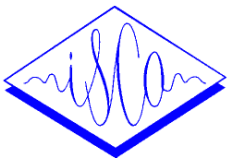

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

An NN-based Approach to Prosody Generation for English Word Spelling in English-Chinese Bilingual TTS

Wei-Chih Kuo, Yih-Ru Wang, Hung-Mao Lu, and Sin-Horng Chen

Department of Communication Engineering, Chiao Tung University, Hsinchu.
cm87532@cm.nctu.edu.tw and schen@cc.nctu.edu.tw

Abstract

In this paper, an RNN-MLP-based scheme to generate proper prosodic information for spelling English words embedded in Chinese text background is proposed. It is extended from the RNN prosody synthesis scheme of an existing Mandarin TTS by adding four MLPs to follow the RNN. It first treats each English word as a Chinese word and uses the RNN to generate eight prosodic parameters for each alphabet of the word. It then uses these four MLPs to refine these prosodic parameters. Experimental results showed that the proposed RNN-MLP scheme led to 36.3, 37.3, 11.6, and 29.1% reductions in RMSE for the synthesized alphabet duration, log-energy level, pitch contour, and pause duration, respectively, over the scheme using the RNN only.

1. Introduction

An English-Mandarin bilingual text-to-speech (TTS) system, which can generate natural synthesized speech for input texts mixed with English words and Chinese text, is important for Chinese societies. Two pronunciation styles of English words are needed to be developed. One is to spell a word alphabet-by-alphabet and another is to read it according to the phonetic symbol string provided by a lexicon. The formal is suitable for the pronunciations of words like “IBM”, “HP” and “TCP/IP”, and the latter is suitable for words such as “UNIX”, “EXCEL” and “apple”. The expansion of an existing Mandarin TTS system to equip with the first style of English word pronunciation is much easier than that for the second English word pronunciation style. This is mainly owing that only a small effort is needed to increase the acoustic inventory by adding waveform templates of 26 English alphabets for the first case, while a labor-intensive work is needed to build a complete wavetable of all possible diphones for the second case. Besides, the synthesis of prosody for the first English word pronunciation style looks much easier. Owing to its simplicity, we therefore consider the first case only in this preliminary study of expanding our Mandarin TTS system [1] to build an English-Mandarin bilingual TTS system, and concentrate all our efforts on the problem of generating proper prosodic information for English word in order to spell it smoothly in the background of Mandarin speech.

The organization of the paper is stated as follows. Section 2 presents the proposed method of expanding the recurrent neural network- (RNN-) based prosody synthesis scheme of our Mandarin TTS system [1] to increase the function of generating proper prosodic information for spelling English words. Performance of the proposed method is evaluated by experiments discussed in Section 3. Some conclusions are given in the last section.

2. The Proposed RNN-MLP Method

Fig. 1 is a block diagram of the proposed RNN-MLP method of generating prosodic information for spelling English words embedded in background Chinese text. It adds four additional three-layer multi-layer perceptrons (MLPs) to follow the four-layer RNN prosodic information generator of the Mandarin TTS system developed previously in Chiao Tung University [1]. In operation, it first treats each English word as a Chinese word and uses the RNN to generate a set of prosodic parameters for each alphabet of the word. This aims at making the synthesized prosodic information match with that of the background Mandarin speech. Then, the four MLPs are employed to refine these prosodic parameters for compensating the mismatch on the prosody pronunciations between the English word and the substituting Chinese word. In the following subsections, we briefly review the function of the RNN and discuss the way of generating the prosodic information for spelling English words in detail.

