A MANDARIN TEXT-TO-SPEECH SYSTEM

Shaw-Hwa Hwang, Sin-Horng Chen, and Yih-Ru Wang

Department of Communication Engineering and
Center For Telecommunications Research
National Chiao Tung University
Hsinchu, Taiwan 300, Republic of China
email:u8011854@cc.nctu.edu.tw, schen@cc.nctu.edu.tw
Tel:+886-35-731822, Fax:+886-35-710116

ABSTRACT

In this paper, the implementation of a high-performance Mandarin TTS system is presented. The system is composed of four main parts: text analysis (TA), prosodic information generation (PIG), waveform table (WT) of 411 base-syllables, and PSOLA-based waveform synthesis (PSOLA). In TA, a statistical model based method is first employed to automatically tag the input text to obtain the word sequence and the associated part-of-speech (POS) sequence. A lexicon containing about 80000 words is used in the tagging process. Then the corresponding base-syllable sequence is found and used to get from WT the basic waveform sequence. Some linguistic features used in PIG are also extracted in TA. In PIG, a four-layer recurrent neural network (RNN) is employed to generate some prosodic information including pitch contour, energy level, initial duration and final duration of syllable as well as inter-syllable pause duration. Finally, in PSOLA the basic waveform sequence is modified using the prosodic information to generate output synthetic speech. The whole system is implemented by software on a PC/AT 486 with a 16-bit Sound Blaster add-on card. Only 3.2 Mbyte memory space is required. It can synthesize speech in real-time for any input Chinese text. Informal listening tests by many native Chinese living in Taiwan confirmed that the synthetic speech sounded very fluent and natural.

1. INTRODUCTION

The general goal of a text-to-speech (TTS) system is to mimic the pronunication style of human beings in order to utter clear, natural, and fluent speech for unlimited input texts. The first TTS system was presented in 1968 for English [4]. Since then, many other TTS systems have been proposed for various languages. In the past, TTS systems usually adopted the rule-based approach to generate prosodic information [1, 2, 3]. Although some of them have been demonstrated to be of high performance, it still remains a general difficulty to manually infer a proper set of rules for synthesizing high-quality synthetic speech. In recent years, a new approach which uses a statistical model or a neural net to automatically learn rules from a large set of training data has been proposed. It is usually referred to as the data-driven approach. Its effectiveness has been confirmed by many successful examples. Basically, it is much simpler than the conventional rule-based approach because the difficulty of manually analyzing the pronunciation rules of human beings is avoided. Now, high performance TTS systems have been demonstrated for many languages including English [4], German [5], French [6], and Mandarin Chinese. In this paper, a real-time implementation of high-performance Mandarin TTS system is presented.

Mandarin Chinese is a tonal language. Each character is pronounced as a syllable. There are only about 1300 phonetically distinguishable syllables, which are the set of all legal combinations of 411 base-syllables and 5 tones. Each base-syllable is composed of an optional consonant initial and a vowel final. Although word which consists of one to several syllables is the smallest syntactically meaningful unit, syllable is the basic pronunciation unit in Mandarin speech. Due to the fact that the total number of base-syllable is only 411, syllable is commonly chosen as the basic synthesis unit in Mandarin TTS. In our TTS system, we also adopted the same approach. One advantage of the approach is that we need not to consider the intra-syllable coarticulation for spectral information synthesis. As for inter-syllable coarticulation, we can also neglect it because it is usually not serious in affecting the quality of synthetic speech. So syllables can be directly concatenated without any spectral smoothing. This only results in a slight degradation on the quality of the synthetic speech. Based on above discussions, we find that spectral information synthesis is a relatively easy task for Mandarin TTS.

We now consider the prosodic information synthesis. Although the phonetic structure of Mandarin syllable is very simple, the prosodic structure of Mandarin sentential utterance is much more complicated. Many factors may affect the generation of prosodic information. They include linguistic features of all levels of syntactical structure, the semantics, the speaking habit and the emotional status of

This work was supported by the National Science Council, ROC, under contract NSC84-2213-E009-097. The authors want to thank Telecommunication Laboratories, MOTC, ROC for supporting the speech database. We also want to thank Academia Sinica for supporting the lexicon.

the speaker, the pronunciation environment, etc. So the generation of proper prosodic information from the input text is not a trivial problem. This makes prosodic information generation be a main concern on developing a high performance Mandarin TTS system. In the past, researches on the task of synthesizing prosodic information for Mandarin TTS were very limited. It is therefore very urgent to work out a good algorithm of prosodic information synthesis in order to develop a high performance Mandarin TTS system. In our TTS system, a novel neural network based prosodic information synthesis algorithm is proposed. Due to the fact that it is the dominating factor affecting the naturalness of the synthetic speech, we will emphasize our discussions on it.

