An NN-based Approach to Prosodic Information Generation for Synthesizing English Words Embedded in Chinese Text

Wei-Chih Kuo, Li-Feng Lin, Yih-Ru Wang, and Sin-Horng Chen[1]

Dept. of Communication Engineering, National Chiao Tung University, Taiwan

schen@cc.nctu.edu.tw

ABSTRACT

In this paper, a neural network-based approach to generating proper prosodic information for spelling/reading English words embedded in background Chinese texts is discussed. It expands an existing RNN-based prosodic information generator for Mandarin TTS to an RNN-MLP scheme for Mandarin-English mixed-lingual TTS. It first treats each English word as a Chinese word and uses the RNN, trained for Mandarin TTS, to generate a set of initial prosodic information for each syllable of the English word. It then refines the initial prosodic information by using additional MLPs. The resulting prosodic information is expected to be appropriate for English-word synthesis as well as to match well with that of the background Mandarin speech. Experimental results showed that the proposed RNN-MLP scheme performed very well. For English word spelling/reading, RMSEs of 41.8/78.2 ms, 30.8/26 ms, 0.65/0.45 ms/frame, and 3.06/4.9 dB were achieved in the open tests for the synthesized syllable duration, inter-syllable pause duration, pitch contour, and energy level, respectively. So it is a promising approach.

1. INTRODUCTION

Text-to-speech (TTS) systems have traditionally been designed to deal with a single target language, for instances English [1], Chinese [4,6] or Japanese, etc. This means all components of a TTS system, such as language model, sound inventory and prosodic model, are optimized to the target language. As foreign loanwords and proper names, or passages in a different language occur in a given text, a single-language TTS system must necessarily fail. To solve the problem, multilingual TTS systems are therefore needed to be developed [2,3].

For the Chinese society, a Mandarin-English bilingual TTS system, which can generate natural synthesized speech for mixed Chinese-English texts, is important nowadays. This is mainly owing to the rapid growth in globalization to bring about a steadily growing number of English terms being directly used in the Chinese society. In the past, English words were translated into Chinese words, such as space shuttle (太空梭) and algebra (代數). But, in recent years, mixed English-Chinese texts or speech are very popular in Taiwan especially for the information processing domain. We give two examples:

(1) 我要進 IBM 公司。(I want to join the IBM corporation.)

(2) 送個 email 給我。(Give me an email.)

Besides, it becomes popular that young generations in Taiwan use short English alphabet strings to replace Chinese words as well as to represent some concepts for daily speech communication and for interactive communication through Internet. We list some of them in the following: BPP (白拋拋, white), SDD (水噹噹, very pretty), LKK (老叩叩, very old), etc.

In all applications of using mixed Chinese-English texts, Chinese is always the primary language. The developments of Mandarin-English polyglot TTS systems are very urgent for Chinese societies. In such systems, two styles of pronouncing English words or alphabet strings embedded in a background Chinese text are needed to be developed. One is to spell a word letter-by-letter and another is to read it according to the phonetic symbol string suggested by lexicons or rules. The formal is suitable for the pronunciation of words like “IBM” and of alphabet strings like “LKK” (very old). The latter is suitable for words such as “UNIX” and “Seven-Eleven”. In both cases, it is required not only to synthesize English speech clearly with high intelligibility but also to make its prosody match well with that of the background Mandarin speech.

In this paper we address the issue of expanding an existing Mandarin TTS system [4] to a polyglot one which can properly pronounce English words embedded in background Chinese text. The study is focused on the problem of generating proper prosodic information for English words in order to make their pronunciations match with the background Mandarin speech.

The organization of the paper is stated as follows. Section 2 presents the proposed approach. Performance of the method is evaluated by experiments discussed in Section 3. Some conclusions are given in the last section.

