# A CHINESE TEXT-TO-SPEECH SYSTEM BASED ON PART-OF-SPEECH ANALYSIS, PROSODIC MODELING AND NON-UNIFORM UNITS

*Fu-chiang Chou*[1], *Chiu-yu Tseng*,[2] *Keh-jiann Chen*[3] *and Lin-shan Lee*[1,3]
[1] Department of Electrical Engineering, National Taiwan University
[2] Institute of History and Philology, Academia Sinica
[3] Institute of Information Science, Academia Sinica
Taipei, Taiwan, Republic of China
*email-addr*: moza@speech.ee.ntu.edu.tw

## ABSTRACT

This paper presents a new Chinese text-to-speech system that produces very natural and intelligible synthetic Mandarin speech based on part-of-speech analysis, prosodic modeling and non-uniform units. The distinguishing features and key technology for the system can be summarized as follows:
(1) A text analysis module for word identification and tagging was developed based on part-of-speech modeling and using heuristic rules to achieve very high accuracy.
(2) The required prosodic parameters for the synthetic speech are derived from a two-stage procedure. The prosodic structures of the input texts are first derived from a statistical model trained by a large speech database, and the prosodic parameters are then determined according to the structures.
(3) A specially designed speech segments inventory constructed with non-uniform and pitch dependent units is used to improve the fluency and intelligibility of the system.

## 1. INTRODUCTION

In resent years, the performance of personal computers has dramatically increased. They also offer multi-media facilities such as audio I/O that makes it possible to have speech output in many applications. Although various efforts have been devoted to the improvement of speech synthesis technologies, TTS has not been widely accepted. The major reason is the naturalness of the generated speech is not sufficient enough, especially for Chinese.

In general, there are three key problems in developing a high-quality Chinese TTS system: the text analysis, the prosodic models and the synthesis units. For Chinese, we know that precise syntactic analysis of the sentence structure is difficult and large amount of computation required. So most of the previous systems just identify the words in the text analysis procedure and no more linguistic information can be used for the further prosodic processing [1][2]. In order to use the higher level linguistic information, the parts-of-speech must be precisely tagged in the text analysis. For the prosodic processing, we construct a statistical prosodic model to generate a hierarchical prosodic structure using the parts-of-speech as the main input features. The necessary prosodic parameters for the input sentence are determined by the generated prosodic structure. Different prosodic parameters can be determined on different level of the structure, including intonation, energy, tone sandhi, syllable duration and pause. For the synthesis units, Mandarin Chinese is a tonal language and each character is pronounced as a syllable. There are about 1300 syllables, which are the legal combinations of 408 base-syllables and 5 tones. Due to the fact that the total number of base syllable is only 408, syllable is commonly chosen as the basic synthesis unit in Chinese TTS. This kind of approach neglects the inter-syllabic coarticulation and causes a discontinuous effect of the synthetic speech [3]. Inter-syllabic units can be used to smooth this effects. The related problem of the units is the synthesis methods. The concatenation of speech segments with the pitch-synchronous Overlap and Add (PSOLA) techniques becomes much popular in recent years. However, experiments show that the method will cause audible distortions if the pitch change is larger than a factor of 2. This is not a special case in Chinese. A speech segments inventory constructed with non-uniform and pitch dependent units was designed to solve these two problems.

The paper is organized as follows. Three main modules of the system are described in section 2-4,. The complete system is described in section 5. The last section is the conclusions.

## 2. PART-OF-SPEECH ANALYSIS FOR WORD IDENTIFICATION AND TAGGING

Although a Chinese word is composed of one to several characters, a Chinese sentence is in fact a string of characters without blanks to mark the word boundaries. However, the basic unit for most linguistic processing is the word. Therefore the first step for text analysis is to

0-8186-7919-0/97 $10.00 

identify the words correctly in the input texts. Usually a lexicon with a large set of entries is used to match the input sentences. But sometimes there are many different successful matchings. The difficulties of word identification include (1) the existence of large number of compound words, word variants, etc. (2) the large number of proper names not included in the lexicon (3) the high degree of ambiguity because the Chinese words are actually not well defined and the segmentation of a sentence into words may not be unique [4]. After identifying words, A grammatical category (part-of-speech) must be assigned to each identified word. However, the occurrence of homographs and multi-function words is around 20% of the text [5]. To disambiguate multi-category words correctly becomes the major issue of tagging.

