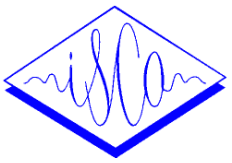

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPECIAL TEXT PROCESSING BASED EXTERNAL DESCRIPTOR RULE

Wu Xiaoru Wang Renhua Hu Guoping
Department of Electronic Engineering & Information Science
University of Science & Technology of China
P.O.Box. 4, HeFei, AnHui Province. P.R.China
Email: Wxr@mail.ustc.edu.cn Tel: 0551-3601363

ABSTRACT

In this paper, a special text processing method is presented, which integrates external descriptor rule and model matching. The letter-to-sound rules for a given kind of special mark can be described by some describable features which can be extracted from this kind of special mark, and then a corresponding model can be built and then the model parameter can be stored into a external form. Since the features can be extracted and described conveniently, a new model can be built easily for improving the ability to analyze special text.

1. INTRODUCTION

In this paper, All kinds of text except for Chinese characters are named as special text. The purpose of special text processing, also named text normalization, is to get enough information, such as pronunciation、rhythm and so on, for pronouncing special text correctly in Chinese TTS. Generally the way of reading a special text sometimes is extremely different from that of reading Chinese characters[1], in which we only need to read every Chinese character orderly according to the pronunciation in lexicon. For special text, the letter-to-sound rule can't be gotten until some specific knowledge is gained, such as some abbreviation, and some new kinds of special text may appear following the development of some subject.

The difficulty for processing special text lie in that computer can't understand the meaning of input text accurately now. In other word, the sentence or word meaning can't be used to help and instruct computer to recognize special text and to pronounce it. So we must extracted as much information as possible from special text itself for concluding the pronunciation and other correlative information of given special text. In Chinese, the main types of special text which are processed with difficulty are following:

A. Some symbols have different pronunciation in different context.
B. A string is consisted of digit、letter and other symbols and denote specific meaning. This kind of special text's sound can't be gotten until specific knowledge is gained.

The method we present in this paper can process special text with accuracy and is very flexible for further development. It possesses following virtues:

- The method of top-down analysis and hierarchical text processing is adopted.

  Different type of text will be processed on different level, such as punctuation mark、number and so on. According to the rules, which are defined in advance, particular marks will be processed at the current level. After being processed, the marks will be labeled as a child level of current level.

- A method, which integrates external descriptor rule and model matching, is adopted. According to different priority of the rules, the input mark will be matched with the every model which is described in an external form. If the matching succeed, the mark will be pronounced following the way given in this model. A few of distinctive and describable rules are summarized based on the intrinsic properties of some kind of marks such as decimal fraction, and then are saved in a form. Based on the rules we can decide if the input mark will be classified as a given class.

- To some marks whose pronunciations can't be defined by a perfect model, an enumerative method is used to avoid the difficulties which we can't solve now. For example, the pronunciation of number sometimes

must be defined only by fully semantic understanding.

- Additionally, in order to avoid mistakes due to mixing SBC case and DBC case or capital letter and lowercase, we'll normalize all input special marks into DBC case and capital letter after filtering out all kinds of control marks and random code.

2. THE CHARACTERISTIC MODELS AND EXTERNAL RULE FORM

The same as the way in which human recognize the special text, when computer processes them, it should also extract some useful correlative features and then match the feature with the models which can describe the feature of specific kind of special text. It has been denoted that the recognition for the given type of special text need specific knowledge, then how should the specific knowledge be expressed for the purpose of processing and extracting conveniently? Based on the idea, two types of characteristic parameter are used to help building the models for reading special text. The first type of parameter is named judgement parameter for concluding to which type the input special text is attributed. The second parameter is named pronunciation parameter for instructing how the special text's sound should be gained.

2.1 segmenting the special text

For the purpose of reading, remembering the other reasons, some fixed separate symbols are inserted into long character string, for example, the WWW address consists of several significant strings which are separated by the fixed symbol(dot). So these kinds of special text always can be segmented into several child-strings. There are two purposes for segmenting string. The first is that every child-string's feature can be extracted and described more easily and then all the features can be integrated to build final model. The other purpose is that the segmentation makes the model matching more easy during the period of special text analysis.

2.2 Building model

Since special text can be segmented into several describable child-strings (named main child-string in following paragraph) and separate symbol (named separated child-string in following paragraphs), then based on the features of each child-string we can build a model. Each model denotes a fixed type of special text. It can be considered as following: "each model is composed of commonness slots and individuality slots to describe the common traits of every main child-string and individual trait of each child-string respectively. Each slot's traits can be expressed by a multi-dimensional vector."

Generally the features denoted by commonness slots are used to restrict the number of child-strings, permitted character string style for all main child-strings and so on. The features denoted by individuality slots are used to restrict string's length, the string style of each child-string and so on. Each slot also takes a pronunciation parameter to instruct how the child-string's pronunciation should be gotten. For improving flexibility of this method, redundant slot is retained to add new feature for new model. For upgrading the efficiency in this method, each model parameter is organized in a tree list form and then stored into an external form.

