Computational Linguistics and Chinese Language Processing
vol.1, no.1, August 1996, pp. 87-100

A Mandarin Text-to-Speech System

Sin-Horng Chen*, Shaw-Hwa Hwang+, Yih-Ru Wang*

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), a waveform table of 411 base-syllables (WT), and PSOLA-based waveform synthesis (PSOLA). In TA, 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 in WT to form the basic waveform sequence of the base-syllables. 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 the pitch contour, energy level, initial duration and final duration of syllables as well as the inter-syllable pause duration. Lastly, in PSOLA, the basic waveform sequence is modified using the prosodic information to generate output synthetic speech. The whole system has been implemented by software on a PC/AT 486 with a 16-bit Sound Blaster add-on card. Only memory spaces of 3.2 Mbyte and 5.5 Mbyte are, respectively, required for the two versions with sampling rates of 10 kHz and 20 kHz. It can synthesize speech in real-time for any input Chinese text. Informal listening tests by many native Chinese living in Taiwan have confirmed that the synthetic speech sounds very fluent and natural.

keyword: speech synthesis, prosodic information, recurrent neural network, PSOLA, Mandarin Text-to-Speech

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. 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 [Klatt 1976, Carlson

* National Chiao Tung University, Hsinchu, Taiwan. E-mail: {schen, yrwang}@cc.nctu.edu.tw
+ Sillicon Integrated System Corporation, Hsinchu, Taiwan. E-mail: u8011854@cc.nctu.edu.tw

and Granstrom 1979, Lee *et al.* 1989]. Although some of them have been demonstrated to have 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 [Mixdorff and Fujisaki 1995], German [Sagisaka 1990], French [Traber 1993], Japanese [Klatt 1982], and Mandarin Chinese.

In this paper, the real-time implementation of a high-performance Mandarin TTS system is presented. Two fundamental issues must be considered first for developing a high-perfromance TTS system. They are the determination of the type of synthesis unit and the speech synthesizer. We will discuss them in more detail in what follows.

In the past, many types of synthesis units have been suggested for TTS. They include phoneme (phone), diphone, demi-syllable, syllable, word, and phrase. Among the many factors used to determine the type of synthesis unit, the memory space required to store all the synthesis units and the complexity of post-processing to smooth out abrupt spectral changes which occur when directly concatenating two synthesis units are most important. Basically, the smaller the type of synthesis unit selected is, the less memory space and the more complicated concatenation rules are required. So a tradeoff between the required memory space and the system complexity exists in determining the type of synthesis unit. Two basic approaches have been studied in the past. One is to choose phone-like synthesis units and to then apply them to a spectral smoothing technique to modify the part near the boundary when we concatenate two synthesis units. Another is to model each synthesis unit by using multiple templates and to then choose a proper one in the synthesis according to the context. Usually, no spectral smoothing is needed in the latter approach. In [Moulines and Charpentier 1990], a context oriented clustering (COC) algorithm was proposed to automatically select, in the training, all the synthesis units from a training database by using the clustering technique to consider contextual coarticulation. A score matching rule is then applied in the synthesis to determine the most suitable synthesis units used for an input text. This method has been proven to be successful in selecting a set of proper synthesis units for Japanese TTS. In [Nakajima and Hamada 1988], a system that generates a spectral parameter sequence by concatenating the spectral parameters of the synthesis units of pseudo-phonemes was proposed. In [Mikuni and Ohta 1986], a decision-tree-based clustering method that combines acoustic

and linguistic knowledges with statistical modeling was proposed. This method can not only find a trainable and consistent set of generalized allphonic models, but also can achieve some local optimality with respect to the limited training data. Experimental results showed that the use of regression trees offers a promising solution for the data scarcity problem.

For Mandarin TTS, determination of proper synthesis units is relatively easy. 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 the legal combinations of 411 base-syllables and 5 tones. Each base-syllable is composed of an optional consonant initial and a vowel final. Although a word which consists of one to several syllables is the smallest syntactically meaningful unit, the syllable is the basic pronunciation unit in Mandarin speech. Due to the fact that the total number of base-syllables is only 411, syllables are commonly chosen as the basic synthesis units in Mandarin TTS. In our TTS system, we also adopt this approach. One advantage of this approach is that we need not consider intra-syllable coarticulation for spectral information synthesis. Moreover, because inter-syllable coarticulation is usually not serious, we can also neglect it by directly concatenating two syllables without any spectral smoothing. This results in only slight degradation in the quality of the synthetic speech.