The diagram for word identification and tagging is illustrated in Figure 1. These two processes are integrated in order to share the linguistic information and achieve the global optimal solution. The input sentences are processed from left to right. The first step is to match the input character string with the lexicon and DM (determinative-measure compounds) rules. With these rules, the matching process works as if the lexicon contains all the DM's. The associated part-of-speech for the matched word is assigned in the same time. If an ambiguous segmentation or tagging occurs, the program looks ahead two more words and applies the disambiguation process for these three word chunks. Some heuristic rules select the most possible chunks and a Markov model then determines the best one. The first word of the chunk with maximal probability will be selected and the process will proceed until the end of the sentence. The probability for a possible word sequence $(W_1,W_2,W_3)$ and the corresponding tagging $(T_1,T_2,T_3)$ is defined as:

$$P(W,T) = \prod_{i=1}^{3} P(W_i)P(T_i|W_i)P(T_i|T_{i-1}) \quad (1)$$

The other types of compounds in Chinese are handled by many different morphological rules in the last procedure. Reduplication of verbs ("打打球" means "play ball"), A-not-A construction ("好不好", means "good or not" ) and derived word with a suffix ( "數位化" means digitize and "人工化" means artificialize) are the common examples that will be identified in the system. A corpus with manual word identification and tagging is used for training and testing. The accuracy is 98.6% for word identification and 95.2% for tagging excluding the unknown words.

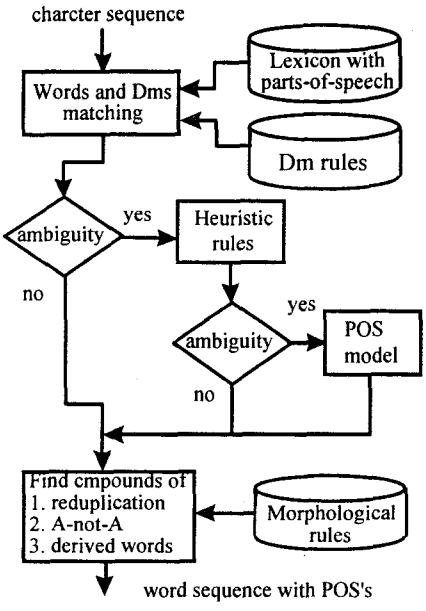


**Figure 1**: The diagram for word identification and tagging.

## 3. PROSODIC MODELING AND PARAMETERS GENERATION

The most important problem of speech synthesis is to generate the required prosodic parameters such as pitch, energy, pause and duration for the speech synthesizer to improve the naturalness. A statistical model to generate hierarchical prosodic structures for Chinese sentences using parts-of-speech as the primary input features was developed for this purpose [6]. The prosodic structure of a sentence can be represented by different levels of break markers in the model. Four levels of breaks were used in this model each with a break index: no break(0), minor break(1), major break(2), and punctuation mark break(3). Two types of prosodic phrases, i.e., minor prosodic phrases and major prosodic phrases, are determined by these break indices. The labeled break indices and the corresponding parts-of-speech of a tagged speech database were used to train a statistical model that can predict the prosodic structure of any input sentence. A grouping method for clustering the parts-of-speech based on statistical behavior of the words in the training database was used to improve the accuracy in prosodic structure prediction. The achieved accuracy is 83.1% for prosodic structure prediction.

Various prosodic parameters can then be determined on different levels of the prosodic structure, including pitch, energy, duration and pause. Global intonation and energy contours are first applied to the major prosodic phrases depending on the type of the ending punctuation mark. These contour patterns were extracted from the training speech database according to the punctuation marks or some special sentence-final particles (such as 嗎, 呢...) in Chinese. A dynamic range variation contour for $F_0$ was also extracted from the training database to realize the diminution effect of $F_0$ range in a sentence. In addition, an important characteristic of Mandarin Chinese is that it is a monosyllabic based tonal language and the interaction among the tones of syllables in a sequence is phonologically predictable at both the word and phrase levels. The $F_0$ patterns of all possible tonal combinations can be extracted from a carefully designed word speech database with some processing to take care of carious factors that may affect the realization of tones in actual speech. These tone patterns are used to construct the fundamental frequency contours of the words located in a minor prosodic phrase. The use of larger units for the tone patterns can improve the fluency of the synthetic speech. The duration d of a syllable is determined by various factors also:

$$d = d_i \bullet r_t \bullet r_p \bullet r_b \quad (2)$$

where $d_i$ is the average duration of the specific syllable $i$ and the other factors are the ratios based on the tone, the position of the syllable in phrase and in sentence, and the break index. The length of the pause between two words is determined by the break index. Furthermore, a small random number is used to adjust the pause length to avoid the effect of unnatural rhythm. The whole mechanism to determine the prosodic parameters is illustrated in Figure 2.

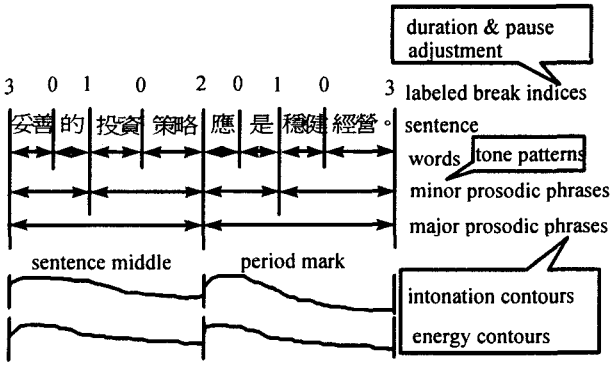


**Figure 2:** The mechanism for the determination of the prosodic parameters according to the prosodic structure.

## 4. SELECTION AND EXTRACTION OF NON-UNIFORM SYNTHESIS UNITS

Conventionally, Mandarin speech synthesizers usually use the syllable as the basic unit for concatenative synthesis. This approach assumes that coarticulation is minimal and negligible at syllable boundaries and no concatenation rules are necessary, which is definitely not a valid assumption for many cases. In the present system non-uniform units were carefully selected to take care of the coarticulation problem for different syllabic combinations. Two additional types of units were selected in addition to the widely used syllables. A disyllabic unit is a larger unit including a syllable and the FINAL part of the preceding syllable, selected for the case that a syllable is severely affected by the preceding syllable. Such units can be extracted when quick change in spectrum or energy is observed on the boundary between two syllables. The third type of units is the diphone, which is adopted only when no suitable larger units can be used. These units are summarized below:

1) 408 syllables (for the situation that the coarticulation effect is insignificant and negligible)
2) 1406 disyllabic units (for heavy coarticulation cases)
3) 220 diphones (to smooth some syllable boundaries)

The segments of these units were extracted from a specially designed speech database that includes most of the syllabic combinations. When some desired units were not found in the database, the unit with similar phonetic structure will be chosen.

Another special feature in the present system is the use of pitch dependent units. The pitch range in Mandarin is relatively large because it is a tonal language. The fundamental frequency values generated by the prosodic model mentioned in previous section are from 70Hz to 250Hz. Such a significant change in pitch periods may cause some distortion when TD-PSOLA is used. This problem is taken care of by multiple speech segments with different pitch periods. Three copies of speech segments were pre-synthesized with a shape-invariant pitch modification method that is less influenced by the changes of pitch periods [7]. The fundamental frequencies of these speech segments are 80Hz, 130Hz and 200 Hz which are equally spaced on log scale. When the syllable sequence and associated prosodic parameters are determined by the prosodic model, the suitable speech segments will be selected from inventory for the TD-PSOLA.

## 5. THE COMPLETE SYSTEM

Figure 3 is the block diagram of the complete system developed on a personal computer with a 16 bits sound card. The system includes three main modules: the text analysis module, the prosody generation module and the speech synthesis module. The text analysis module transforms an input sentence into a word sequence with parts-of-speech tagged. The text preprocessing sub-module handles problem of non-Chinese characters. Non-Chinese characters such as number (1,234, 363-5251, ...), symbols ($1.00, 95%, ...) and foreign languages (PM, TEL, ...) must be transformed to suitable Chinese character sequence. The word identification and tagging sub-module then generates the tagged word sequence as described in section 2. The prosody generation module generates the corresponding prosodic structure and required prosodic parameters. The prosodic model and the prosodic templates are both trained and extracted from a speech database of a male speaker. The resulted prosodic speaking style can be easily identified when used in conjunction with the speech units extracted from the same speaker. A new model is under construction for a female speaker. The speech synthesis module finally selects the suitable speech segments based on the syllabic combinations and $F_0$ values. These speech segments are concatenated to generate the output speech by TD-PSOLA algorithm. Since the pitch change is smaller than a factor of 2 by using pitch-dependent units, it does generally not result in audible distortions in the output speech. Preliminary listening tests have confirmed that the synthesized speech sounded very fluent and natural. A formal listening and diagnostic test will be designed to evaluate the system and used for further adjustment.