2. THE PROPOSED APPROACH

Figure 1 is a block diagram of the proposed approach. It adds two sets of three-layer multi-layer perceptrons (MLPs) to follow a recurrent neural network (RNN) prosodic information generator developed previously in our Mandarin TTS system [4]. The part of Chinese text is treated as before by solely using the RNN to generate all prosodic information needed for synthesizing its speech. For an English word, the RNN and one of these two MLP sets are used to generate the prosodic information needed for spelling or reading it. Specifically, it first treats the English word as a Chinese word and uses the RNN to generate an initial set of prosodic parameters for each of its constituent syllables. Here it aims at making the initially synthesized prosodic information matched well with that of the background Mandarin speech. Then, it employs the MLP set to refine these initial prosodic parameters to correct the distortions

[1] The corresponding author

caused by the improper RNN input settings due to the mismatches between the phonetic structures of English alphabets/syllables and Mandarin syllables. In the following subsections, we briefly review the function of the RNN and discuss the ways of generating the prosodic information for spelling and for reading an English word in detail.

A. The RNN prosodic information synthesizer

The RNN is originally designed to generate prosodic information for synthesizing a pure Chinese text. It operates with character-synchronized clock. It accepts two types of inputs extracted from the context of the current character. One is a set of word-level linguistic features and another is a set of syllable-level features. 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. The RNN has been confirmed to perform very well for pure Chinese input texts. [4]

B. Prosody generation for spelling an English word

To generate the prosodic information for spelling an English word, we first use the RNN to generate a set of initial prosodic parameters for each of its constituent syllables. For preparing the RNN inputs, we analyze the phonetic structures of 26 English alphabets and compare them with the generic phonetic structure of Mandarin syllables to construct a lookup table showing the mapping of each alphabet to a Mandarin-like syllable or two-syllable sequence. To explain the way of constructing the table more clearly, we briefly introduce the phonetics of Mandarin speech as follows. Fig. 2 shows the generic phonetic structure of Mandarin syllables. Mandarin Chinese is a tonal and syllabic language. Each character is pronounced as a syllable. A syllable is generally composed of two parts: a base-syllable and a tone. There are in total 411 base-syllables, each of which can have up to five different syllabic tones. All base-syllables are in a simple initial-final (or roughly consonant-vowel) structure. The initial, if it exists, consists only of a single consonant. There are in total 22 initials including a null one. The final is composed of three components including an optional preceding medial, a vowel nucleus, and an optional nasal ending. There are in total 39 finals. From above discussions, we find that the phonetic structures of all alphabets, except the six alphabets {F, H, L, S, W, X}, are matched or roughly matched with the generic initial-final structure of Mandarin syllable. We can therefore easily assign a Mandarin-like syllable to each of them. For these six mismatching alphabets, we process them in two ways. For the two alphabets, “L” and “W”, we regard them as bi-syllabic alphabets and assign a pair of Mandarin-like syllables to each of them. Note that, although “L” is actually a monosyllabic alphabet, it is commonly pronounced as a bi-syllabic alphabet in Taiwan. For the four alphabets of {F, H, S, X}, we neglect their ending consonants and roughly assign a Mandarin-like syllable to each of them. Table 1 shows the mappings of all 26 English alphabets to Mandarin-like syllables with form of initial-final/tone. With the help of Table 1, we can extract all word-level and syllable-level linguistic features for each syllable of the English word and generate a set of initial prosodic features by the Mandarin RNN prosody generator.

We then refine the initial prosodic information of the English word by four additional MLPs using some new input linguistic features. The processing is performed separately for the four subsets of initial prosodic parameters: four pitch parameters, one syllable duration formed by combining the initial and final durations, one inter-syllable pause duration, and one log-energy level. Parameters in each subset are processed by an MLP. Each MLP is designed to map from the distorted initial prosodic parameters to the correct ones with the help of some new input linguistic features extracted from the context of the current English syllable. These additional linguistic features include the length (i.e., number of syllables) of the English word, the position of the current syllable 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 initial type of the following syllable, two indicators showing whether there are PMs before and after the current word, and special Mandarin word or POS preceding/following the current word. These four MLPs can be trained by the back-propagation (BP) algorithm using a real-speech database containing utterances of mixed English-Chinese texts.

C. Prosody generation for reading an English word

The process of generating prosodic information for reading an English word is similar to that for spelling an English word. We first treat the English word as a Chinese word and use the RNN to generate an initial set of prosodic parameters for each of its constituent syllables. There is no problem about finding the word-level input features if we take the syllable count as the word length. Similar problem of mismatch in syllables’ phonetic structure is encountered on preparing the syllable-level input features. We solve the problem by assigning to each English syllable a set of three pseudo codes of initial type, final type and tone. The initial type is assigned according to the leading consonant. The final type is assigned according to the vowel nucleus. And the tone is manually assigned.