The remainder of the paper is stated as follows. Section 2 presents the the proposed Mandarin TTS system. The real-time implementation of the system by software on a PC/AT 486 is presented in Section 3. Performance evaluation of the system is also discussed. Some conclusions are given in the last section.

2. SYSTEM DESCRIPTON

Fig.1 shows the block diagram of the proposed system. It is functionally composed of four main parts: text analysis (TA), prosodic information generation (PIG), waveform table (WT) of 411 base-syllables, and PSOLA-based speech synthesis (PSOLA). Input Chinese text in the form of character sequence is first tagged in TA to obtain the best word sequence and the best part-of-speech (POS) sequence simultaneously. The corresponding syllable sequence is then extracted and used in WT to find the basic waveform sequence. Some word-level and syllable-level linguistic features are also extracted and used in PIG to generate the prosodic information. Last, the basic waveform sequence is modified in PSOLA by using the prosodic information to generate the output synthetic speech. In the following, all the four main parts are discussed in detail.

2.1. Text Analysis

The task of TA is to analyze the input text to extract some linguistic features needed for synthesizing both spectral information and prosodic information. In our system, this is realized by first tagging the input text to obtain the word sequence and the POS sequence and then extracting the required linguistic features from them. Fig.2 shows the block diagram of TA. Input text in the form of character sequence represented by the Big-5 code is first tagged by using a statistical model based method to find the best word sequence and the best POS sequence simultaneously. An 80000-word lexicon containing 1- to 5-syllabic words and a POS bigram model calculated from a database containing utterances of short sentences and paragraphic text are used in the tagging process. It is an optimal search procedure which maximizes the following objective function

$$S(W,T) = L^2(w_1) + \eta \log P(t_1) + \sum_{i=2}^{M} [L^2(w_i) + \eta \log P(t_i|t_{i-1})] \quad (1)$$

Here $W = (w_1, w_2, ..., w_M)$ and
$T = (t_1, t_2, ..., t_M)$ are respective a candidate word sequence and an associated POS sequence of the input sentence, $L(w_i)$ is the number of characters in word w_i, $\eta(= 0.33)$ is a weighting factor, M is the number of words of the sentence, and $P(t_1)$ and $P(t_i|t_{i-1})$ are the initial and the transition probabilities of the POS bigram model. The optimal search procedure is efficiently implemented by a Viterbi search algorithm.

After obtaining the optimal word and the associated POS sequences, we then use two additional rules to construct two types of compound words which are not contained in the lexicon. One is for the character-duplicated compound word and the other is for the determiner-measure compound word.

Two sets of linguistic features are then extracted from the word and the POS sequences. One is the syllable sequence which is extracted from the word sequence by looking up the lexicon. It will be used in WT to obtain the basic waveform sequence. The other consists of two subsets of syllable-level and word-level linguistic features and is used in PIG to synthesize proper prosodic information. The subset of syllable-level linguistic features contains four sequences of *consonant types of syllables, vowel types of syllables, tones of syllables,* and positions of syllables in the corresponding words. The subset of word-level linguistic features includes *the POS sequence, and two sequences of word lengths* and *punctuation marks.*

2.2. Prosodic Information Generation

The task of PIG is to generate proper prosodic information by using the linguistic features generated in TA. In our system, an RNN-based approach is adopted. It employs a four-layer RNN with two hidden layers to simulate human's prosody pronunciation mechanism for generating all prosodic information required in our system. They include pitch contour, energy level, initial duration and final duration of syllable as well as inter-syllable pause duration. Fig.3 shows the block diagram of the RNN. It can be functionally partitioned into two parts. The first part consists of the input layer and the first hidden layer and is taken as a prosodic model to explore the prosodic phrase structure of the synthetic speech by using the input word-level linguistic features. It operates on a clock synchronized with word to generate outputs representing the phonologic state of the prosodic phrase structure at the current word. The second part consists of the second hidden layer and the output layer. It operates on a clock synchronized with syllable to generate the prosodic information by using the prosodic state fed-in from the first part and the input syllable-level linguistic features. Trained with a large set of real utterances accompanying with the associated texts, the RNN prosody synthesizer can automatically learn many prosody phonologic rules of human beings including the well-known F0 sandhi rule of Tone 3 change. It can therefore be used to generate proper prosodic information required for synthesizing natural and fluent speech.