For the determination of a speech synthesizer, three types of candidates can be chosen. They include linear prediction (LP)-based vocoders, waveform coders, and PSOLA-based waverform synthesizer. The speech quality, the flexibility of adjusting prosodic parameters, the memory space required to store all the synthesis units, and the computational complexity are the four major factors considered in determining a speech synthesizer in a TTS system. Basically, the LP-based vocoder has fair speech quality, low memory space requirements, moderate-high complexity, and a moderate degree of flexibility in adjusting prosodic parameters. On the other hand, the PSOLA-based waveform synthesizer has good speech quality, low computational complexity, and high flexibility in prosody adjustment. However, it requires a relatively large memory space to store the waveform templates of all the synthesis units. Due to the fact that the cost of semiconductor memory has dropped very quickly in recent years, this factor has become unimportant. So, the PSOLA-based synthesizer is very popular now. We also adopt it in our TTS system.

We will now consider prosodic information synthesis. Although the phonetic structure of Mandarin syllables is very simple, the prosodic phrase structure of Mandarin speech is much more complicated. Many factors may affect the generation of prosodic

information. They include the linguistic features of all levels of syntactic structure, the semantics, the speaking habits and the emotional status of 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 a primary concern in developing a high performance Mandarin TTS system. In the past, research on the task of synthesizing prosodic information for Mandarin TTS has been 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 [Tzoukermann 1994] is used.

The remainder of the paper is organized as follows. Section 2 presents the proposed Mandarin TTS system. The real-time implementation of the system using 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), a waveform table of 411 base-syllables (WT), and PSOLA-based speech synthesis (PSOLA). Input Chinese text in the form of character sequences represented by the Big-5 code 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 four main parts will be 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 simultaneously obtain the word sequence and the POS sequence and by then extracting the required linguistic features from them. Fig. 2 shows the block diagram of TA. Input text in the form of character sequences 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

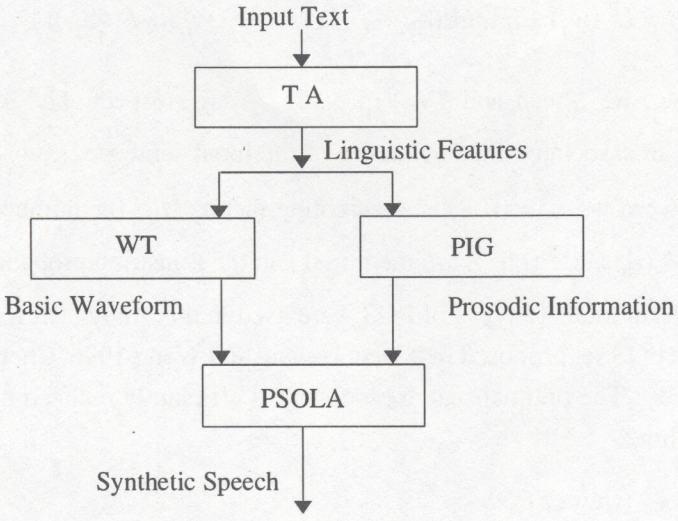

Figure 1 *The block diagram of the proposed TTS system*

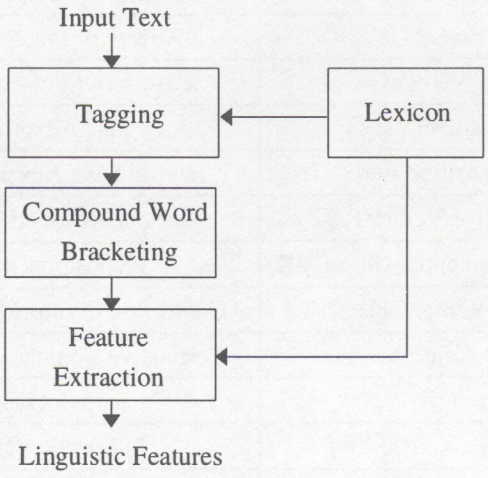

Figure 2 *The block diagram of TA*

containing 1- to 5-syllabic words and a POS bigram model calculated from a database containing 655 utterances of short sentences and paragraphic texts are used in the tagging process. The lexicon was provided by Academia Sinica, Taiwan, ROC. It is an optimal search procedure which maximizes the following objective function:

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

Here $W = (w_1, w_2, ..., w_M)$ and $T = (t_1, t_2, ..., t_M)$ are, respectively, 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, η (= 0.33) is a weighting factor, M is the number of words in 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. In total, 42 types of POS were used in this study. They were obtained by modifying the POS set proposed in [Chen, Hwang and Wang 1996, Chen 1989]. They are listed in table 1. The optimal search procedure is efficiently realized by using a Viterbi search algorithm.

Table 1. *The 42 types of POS*

Active Intransitive Verb(VA)	Adverb of Quantity(DA)
Active Pseudo-Transitive(VB)	Adverb of Evaluation(DB)
Active Transitive Verb(VC)	Negation(DC)
Ditransitive Verb(VD)	Adverb of Time(DD)
Active Verb with a Sentential Object(VE)	Adverb of Degree(DE)
Active Verb with a Verbal Object(VF)	Adverb of Place(DF)
Classificatory Verb(VG)	Adverb of Manner(DG)
Stative Intransitive(VH)	Aspectual Adverb(DI)
Stative Pseudo-Transitive Verb(VI)	Interrogative Adverb(DJ)
Stative Transitive Verb(VJ)	Sentential Adverb(DK)
Stative Verb with a Sentential Object(VK)	Preposition(P)
Stative Verb with a Verbal Object(VL)	Coordinate Conjunction(CA)
Nonpredicative Adjective(A)	Correlative Conjunction(CB)
General Noun(NA)	Particle(T)
Special Noun(NB)	Interjection(I)
Place Noun(NC)	Bound(B)
Verb-Complement Compound(VR)	Time Noun(ND)
Determiner(NE)	Sentence
Special Verb1(is, are, am)(V1)	Measure(NF)
Special Verb2(has, have)(V2)	Localizer(NG)
Determiner-measure Compound(DM)	Pronoun(NH)

After obtaining the optimal word sequence and the associated POS sequence, we then use two additional bracketing rules to construct two types of compound words which are not contained in the lexicon. One is for character-duplicated compound words and the other is for determiner-measure compound words. Fig.3 displays two typical examples of tagging results of TA.

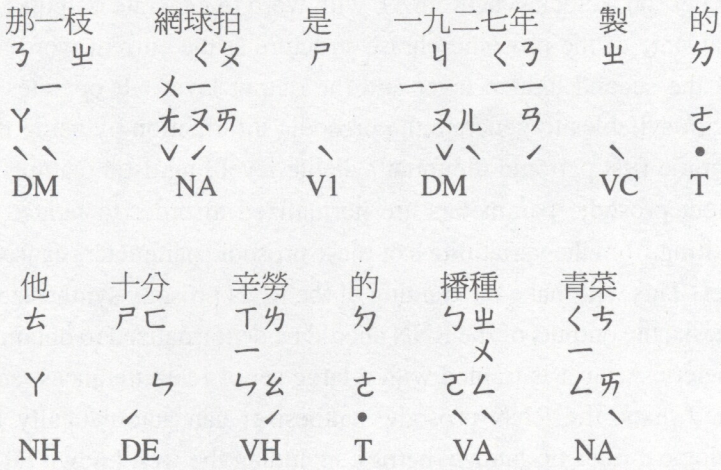

Figure 3 *Two typical Examples of tagging results of TA*

Two sets of linguistic features are then extracted from the two sequences of words and POS. 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. We note that the problem of Pohin characters is serious only for monosyllabic words and is not solved here. The other set 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: *the consonant type of the syllable*, *the vowel type of the syllable*, *the tone of the syllable*, and *the position of the syllable in the corresponding word*. The subset of word-level linguistic features includes *the POS sequence, and two sequences of word length* 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 the simulate human's prosody pronunciation mechanism to generate all the prosodic parameters required in our system. They include