2.1. The RNN prosodic information generator

The RNN is originally designed to generate the prosodic information for pure Chinese text. It operates in a character-synchronized mode. It accepts two types of inputs extracted from the context of the current character. One is a set of word-level linguistic features including the lengths and the part-of-speeches (POSs) of the current and following words, and an indicator showing the type of punctuation mark (PM) located after the current word. The other is a set of syllable-level features including the tones and the types of initials and finals of the current and following words, and an indicator showing the location of the current syllable in the current word. It generates eight outputs including four parameters representing the pitch contour, one parameter representing the energy level, two parameters representing, respectively, the initial and final durations, and one parameter representing the pause duration following the current syllable.

2.2. The method of prosody generation for English words

To generate the prosodic information for spelling an English word, we first treat the word as a Chinese word and use the RNN to generate a set of prosodic parameters for each of its constituent alphabets. In this approach, there is no problem on finding the word-level input features, but some troubles will be encountered at preparing the syllable-level input features. This mainly results from the mismatches between the phonetic structures of sounds of some English alphabets and that of Mandarin syllables. All Chinese syllables have very regular phonetic structure. It can be firstly decomposed into a base-syllable and a tone. The base-syllable part can be further divided into an optional initial and a final. An initial comprises a single consonant. A final consists of an optional medial, a

vowel nucleus and an optional nasal ending. Obviously, the six alphabets, including “F”, “H”, “L”, “S”, “W” and “X”, are not matched with the initial-final structure of Mandarin base-syllable. All other alphabets are matched or roughly matched with the initial-final structure. The RNN was originally designed and trained to accept features of syllables with the initial-final structure. For making the RNN operable for English alphabets, we assign three pseudo codes of initial type, final type and tone for each of 26 English alphabets as shown in Table 1. It is noted that the two alphabets, “L” and “W”, are regarded as bisyllabic alphabets. By this arrangement, we can generate a set of eight prosodic parameters for each alphabet of an English word by the RNN. We can also expect that the synthesized prosodic information of the English word matches well with the global trend of the prosodic information of the background Mandarin speech generated by the same RNN.

The prosodic information of each English alphabet is then further processed to compensate the mismatch on the prosody pronunciations between the English word and the substituting Chinese word. as well as the effect of possible improper settings at the syllable-level input of the RNN due to the phonetic structure mismatches discussed above. The processing is performed separately for the four subsets of prosodic parameters: four pitch parameters, one alphabet duration, one inter-alphabet/syllable pause duration, and one log-energy level. Four MLPs are employed to realize the parameter-refinement processing. All four MLPs have the same three-layer structure. They are designed to map from the distorted parameters to the correct ones with the help of some input linguistic features extracted from the context of the English word. These additional input features include the word length, the position of the current alphabet in the word, the identity of the current alphabet, pitch means and log-energy levels of the two syllables before and after the word, the broad type of initial in the syllable following the word, and two indicator showing whether there are PMs before and after the word. These four MLPs can be trained using a real-speech database containing utterances of mixed English-Chinese texts.

3. Simulations

Performance of the proposed RNN-MLP method of prosodic information synthesis for spelling English words embedded in background Chinese text was examined by simulations. An English-Mandarin bilingual speech database was used in the test. The database consisted of 539 sentential utterances and (1) is a typical example.

(1) Jie1 Shia4 Lai2 Rang4 Uo3 Men5 Kan4 Kan4 Iou2 **CNN** Suo3 Ti2 Gong1 De5 Bau4 Dau4。

接下來讓我們看看 CNN 所提供的報導。

(Then, let's take a look at the CNN report.)

All utterances were generated by a single female speaker. They were all spoken naturally at a speed of 3.5 syllables/s. There are, in total, 13540 characters including 1872 English alphabets and 11668 Chinese characters. The database was divided into two parts: a training set and an open test set. These two sets consisted of 1485 and 387 English alphabets, respectively.

All speech signals were digitally recorded at a 20-kHz sampling rate. They were manually segmented into syllable sequences. Further preprocessings were then performed to find all initial-final boundaries and to detect energy and pitch contours. Eight prosodic parameters were then extracted for each syllable/alphabet. They included four orthogonally-transformed coefficients of pitch contour, initial and final durations, maximal log-energy level, and the pause duration following the current syllable.