After obtaining the initial prosodic parameters of all syllables, we then use a set of four MLPs to refine them. The refinement process is the same as that of the English word-spelling case except that different additional input linguistic features are used. For the MLP of pitch level synthesis, the position of accent syllable in the word and the vowel type of the accent syllable are used. For the MLP of energy level synthesis, the vowel type of the accent syllable is used. For the MLP of word duration synthesis, the total numbers of syllables and phonemes in the word are used.

3. Simulations

Performance of the new method of prosodic information synthesis for English words embedded in a Chinese text was examined through simulations. Two English-Mandarin bilingual speech databases were used in the test. They were used in the prosody syntheses for English word spelling and reading, respectively. They are referred to as DB-spelling and DB-reading. The DB-spelling consisted of 539 sentential utterances. The DB-reading consisted of 845 sentential utterances.

All utterances were generated by a single female speaker. Utterances in the DB-spelling database were read slowly in a rate of 1 to 4 (ave. 3) syllables/s. The database contains, in total,

13540 characters including 1872 English characters and 11668 Chinese characters. It was divided into two parts: a training set and an open test set. These two sets consisted of 1485 and 387 English characters, respectively. Utterances in the DB-reading database were spoken naturally at a speed of 2 to 5.5 (ave. 4.5) syllables/s. The database contains 923 English words and 15908 Chinese characters. The syllable number of English word ranged from 1 to 6. The database was also divided into two parts: a training set and an open test set. These two sets consisted of 740 and 183 English words, respectively.

All speech signals were digitally recorded using a 20-kHz sampling rate. They were then manually segmented into syllable sequences. Further preprocessings were then performed to find all initial-final boundaries and to detect energy and pitch contours. The eight prosodic parameters, including four orthogonally transformed coefficients of pitch contour, maximal energy level, initial and final durations, and pause duration, to be synthesized for each syllable/character were then extracted. All texts were also manually processed to firstly be segmented into word sequences and then tagged to obtain part-of-speech (POS) sequences. Besides, syllable sequences were obtained by table lookup using a Chinese lexicon and an English lexicon.

A. The case of English word spelling

Table 2 shows the RMSEs of the prosodic parameters synthesized by the RNN scheme and by the RNN-MLP scheme. It can be found from Table 2 that the RNN-MLP scheme outperformed significantly the RNN scheme. RMSEs of 39.4 (41.8) ms, 25.9 (30.8) ms, 0.56 (0.65) ms/frame, and 2.16 (3.06) dB were achieved in the closed (open) test by the RNN-MLP scheme for the synthesized syllable duration, inter-syllable pause duration, pitch contour, and energy level, respectively. This confirmed that the RNN-MLP scheme was effective on compensating the effect of improper RNN input settings as discussed in Section 2. A typical example of the synthesized syllable pitch mean is displayed in Figure 3. The text is "Ni3 Hui2 Gu4 Hou4，Zhi3 Yao4 Ba3 Fu4 Dang3 Ming2 Shi4，d o c De5 Dang3，Shan1 Chu2 Diao4 Jiu4 Ke3 Yi3 Le5。" (你回去後，只要把副檔名是，d o c的檔，刪除掉就可以了。). As shown in the figure that the trajectory of the synthesized syllable pitch mean matches well with their original counterpart. So the proposed method is a promising one.

B. The case of English word reading

Table 3 shows the experimental results for the English word-reading case. It can be found from Table 3 that the RNN-MLP scheme outperformed significantly the RNN scheme. RMSEs of 71.4 (78.2) ms, 25.7 (26) ms, 0.45 (0.45) ms/frame, and 3.41 (4.90) dB were achieved in the closed (open) test by the RNN-MLP scheme for the synthesized syllable duration, inter-syllable pause duration, pitch contour, and energy level, respectively. This confirmed the effectiveness of the RNN-MLP scheme. A typical example of the synthesized prosodic parameters is displayed in Figure 4. The text is "Zhe4 Zhen1 Zhu1 Xiang4 Lian4，Dai4 Zai4 Jing3 Shang4，**ennoble** Le5 Ta1 Zheng3 Ge5 Ren2 De5 Qi4 Zhi2，Gan3 Jue2 Fei1 Chang2 De5 Yong1 Rong2 Hua2 Qui4。" (這珍珠項鍊，戴在頸上，ennoble了他整個人的氣質，感覺非常的雍容華貴。). Can be seen from the figure that the use of MLPs greatly improves the RNN synthesized pitch means for all three English syllables.

