A Hierarchic Processing Model In Chinese TTS

Bo Yin  Ren-Hua Wang
Department of Electronic Engineering & Information Science
University of Science & Technology of China
P.O.Box 4, Hefei, 203027
Email: rhw@ustc.edu.cn
Tel: 0551—3603645

Abstract

This paper puts forward a kind of hierarchy text processing model aimed at Chinese TTS system, and defines corresponding hierarchy labeling system. The actual realization on the hierarchic processing is also given in detail, and the processing tactics on the sub-phrase layer is specially discussed.

1. Introduction

In TTS system how to make the output speech with high naturalness is always the most crucial problem. In one hand, it lies on how to generate fluent natural voice from appropriate prosody knowledge and right algorithms; in the other hand, our understanding of target language, including the information of the interior syntax structure and the prosody layer structure of special text in concrete environment and etc, directly determines whether the synthesized voice has intelligibility and naturalness as human voice does.

Along with the advancement of speech synthesis technology, we get more and more prosody knowledge, which can make voice more expressive. Therefore, we realize the comprehension of the fine structure of text’s syntax and semantics is of great consequence for the promotion of synthetic effect.

Because of the deficiency of the adopted synthesis algorithm’s expressive force, the most Chinese TTS systems currently adopt simpler text processing, e.g. generate phonetic letters first by word segmentation, then convert them to speech waveforms, and a few systems give the boundaries of breath groups. That is, in congener Chinese TTS system, there is no similar front text process.

One of our efforts is to put forward a kind of accurate and feasible text-analysis strategy, which can correctly analyze and describe the layer’s information for each language and mapped them into prosodic layer, establish an effective and general model, and then to provide more valuable information to speech synthesizer.

Because it is related to the natural language processing and analysis, some knowledge and methods on computational linguistics are inevitably adopted. But, different from the usual natural language understanding, our work doesn’t concern about the sentence’s concrete meaning. In fact, how to express the sentence on the prosodic layer is concerned most, since all the analysis on the philological layer work for getting the expressive information on prosodic layer. Because of this, a lot of statistic methods and rules are adopted in the process, including machine-learning and rules-table-generating.

2. The establishment of the model

2.1. The idea of the model

2.1.1. The idea of hierarchy process

“Chinese is a pile of concepts”. In fact, not only in Chinese, but also in other natural language, it is always true that the small language units are assembled into bigger units to express specific meaning. However this constitutional relationship based on layers is more obvious in Chinese. So we define some distinctive layers and adopted them as different standards to process text. With this idea, all layer’s information can be organized in a form of parsing tree, in which higher level’s information is obtained from ones of lower levels. Then the closely linked degree between the syllables in a context can be obtained from the sentence’s parsing tree.
Compared with other methods, our method can describe natural language in more details. Additionally, it is easier to process and expand with the method.

2.1.2. The idea of modularized processing with step by step

Hierarchic processing directly results in the idea of modularized processing with step by step. Since the range of each layer is defined in advance, we can adopt different algorithm on different layer, then process each layer in sequence according to the dependent relationship of neighboring layers. Different module has been developed for the processing of different layer. These modules exist in the form of individual binary file, which are loaded and run by host program. The data transfer between the neighboring layers is the tree to be processed.

The direct advantage of this scheme is that, the whole system is easier to be clipped and maintained. We can choose appropriate algorithm module in different application occasions, and achieve both good performance and high efficiency. Meanwhile, the different module can be developed by different person, then both the maintenance and the performance contrast test of different algorithms are easier to test. Above all, the period of changing research state to application state will be shortened significantly.

2.1.3. The combination of statistic methods and rules

Because of the uncertainty in natural language, strictly defined rules can not be adopted. Especially now people haven’t enough research fruits about natural language, statistic method is a good way to try. It not only can be directly used in algorithm, but also be an important source of most rules.

When these statistic methods and rules are applied, the separation of data and algorithm is emphasized. All the rules are described by exterior rules-table. That is crucial for both system stability and the condition that different analyzing methods must be used on different text materials.

2.1.4. The introduction of uncertainty degree

Because of the limitation of natural language processing, for a concrete environment, an accurate solution is hard to beachieved. Meanwhile, the adopted method’s limitation adds to the imprecision. And, when some problem is uncertain, another problem then is affirmative. To describe this occasion, uncertainty degree is introduced. It is decided by the cost in rules-table, the coefficient gotten by statistics, and some target text’s character.

2.2. Hierarchy labeling system

To realize hierarchy processing, firstly we need a labeling system to distinguish each layer of text prosodic structure. But the existent labeling systems are all built for either philological purpose or prosodic one. Due to this reason, we put forward our own hierarchy labeling system with reference to TOBI labeling system and some others’ Chinese prosody labeling method, which gives attention to both of philological and prosodic application.

The establishment of this labeling system makes hierarchy processing to be feasible. And it is of great significance in labeling text, sharing resources and transmitting text with prosodic information.

We define the following layers:

L0: syllable layer, In Chinese, it represents a Chinese character.
L1: speech foot layer, it is well known as the layer of rhythm.
L2: sub—phrase layer, syntax words combine with mono-syllable word, and appear as one pronouncing unit.
L3: master—phase layer, it is relatively independent syntax unit.
L4: Breath—group layer, it happens to take a breath in a long sentence.
L5: Sentence layer, it is the top layer of all the concerned layers.

For example, the segmentations of sentence“我们的最终目标是得到高自然的语音” are as follows:

L0: 我们的最终目标是得到高自然的语音
L1: 我们的最终目标是得到高自然的语音
L2: 我们的最终目标是得到高自然的语音
L3: 我们的最终目标是得到高自然的语音
L4: 我们的最终目标是得到高自然的语音

Because this sentence is simple, the segmentations on L3 and L4 are same. In fact, the
segmentation on neighboring layer can be identical. Then some layers can be merged on some occasion.