Before testing the proposed method, we calculated the root mean square errors (RMSEs) of two baseline schemes for reference. Scheme-B1 directly used waveform templates of 26 English alphabets in the acoustic inventory without prosody modification. Scheme-B2 used the means of prosodic parameters of 26 English alphabets calculated over the whole database. Table 2 shows the results. It can be seen from Table 2 that Scheme-B2 was better than Scheme-B1. We also find that the RMSE of alphabet duration in Scheem-B1 was very large. This shows that all waveform templates of 26 English alphabets were too long.

We then examined the performance of the proposed RNN-MLP method. Table 3 shows the experimental results. Can be seen from Table 3 that the RNN-MLP method greatly improved the performance of the RNN method. About 36.3, 37.3, 11.6, and 29.1% reductions in RMSE for the synthesized alphabet duration, log-energy level, pitch contour, and pause duration were achieved. By comparing the data shown in Tables 2 and 3, we find that the RNN-MLP method outperformed Scheme-B1 and Scheme-B2.

Table 4 shows the RMSEs of the synthesized prosodic parameters by the RNN for Chinese texts. By comparing the data shown in Tables 3 and 4, we find that all synthesized prosodic parameters except the pause duration were better for English alphabets than for Chinese characters. This may resulted partially from the smaller linguistic variability of English words in the database and partially from the fact that the speaker uttered English words more steadily and carefully.

A typical example of the synthesized prosodic parameters for a mixed English-Chinese sentence is displayed in Figure 2. The text is “現在的年輕人，身上穿的是，USNS 的衣服，IBS 的牛仔褲，腳上則是 NB 的鞋子。”. Clearly, the trajectories of the synthesized prosodic parameters match well with their original counterparts.

4. Conclusions

In this paper, we proposed to extend an existing RNN-based prosody synthesis scheme for Mandarin TTS to an RNN-MLP scheme for English-Mandarin bilingual TTS. Experimental results have confirmed that it is very promising for the generation of prosodic information to spell English word embedded in background Chinese text. Further study to tackle the more difficult task of reading English words embedded in Chinese text will be done in the future.

Acknowledgement

This work was supported by NSC under contract EX-91-E-FA06-4-4.

References

[1] Sin-Horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang, “An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 6, pp. 226-239, May 1998

[2] S. J. Lee, K. C. Kim, H. Yoon, and J. W. Cho, “Application of fully recurrent neural networks for speech recognition,” in *Proc. ICASSP*, 1991, pp. 77-80

Table 1: Assignment of initial, final and tone for 26 English alphabets. (ϕ : empty)Character

	Phonogram	Initial type	Final type	Tone
A	[e]	ϕ	ei	1
B	[bi]	b	i	1
C	[si]	sh	i	1
D	[di]	d	i	1
E	[i]	ϕ	i	1
F	[ef]	ϕ	ei	2
G	[i]	j	iu	1
H	[eʧ]	ϕ	ei	2
I	[aɪ]	ϕ	ai	1
J	[ʤe]	i	e	1
K	[ke]	k	ei	1
L	[el]	ϕ	ei	3
M	[em]	l	o	5
		ϕ	ei	4
N	[en]	ϕ	en	1
O	[o]	ϕ	ou	1
P	[pi]	p	i	1
Q	[kju]	k	iou	1
R	[ar]	ϕ	a	3
S	[es]	ϕ	`e	2
T	[ti]	t	i	1
U	[ju]	ϕ	iou	1
V	[vi]	m	i	1
W	[´dʌblju]	d	ai	1
X	[eks]	l	iou	1
		ϕ	ei	2
Y	[waɪ]	ϕ	uai	1
Z	[zi]	r	i	4

Table 2: The RMSEs of two baseline schemes.