4. Conclusions

In this paper, we discussed an approach to extending an existing RNN-based prosody synthesis scheme for Mandarin TTS to an RNN-MLP scheme for Mandarin-English mix-lingual TTS. Experimental results have confirmed that it is promising for the generation of prosodic information to spell and read English words embedded in background Chinese texts. Further study to implement the complete Mandarin-English mix-lingual TTS system will be done in the future.

Acknowledgement

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

REFERENCES

[1] D. H. Klatt, "Review of Text-to-Speech Conversion for English," J. Acoust. Soc. Am.,vol.82, no.3, pp.137-181, Sept. 1987.

[2] R. Sproat, Multilingual Text-to-Speech Synthesis: The Bell Labs Approach, Kluwer Academic Publishers, 1998.

[3] A. Bonafonte, I. Esquerra, A. Febrer and F. Vallverdu, "A Bilingual Text-to-Speech System in Spanish and Catalan," in the proc. of Eurospeech97, pp.2455-2458, Rhodes, Greece, 1997.

[4] 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.

[5] Y. R. Chao, "A Grammar of Spoken Chinese," Berkeley, CA: University of California Press, 1968.

[6] Lin-shan Lee, Chiu-yu Tseng, Ming Ouh-young, "The Synthesis Rules in a Chinese Text-to-speech System," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 9, September 1989, pp. 1309 – 1319.

Table 1: Assignment of initial type, final type and tone for 26 English alphabets (ϕ: null initial)

Alpha-bet	Phono-gram	*Initial*	*Final*	Tone	Alpha-bet	Phono-gram	*Initial*	*Final*	Tone
A	[e]	ϕ	ei	1	N	[ɛn]	ϕ	en	1
B	[bi]	b	i	1	O	[o]	ϕ	ou	1
C	[si]	x	i	1	P	[pi]	p	i	1
D	[di]	d	i	1	Q	[kju]	k	iou	1
E	[i]	ϕ	i	1	R	[ar]	ϕ	a	3
F	[ɛf]	ϕ	ei	2	***S***	[ɛs]	ϕ	ei	2
G	[ʤi]	j	iu	1	T	[ti]	t	i	1
H	[eʧ]	ϕ	ei	2	U	[ju]	ϕ	iou	1
I	[aɪ]	ϕ	ai	1	V	[vi]	m	i	1
J	[ʤe]	j	iue	1	***W***	[ˋdʌblju]	d	ai	1
K	[ke]	k	ei	1			l	iou	1
L	[ɛl]	ϕ	ei	3	***X***	[ɛks]	ϕ	ei	2
		l	ou	5	Y	[waɪ]	ϕ	uai	1
M	[ɛm]	ϕ	ei	4	Z	[zi]	l	i	4

Table 2: RMSEs of the synthesized prosodic parameters for the English word spelling case.

	Inside				Outside			
	Duration (ms)	Pause (ms)	Pitch (ms/frame)	Energy (dB)	Duration (ms)	Pause (ms)	Pitch (ms/frame)	Energy (dB)
RNN	58.1	33.9	0.93	5.17	60.7	33.6	1.02	4.88
RNN -MLP	39.4	25.9	0.56	2.16	41.8	30.8	0.65	3.06

Table 3: RMSEs of the synthesized prosodic parameters for the English word reading case

	Inside				Outside			
	Duration (ms)	Pause (ms)	Pitch (ms/frame)	Energy (dB)	Duration (ms)	Pause (ms)	Pitch (ms/frame)	Energy (dB)
RNN	141.7	28.3	1.24	5.27	154.1	28.2	1.21	5.26
RNN -MLP	71.4	25.7	0.45	3.41	78.2	26.0	0.45	4.90