In the front process, there may be one more layer called lexicon layer, which is the result of lexicon-matching. We shield this layer to the back process, then the back has no relation with the lexicon. It can be called L1.5. Otherwise, L6, macro sentence layer, is added for the convenience.

3. Process flow and realization

After the front text-analysis, according to the information on the parsing tree, the back prosodic generation/speech synthesis module can synthesize out the final speech now.

According to the idea of hierarchy process model, we design the KD2000 Chinese TTS system[1] and put it into practice. The results show that this system is a victory and highly effective. The good performance comes up with the hierarchy process model in setting up systems.

The application systems based on KD2000 have been put into practice in a lot of fields such as telecom, civil aviation and personal user, etc.

The following is a practical hierarchy process flow:

Fig 1 shows the hierarchy structure in the sentence “司马光还在21世纪的第三天。” Each node has the following information: layer, character, pronunciation, word type, rhythm, word frequency, etc. The hierarchy structure describes the connection tightness of neighboring nodes.

Three main steps in the whole process flow are introduced at follows.

3.1. Text process in advance

The text process in advance is comprised by sentence segmentation, character set transform, process on special symbol, process on text mark.

The process on special symbols is that, in TTS, how to give the correct prosodic information for the symbols appear in the text but not the Chinese characters. This process is an important part in the whole TTS system’s front process. If it doesn’t work well, the synthesis system’s tone is not good, and the synthesized speech has bad intelligibility.

The process on text-labeling is for explaining the control tag in the text, such as pronunciation, speaking speed, etc.

3.2. Automatic segmentation

Segmentation is the basis of the further analysis. Whether it is correct or not affects the final analyzing results.

The main lexicon for segmentation here has about 57000 items, including all kinds of prosodic and syntax information.

Because of being lack of ability to understand the meaning of the input text, the basis of sentence segmentation strategy at present is still dependent on the syntax lexicon with attributes of the classified words. It is possible to select the best segmentation path from all possible segmentation paths at the help of appropriate evaluating functions and Full-Expended-Words-Segmentation-Net we set up. For not affected by the fake words, the evaluating functions give attention to both Segmented Word-Frequency(SWF) and Form Word-Frequency (FWF)[1].

By this method, we segment one month’s text in China Daily for test. The correctness ratio is 96.2%.

In addition, we have special process for the unregistered words, polyphones, word type adjustment and etc. These all get good effects.

3.3. The generation of the layers

Now that the information of hierarchy is especially useful to the synthesis speech, how to generate correct layers and the descriptive information of each layer becomes one of our main research subjects. Because no idea can be mentioned before
about adopting information of several levels to improving output speech, we are fishing for the better way gradually.

As the junction of philological level and speech level, Sub—phrase layer carries a lot of syntax and prosodic information. In the view of prosodist, L2 is the reflection of prosody word, and it has great significance to the naturalness and intelligibility of speech. In the opinion of philologist, L2 is mainly determined by philological knowledge. So we adopt describing rules and corresponding costs to generate the optimizing path for L2. The rules are originated from both of mature syntax rules and statistical information. Because of the different contribution of each rule a cost coefficient are set up to embody the factor.

The rules-table describes dualistic information and corresponding costs. We can merge the L1/L2 nodes according to the rules described in the table. For some certain path, its cost is:

$$S = \sum_{m} \log(S_{m,m+1})$$

$S_{mn}$ is the cost to connect the node m and the node n.

$$S_{mn} = \sum_{j} R_j(C_m,C_n) \cdot Weight_{R_j}$$

$Weight_{R_j}$ is the weight when the jth rule is applied.

When merging the nodes we choose the rule, which makes the value of $Weight$ maximum locally.

When choosing the optimum path, we have tried many kinds of algorithm, including sliding dynamic window algorithm, $A^*$ searching algorithm, etc. Because matching one rule needs many information, considering both performance and efficiency, we adopt $A^*$ algorithm.

The rules are originated from both of mature syntax rules and statistical information. Because of the different contribution of each rule the cost coefficients are set up to embody the factor.

Each rule will possibly be linked to the other classified base, so the whole rules-table is designed as a mapping of 2D table. Such design doesn’t waste space, and give the possibility to the sharing of the main base.

This algorithm has fine effect in practice. Its shoot ratio can achieve 92.3% (for close text material) and 82.1% (for open text material).

The decision of L3 depends on the language understanding, which is also a difficult subject. We can avoid the problem at certain degree by adopting the statistical method and the information of L2. We introduce C45 Decision Tree to generate L3 and it’s lingual and prosodic information at the help of the L2 information. The C45 Decision Tree is built up by C45 arithmetic and the training text material which have been segmented and labeled by some lingual experts in advance. In the limit of maximum number of the words, at each division among L2 layers we calculate the probability in which it belongs to the boundary of L3.

There are about ten thousand training instances. The embranchments of Decision Tree are 468. Error ratio is 13.4%.

Through tests, we find the existent synthesis system can not express the L4 effectively. So the actual system hasn’t the process on L4.

4. Summary

This paper puts forward a kind of hierarchic processing model for Chinese TTS system, and defines corresponding hierarchic prosody labeling system with reference to TOBI labeling system. Hierarchic processing directly results in the idea of modularized processing, which makes the system easier to be clipped and maintained. Based on the idea of hierarchy process model, the KD2000 Chinese TTS system achieves very good performance.

Further research work includes: The analysis on fine structure of prosodic layer, automatically labeling on text material, the analysis on the mood, and context-sensitive analysis etc. These work now are undergoing.

Reference