2.3 Input special text analysis and model matching

Based on the characteristic model built in advance, the style and pronunciation of input special text can be concluded. At first depending on the separate symbols defined in the model, the input special text can be segmented into several strings, then the feature can be extracted from each child-string and give the feature value to the corresponding multidimensional vector. At last the multidimensional vector will be compared with the trait parameters which have been defined in the model. The matching strategy we adopt is full matching, which means that the model can't be selected as the appropriate one until all of the matching between each trait parameter and the given model parameter succeed. When the correct model is gotten, the pronunciation of each child-string is gained under the instruction of the model's pronunciation parameter and then the input special text's sound can be gotten by organizing pronunciation of each child-string in specific way defined in the model.

2.4 Adding new model

Because all of the models can be built under the same way and will be stored in an external form finally, users can add new model with the method of processing new types of special text which aren't defined in the external form. In fact when adding new model, only a model with empty slots should be constructed and then every slot should be endued with

appropriate parameter. When some undefined trait parameters must be used to restrict new model, they can be defined with the macro-way and then add the new parameters into the corresponding expanding slot..

2.5 An example

The course of building and matching model will be illustrated by the example. How to get the pronunciation of the email address "user12@263.net". Firstly a new model named email address is built and the string "@." 、"CHAR#||#DIG"and "#>4" are added into the first、second and third commonness slots to denote the separate symbols、the same character of every main child-string and permitted number of child-string respectively. The symbol "#" is a transferred meaning symbol to denote that the string between "#" has fixed and defined meaning.

During the period of matching model, the string ""user12@263.net"" is separated by the symbol '@' or '.' and then the feature is extracted from each child-string. The features of each child-string are as the same as that described in the model named "Email_address". So the pronunciation of the input string can be gotten as the way described in this model.

3. HIERARCHICAL TEXT PROCESSING METHOD

There are two reasons for processing input special text with hierarchical idea. The first one results from the internal laws for pronouncing special text because the relationship between some special text is closer than that of others. This kind of relationship should be taken into take into account to get prosodic information for improving the naturalness of speech when synthesizing special text. The other reason for processing input special text with hierarchical method lies in increasing the accuracy of processing special text and reducing the computational load. The special text, which are simple and can be processed with high credibility, will be analyzed at first. The complicated and dubious string should be processed later.

For realizing the idea of hierarchical text processing, a method of top-down text analysis is adopted. Each of rules and models is tagged with different priority according to different layer, so input special text can be processed in different layers with corresponding rules or models. In each layer, the input special text will be determined whether or not it should be split into different kinds of child-string.

Following example will be given to illustrate the hierarchical text processing method. The input special text is "a=12.3cm,b=40cm" and this string is set as root string. At first punctuation-rule has higher priority than that of the others when processing this string, so the string is split into three first floor child-string by this rule (shown in fig1). Then digit and unit-symbol processing-rule are used to analyze every first floor child-string, each first floor child-string is split into three second floor child-strings by the rules, finally in second floor child-string the number-string and unit-string will be united into one string. The processing result is shown in fig 1.

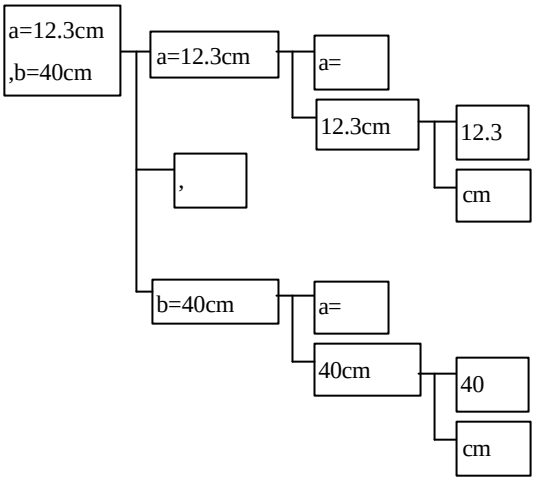

Fig. 1. Hierarchical text analysis process

It should be mentioned that some kinds of symbols have different functions at different context. For example the symbol '/' can sever as a punctuation and also as a specific symbol in fraction. So if we just see it as a punctuation symbol and use it to separate two numbers which all are parts of a friction, improper processing result will be inevitable. To avoid the mistake, some symbols should be tagged with different priority based on the context in which the symbols lie.

4. EVALUATION OF THE METHOD

To provide a set of quantitative evaluations for this method, we have performed a series of experiments. About 40 sentences, all of which contain special text, are extracted from different kinds of article by three persons who don't know why the sentences are needed. When the special text in the sentences are processed, only one mistake happen. The experiments indicate that special text can be processed accurately in the

method we propose in the paper.

5. CONCLUSION

In this paper a method is presented, which integrates external descriptor rule and model matching, and the commonness slots and individuality slots are adopted to describe the traits of the model. All the model parameters are organized in a tree list form and stored into an external form. The experiments show that very high accuracy can be gained for processing special text with the method.

Now the job of building new models and extracting new characteristic parameters for new models can't be performed without human's instruction. In the future we hope the tasks can be done by computers automatically with statistical or other method for improving the flexibility of analyzing text. Maybe the statistical way and non-deterministic finite automata[2] can help computer to extract feature and built model to certain extent.

Keywords: special text, external descriptor rule, model matching

5. REFERENCE

[1] Richard Sproat. Multilingual text analysis for text-to-speech synthesis. ICSLP96,Vo3, pages 1365-1368.

[2] Sangho Lee. A text analyzer for Korean text-to-speech systems. ICSLP96,Vo3, pages 1692-1693