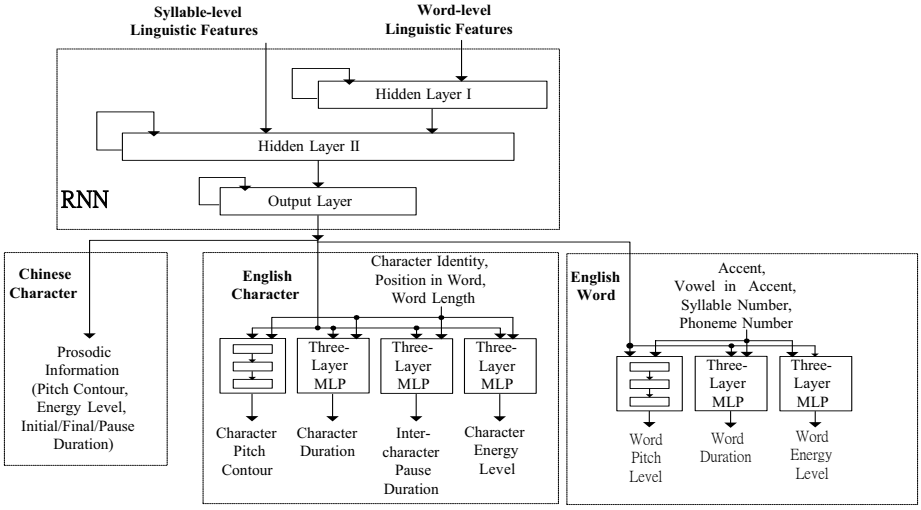

Figure 1: A block diagram of the proposed approach to prosodic information generation for Mandarin-English mix-lingual TTS.

Tone			
Initial	Final		
(Consonant)	(Medial)	Vowel	(Nasal ending)

Figure 2: The generic phonetic structure of Mandarin syllables.

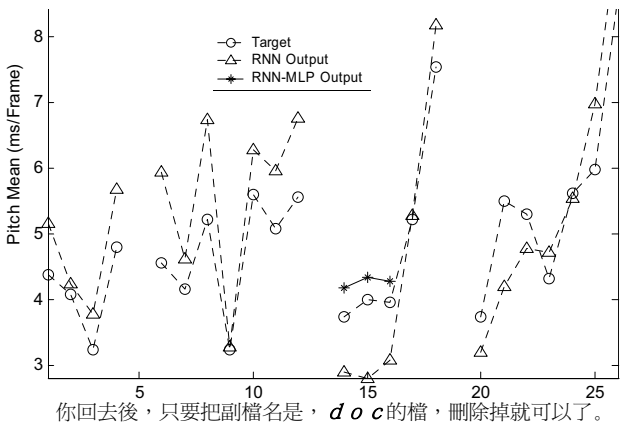

Figure 3: A typical example of synthesized syllable pitch mean for the English word-spelling case. The text is: "你回去後，只要把副檔名是，ｄｏｃ的檔，刪除掉就可以了。".

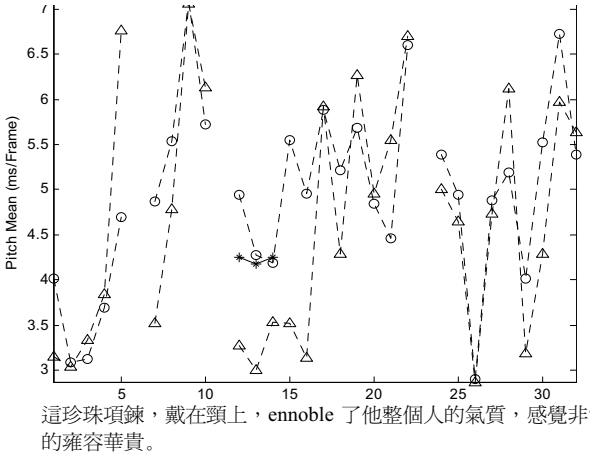

Figure 4: A typical example of synthesized syllable pitch mean for the English word-reading case. The text is: "這珍珠項鍊，戴在頸上，ennoble 了他整個人的氣質，感覺非常的雍容華貴。".