tbody>
</table>

From the table, we can see that the perceptual effect of the speech synthesized by CT-TTS is much worse than NT-TTS. That is the reason why we propose this method of building decision tree. The effect of BT-TTS is greatly improved and close to the effect of NT-TTS.

### 4.2. Objective measure

Because keeping the quality of NT-TTS is the aim, we regard the final units of targets, which obtained by NT-TTS, as the criterion. Then we compared the difference of the final units between BT-TTS and NT-TTS, which is defined as BT-NT, and the difference between CT-TTS and NT-TTS, which is defined as CT-NT.

<table>
<thead>
<tr>
<th></th>
<th>BT-NT</th>
<th>CT-NT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal Num</td>
<td>2606</td>
<td>1007</td>
</tr>
<tr>
<td>Non Equal Num</td>
<td>482</td>
<td>2081</td>
</tr>
<tr>
<td>Equal Ratio</td>
<td>84.4%</td>
<td>32.6%</td>
</tr>
</tbody>
</table>

Seen from the table, there is much difference in final units between CT-TTS and NT-TTS and few differences between BT-TTS and NT-TTS. This shows that the final units of BT-TTS are more close to the one of NT-TTS. So the quality of BT-TTS is also more close to the one of NT-TTS.

As a result of the subjective and objective experiments, the synthesis quality of BT-TTS is much better than that of CT-TTS and almost close to that of NT-TTS. Furthermore, BT-TTS is more efficient than NT-TTS.

### 5. CONCLUSION AND FUTURE WORK

As compared with conventional methods, this technique of building decision tree we introduced has two relative advantages. Firstly the decision tree is more robust because of its great deal of training data, which cover more phonetic and prosodic information. Secondly, as a matter of fact, there should be duplicated units in two classes. The speech quality by adopting this decision tree is improved greatly than CART tree, and is tolerably comparative to the effect without decision tree. From the generation process of the decision tree, it can be found that not all of units are classified and some units are never present in training data. Our future work is to analyze those units, and to simplify the corpus based on the analysis result. Further more, the efficiency of algorithm is considerable in previous system without decision tree because the target cost is calculated in real time of synthesizing. But in this system the cost is calculated off-line at the time of training data generation, and therefore the efficiency is not much important. More detailed and complicated model can be adopted, and more accurate algorithm can be obtained.

### 6. ACKNOWLEDGEMENT

This research is supported by the National Natural Science Foundation of China (69975018).

### 7. REFERENCES