2.3. Waveform Table

The function of WT is to provide the basic primitive waveforms of the synthetic speech. It stores waveform templates of all 411 base-syllables which are the basic synthesis units used in our system. All waveform templates of syllables are obtained via semi-automatically selecting from the training set which contains many sentential and paragraphic utterances. Before stored in the WT, each selected waveform template is further processed to normalize its energy contour to the average of all energy contours of the same base-syllables in the training set. A segmental k-mean algorithm is employed to obtaine the average energy contours of all base-syllables. In synthesis, all constituent waveform templates of the input syllable sequence are sequentially extracted from WT, directly concatenated together, and sent to PSOLA for further modification.

2.4. PSOLA-based Speech Synthesis

Recently, the PSOLA-based speech synthesizer is widely used in TTS. It can generate high-quality synthetic speech in low computational complexity. Due to its superiority, a PSOLA-based speech synthesizer is adopted in our TTS system. It generates the output synthetic speech by modifying the input basic primitive waveform sequence to let its prosodic parameters match with the input prosodic information given by PIG. Modifications include changing the pitch contour for each syllable, adjusting the durations of the initial consonant and the final vowel of each syllable, scaling the energy level of syllable, and setting the inter-syllable pause duration. Finally, the output synthetic speech is generated from a 16-bit Sound Blaster add-on card.

3. EXPERIMENTAL RESULTS

A speech database provided by Telecommunication Laboratories, MOTC, ROC was used to realize our Mandarin TTS system. The database contains 655 sentential and paragraphic utterances pronounced by a single male speaker. The database is divided into two parts. The first part composing 28191 syllables was used to construct a real-time version of the TTS system. The second part containing 7051 syllables was then used to quantitaively evaluate its performance. All speech signals and the associated texts were manually pre-processed in order to extract the acoustic features and the linguistic features required to train and test the system.

The whole system is implemented by software on a PC/AT 486 with a 16-bit Sound Blaster add-on card. Table 1 lists the memory space of the system. Only 3.2 Mbyte RAM is required. It can synthesize speech in real-time for any input Chinese text. Table 2 displays the experimental results of the prosodic information synthesis. Fig.4 shows a typical segment of the synthesized pitch contour. Can be seen from the figure that, for most syllables, the synthesized pitch contours match quite well with their original counterparts. It is worth to note that the tone of the 8th syllable has been changed from Tone 3 to Tone 2 in the original speech. As shown in the figure that the system synthesizes it correctly. By carefully listening examining the original and the synthetic speeches of the database, we found that about 86% of 3-3 tone pairs and 77.4% of 3-3-3 tone trigrams have been correctly synthesized. An informal subjective listening tests using various texts which are not included in the database was also derived to examine the performance of the system. Many native Chinese listeners living in Taiwan confirmed that the synthetic speech sounded very fluent and natural.

Finally, sound of a new female speaker has been added to the system. A training set containing 1000 syllables was collected. To adapt the system to the new speaker, we first selected waveform templates of the 411 base-syllables from the new training set to replace the WT. Then, the statistics (means and standard deviations) of the prosodic parameters used in the system were calculated. Instead of retraining the RNN, we simply synthesize the prosodic information by denormalizing the outputs of the RNN using the statistics of the new speaker. This makes great savings on both the time to retrain the RNN and the work to collect and process a large training set. Informal listening test confirmed that all synthetic speech still sounded very fluent and natural. This shows that the PIG is very roboust.

4. CONCLUSIONS

We have presented in the paper a high-performance Mandarin TTS system. It is a software package run under the PC/DOS environment. It can transfer any Chinese text into natural and fluent Mandarin speech in real-time. Several applications of the system are now under developing. Some advantages of the system can be found. First, the pronunciation rules of prosody generation are automatically inferred. The difficulty of manually analyzing pronunciation rules in the rule-based approach is avoided. Second, the RNN-based prosodic information generator can be easily adapted to a new speaker without intensively retraining. Third, the synthetic speech sounds very natural and fluent. Last, it is a real-time implementation by software only. It is therefore very easy to develop some applications.

5. REFERENCES

[1] Carlson, R., Granstrom, B.(1979). "A text-to-speech system based entirely on rules," Proc. ICASSP, pp.686-688,1976.

[2] Lin-Shan Lee, Chiu-Yu Tseng, and Ming Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," IEEE Trans. ASSP. Vol-37, 1989. pp.1309-1320.

[3] H. Mixdorff and H. Fujisaki, "A Scheme for a Model-based Synthesis by Rule of F0 Contours of German Utterances," in Proc. EUROSPEECH 95, pp.1823-1826, 1995.

[4] Klatt, D. H. "The Klatt-Talk Text-to-Speech System," Proc. ICASSP, pp.1589-1592, 1982.