4 parameters representing the pitch contour of the syllable, 3 parameters, respectively, representing the energy level, initial duration, and final duration of the syllable, and 1 parameter representing the inter-syllable pause duration. Fig. 4 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 phonological 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 syllables to generate the prosodic information by using the prosodic state fed in from the first part and the input syllable-level linguistic features. It is noted that all the output prosodic parameters are normalized in order to reduce the system complexity resulting from the variabilities of these prosodic parameters caused by lexical phonetic features. This will make the training of the RNN prosody synthesizer easier. Of course, in synthesis, the outputs of the RNN should be denormalized to obtain the desired prosodic parameters. After it is trained with a large set of real utterances accompanied by the associated texts, the RNN prosody synthesizer can automatically learn many prosodic phonological 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 to synthesize natural and fluent speech.

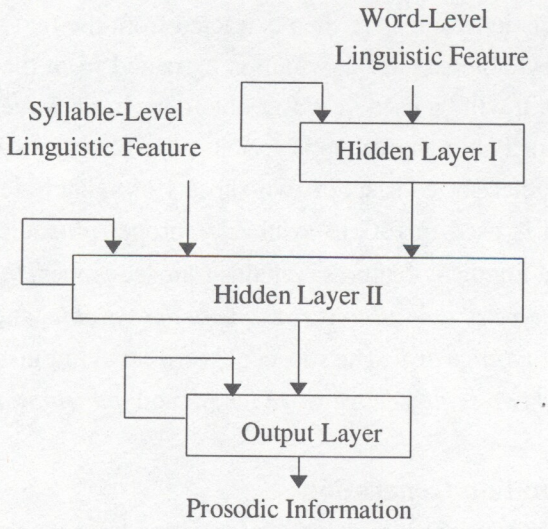

Figure 4 *The block diagram of PIG*

2.3 Waveform Table

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

2.4 PSOLA-based Speech Synthesis

Recently, the PSOLA-based speech synthesizer has been widely used in TTS [Chang *et al.* 1989]. It can generate high-quality synthetic speech with 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 enable its prosodic parameters to match 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 for each syllable, scaling the energy level for each syllable, and setting the inter-syllable pause durations. Finally, the output synthetic speech is generated from a 16-bit Sound Blaster add-on card.

3. Experimental Results

A speech database provided by the Telecommunication Laboratories, MOTC, ROC, was used to realize our Mandarin TTS system. The database contained 655 sentential and paragraphic utterances pronounced by a single male speaker. The database was divided into two parts. The first part, composed of 28191 syllables, was used to construct a real-time version of the TTS system. The second part containing 7051 syllables was then used to quantitatively evaluate its performance. All the 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 was implemented using software on a PC/AT 486 with a 16-bit Sound Blaster add-on card. Table 2 lists the memory space of the system. Only 3.2 Mbyte

and 5.5 Mbyte RAMs were, respectively, required for the two versions with sampling rates of 10 kHz and 20 kHz. It was able to synthesize speech in real-time for any input Chinese text.

Fig. 5 displays a typical example of the synthesized pitch contours of syllables for an input sentential text. It can be seen from the figure that most of the synthesized pitch contours of the syllables resemble their original counterparts in both level and shape.

Table 2. *Memory space requirements for the TTS system*

Sample Rate	10KHz	20KHz
Program	74.575K	74.575K
TA(Lexicon)	732.278K	732.278K
PIG(RNN-Weights)	39.072K	39.072K
WT	2.3M	4.6M
Total	3.146M	5.446M