Scheme	alphabet duration (ms)	pause (ms)	pitch (ms/frame)	energy (dB)
B1	470.1	-	0.74	3.89
B2	40.2	124.5	0.59	2.19

Table 3: Experimental results of the RNN method and the proposed RNN-MLP method.

		Alphabet duration	pause (ms)			pitch (ms/frame)	energy (dB)
			Before-word	Intra-word	After-word		
RNN	Training	55.0	130.3	27.1	103.4	0.66	3.35
	Testing	57.6	139.0	26.5	105.3	0.69	3.22
RNN-MLP	Training	31.5	--	18.1	--	0.45	1.70
	Testing	36.7	--	18.8	--	0.61	2.02

單位 (ms)		詞前	詞中	詞尾
RNN	Training	130.3	27.1	103.4
	Testing	139.0	26.5	105.3
RNN-MLP	Training	--	18.1	--
	Testing	--	18.8	--

Table 4: The RMSEs of the synthesized prosodic parameters by the RNN for Chinese texts.

	Training	Testing
pitch (ms/Frame)	0.84	1.06
pause (ms)	23.7	54.5
initial duration (ms)	17.2	18.5
final duration (ms)	33.3	36.7
log-energy level (dB)	3.39	4.17

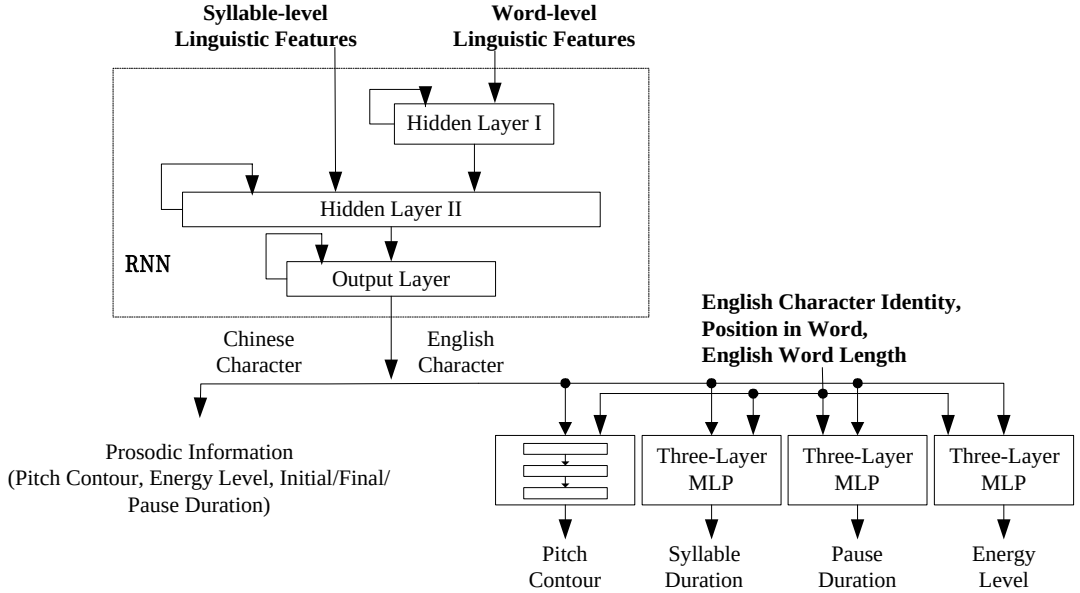

Figure 1: A block diagram of the proposed RNN-MLP method of generating prosodic information for spelling English words embedded in background Chinese text.