[5] C. Traber, "Syntactic Processing and Prosody Control in the SVOX TTS system for German," EUROSPEECH'93, pp.2099-2102. 1993.

[6] Evelyne Tzoukermann. "Issues in Text-to-Speech for French," Int. Conf. on Computational Linguistics, Kyoto, Japan, 1994.

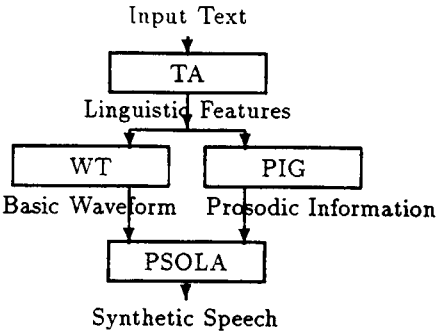

Fig.1 The block diagram of the proposed TTS system

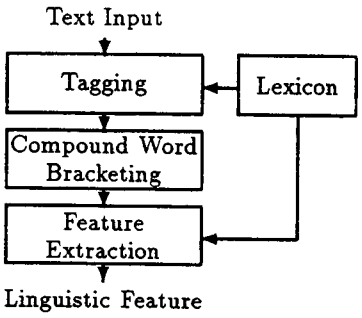

Fig.2 The block diagram of TA

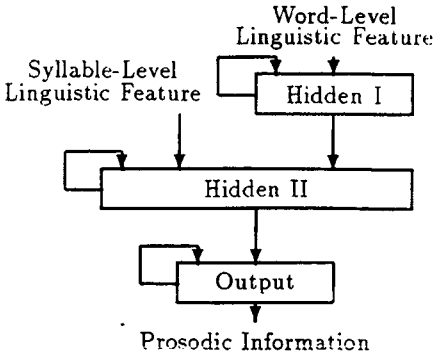

Fig.3 The block diagram of PIG

Table 1. Memory space allocation of the systemt(Unit:Byte)

Program	74.575K
TA(Lexicon)	732.278K
PIG(RNN-Weights)	39.072K
WT	2.3M
Total	3.146M

Table 2. The RMSEs of the five synthesized prosody parameters.

	Inside Test	Outside Test
F0 Contour	0.84ms/Frame	1.06ms/Frame
Pause Duration	23.7ms/Syllable	54.5ms/Syllable
Initial Duration	17.2ms/Syllable	18.5ms/Syllable
Final Duration	33.3ms/Syllable	36.7ms/Syllable
Energy Level	3.39dB/Syllable	4.17dB/Syllable

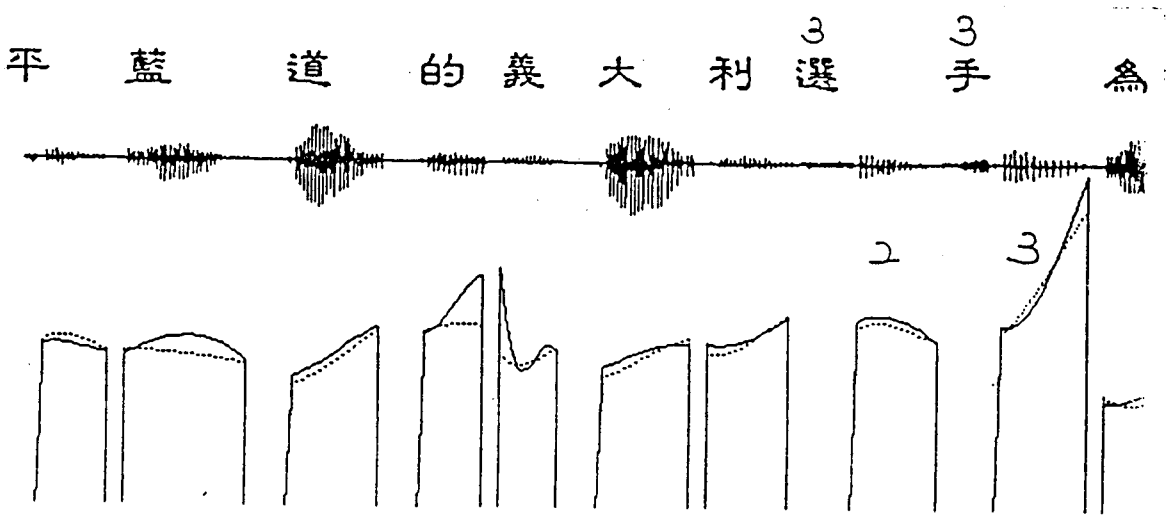

Fig.4 The Original and Synthesized Pitch Contour.