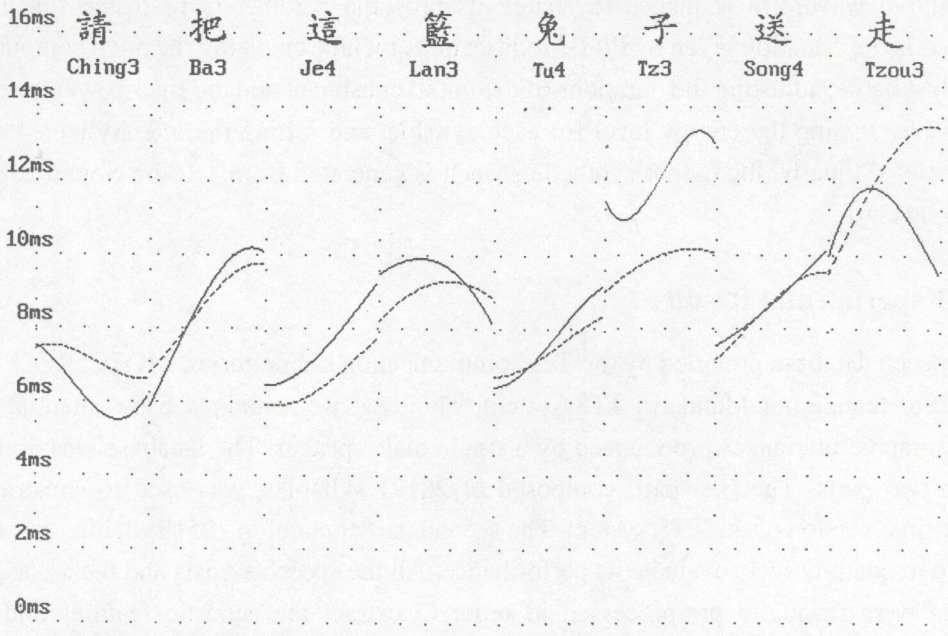

Figure 5 *The original(solid line) and the synthesized(dotted line) pitch contours for an input sentential text*

(Note that the duration of the syllable is normalized to a fixed length.)

This confirms that the declination effect has been correctly learned by the RNN prosody synthesizer. We also find from the figure that the synthesized pitch contour of the first syllable deviates from the standard pattern of Tone 3 and looks like a Tone 2. This shows that the sandhi rule of Tone 3 change has been correctly implemented here. Actually, by carefully examining all the synthesized pitch contours of 3-3 tone pairs and 3-3-3 tone trigrams, we find that the sandhi rule of Tone 3 change has been correctly learned by the RNN prosody synthesizer.

An informal subjective listening test using various texts which were not included in the database was finally 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. It is worth noting that, although many testing texts used in the listening test were not well tagged by TA because they contained some words not included in the lexicon, unnaturalness of the synthetic speech was not often heard. This result is explained as follows. It is known that the prosodic phrase structure of Mandarin speech is not isomorphic to the syntactic structure. Basically, the prosodic phrase structure of an utterance is affected by both the syntactic information of the text and the durational information of the current pronunciation. So, a bad tagging does not have to result in catastrophic prosodic information synthesis. A good prosody synthesizer can use the given syntactic information and the corresponding duration information to choose proper prosodic information. Experimental results show that the RNN prosody synthesizer performs well for most cases of bad tagging.

Lastly, the sound of a new female speaker was 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 waveform table in the WT. Then, the statistics (means and standard deviations) of all the prosodic parameters used in the system were calculated. Instead of retraining the RNN, we simply synthesized all the prosodic parameters by denormalizing the outputs of the RNN using the statistics of the new speaker. This leads to great savings in both the time needed to retrain the RNN and the work required to collect and process a large training set. Fig. 6 displays a typical example of the synthesized pitch contours of syllables for the same input text used in the previous example. It can be seen from the figure that, although the pitch period level of the new speaker is much lower than that of the original male speaker, most of the synthesized pitch contours of syllables still well resemble their original counterparts in both level and shape. Informal listening tests confirmed that the synthetic speech also sounded quite fluent and natural. So, PIG is shown to be very robust.