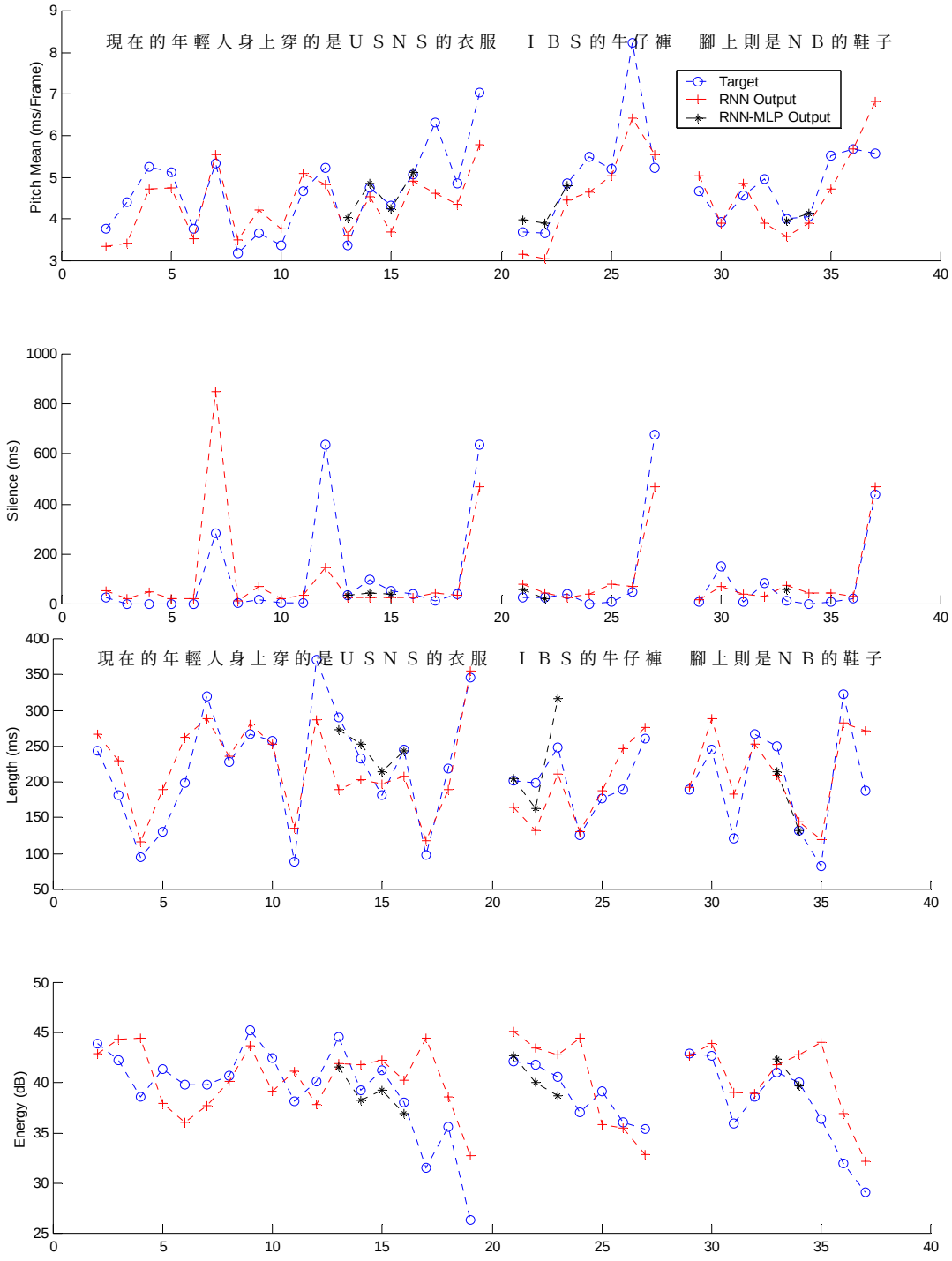

Figure 2: Typical example of prosodic parameter sequence of: (a) the pitch mean, (b) intersyllable pause duration, (c) duration of syllable, and (d) energy level. The text is: “現在的年輕人，身上穿的是，USNS 的衣服，IBS 的牛仔褲，腳上則是 NB 的鞋子。”.