## 6. CONCLUSIONS

In this paper, we have presented a high-quality Chinese text-to-speech system. The system can transfer Chinese text into natural Mandarin speech based on part-of-speech analysis, prosodic modeling and non-uniform units. These technologies significantly improve the naturalness and quality of the TTS system. The system is also modularized for easily incorporating to many applications with speech output. Although these speech applications are not popular now, we believe the people will like to listen to the natural-sounding Mandarin speech from the computer. Many applications are under construction with the Chinese TTS system presented here.

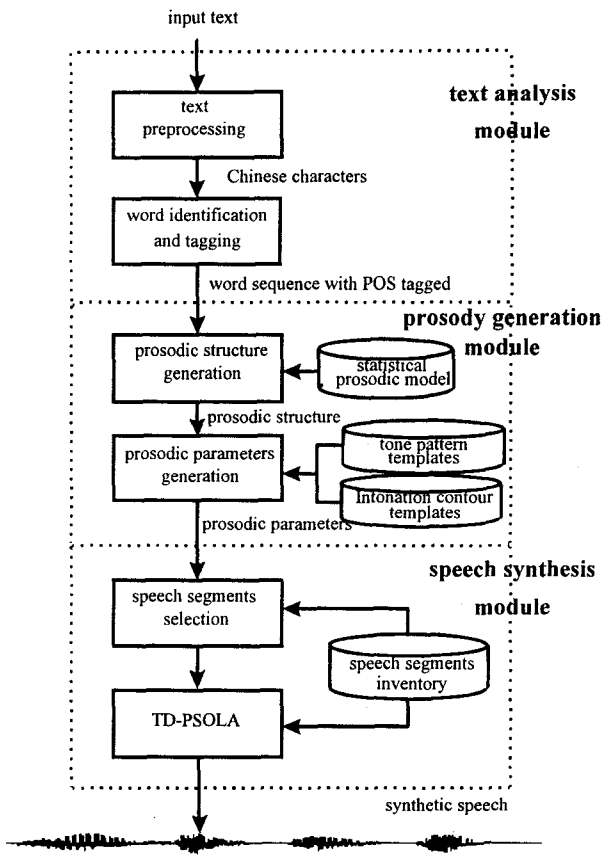


**Figure 3**: The block diagram of the text-to-speech system

## REFERENCES:


[1] Ren-hua Wang, Qinfeng Liu and Difei Tang, "A New Chinese Text-to-Speech System with High Naturalness", ICSLP, 1996, pp. 1441-1444

[2] John Choi, Hsiao-wuen Hon, etc., "Yanhui, a Software Based High Performance Mandarin Text-to-Speech System", ROCLING, 1994, pp. 35-50

[3] Shaw-hwa Hwang, Sin-horng Chen and Yih-ru Wang, "A Mandarin Text-to-Speech System", ICSLP, 1996, pp. 1421-1424

[4] Keh-jiann Chen and Shing-huan Liu, "Word Identification for Mandarin Chinese Sentences", COLING-92, pp.101-107

[5] Li-ping Chang and Keh-jiann Chen, "The CKIP Part-of-Speech Tagging System for Modern Chinese Texts", ICCPOL, 1995, pp.172-175

[6] Fu-chiang Chou, Chiu-yu Tseng and Lin-shan Lee "Automatic Generation of Prosodic Structure for High Quality Mandarin Speech Synthesis", ICSLP, 1996, pp. 1624-1627

[7] R. J. McAulay and T.F. Quatieri, "Shape Invariant Time-scale and Pitch Modification of Speech", IEEE Trans. On Signal Processing, 1992, pp. 497-510