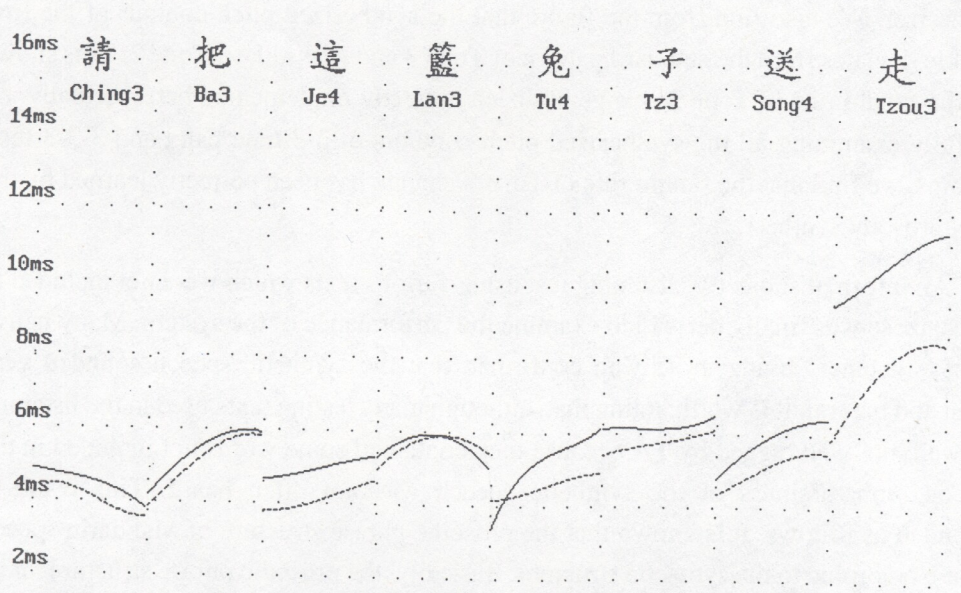

Figure 6 *The original(solid line) and the synthesized(dotted line) pitch contours for the same input text in Fig 5. of the new speaker*

4. Conclusions

We have presented in this paper a high-performance Mandarin TTS system. It is a software package run under both DOS and Windows environments. It can transfer any Chinese text into natural and fluent Mandarin speech in real-time. Several applications of the system are now being developed. 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 intensive retraining. Third, the synthetic speech sounds very natural and fluent. Last, it is a real-time implementation using software only. It will, therefore, be very easy to develop related applications.

5. Acknowledgments

This work was supported by the National Science Council, ROC, under contract NSC84-2213-E009-097. The authors want to thank the Telecommunication Laborato-

ries, MOTC, ROC, for supporting work on the speech database. We also want to thank Academia Sinica for supporting work on the lexicon.

References

Carlson, R. and Granstrom, B., "A text-to-speech system based entirely on rules," *Proc. ICASSP*, 1979, pp.686-688.

Chang, L.L. *et al.*, "Part of Speech(POS) Analysis on Chinese Language," Technical Report, the Institute of Information Science, Academia Sinica, ROC. 1989.

Chen, K.J., "The Indentification of Thematic Roles in Parsing Mandarin Chinese," Proceedings of ROCLING II, 1989, pp.121-146.

Chen, S.H., S.H. Hwang and Y.R. Wang, "An RNN-Based Prosodic Information Generation for Mandarin Text-to-Speech," Submitted to IEEE Trans. on Speech and Audio Signal Processing, 1996.

Klatt, D.H., "Linguistic uses of segmental duration in English: Acoustic and Perceptual Evidence," *J. Acoust. Soc. Am.*, Vol.59, 1976, pp.1208-1221.

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

Lee, Lin-Shan, 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.

Mikuni, I. and K. Ohta ,"Phoneme based text-to-speech synthesis system," *Proc. ICASS*P, 1986, pp.2435-2437.

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

Moulines, E. and F. Charpentier, "Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones," *Speech Communication*, Vol. 9, Dec 1990, pp.453-467.

Nakajima, S. and H. Hamada, "Automatic generation of synthesis units based on context oriented clustering," *Proc. ICASSP*, 1988, pp.659-662.

Sagisaka, Y., "On the Prediction of Global F0 Shape for Japanese Text-to-Speech," *ICASSP*, 1990, pp.325-328.

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

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

Wang, W., W. Campbell, N. Iwahashi and Y. Sagasaka, "Tree-Based Unit Selection for English Speech Synthesis," *ICASSP'93*, Vol.2, 1993, pp.191-194.