Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Duration modeling for English letters embedded in Chinese speech

Wen-Hsing Lai

National Kaohsiung First University of Science and Technology, No. 2, Jhuoyue Rd., Nanzih
District, 811 Kaohsiung, Taiwan
lwh@ccms.nkfust.edu.tw

A review of existing multilingual TTS (Text-To-Speech) systems shows that the secondary language inserted into the primary language sounds more like isolated individual words in an alien language environment and not congruous with the primary language's prosody. Since the letter-by-letter spelling of English words or acronyms appears in Chinese speech quite often, a duration modeling approach for English letters embedded in Chinese speech is proposed to make the English congruous with the primary language's tempo. It takes several major factors as additive factors and estimates all model parameters by an EM (expectation-maximization) algorithm. Experimental results showed that the standard deviation of the duration from the test set was greatly reduced from 59.82 to 9.37 ms by the duration modeling while eliminating effects from factors. The root mean squared error between the original and estimated durations was 9.35 ms for the open tests. Experimental results have confirmed its effectiveness on isolating several main factors that seriously affects the duration. Moreover, the estimated value of the factors agreed well to our prior linguistic knowledge. Besides, the hidden state labels produced by the EM algorithm were linguistically meaningful.

1 Introduction

As global communication and multiethnic societies grow, the demand for multilingual capability increases. Code-switching (alternation between two or more languages) is a quite common phenomenon in many multilingual societies. In Taiwan, at least three languages – Mandarin, Taiwanese and English – are frequently mixed and spoken in daily conversations and writings. Polyglot TTS systems which can process mixed-language input and generate mixed speech with coherent prosody are highly demanded. A review of existing multilingual TTS systems shows that the secondary language inserted into the primary language sounds more like isolated individual words in an alien language environment and not congruous with the primary language's prosody. Therefore, we would like to analyze the prosody of mixed language and explore novel prosodic modeling approach to solve the problems we encountered in existing multilingual TTS. Besides, duration modeling is also important in automatic speech recognition (ASR) [1]. In ASR, state duration models are usually constructed to assist in the HMM-based speech recognition.

Many methods have been proposed to solve the multilingual problem [2] - [8]. For example, text normalization and text processing are performed separately, but treated together in prosody prediction, which ensures a coherent intonation in mixed language situation [2]. Some rely on Phoneme Mapping algorithm to make foreign phoneme sequences pronounceable [3] - [4]. Others include treating each English word as a Chinese word and using an RNN-MLP-based scheme to generate proper prosodic information for spelling English words embedded in Chinese text background [5]. A statistical model using EM [9] via considering some major factors is proposed in this paper. The goal is to separate the effects of the factors so as to better understanding the mechanism of duration generation in Mandarin and English mixed speech.

In processing mixed Mandarin and English, there are two kinds of pronunciation styles of English words. One is to spell a word letter-by letter and the other is to read it according to its phonetic symbol. Since the spelling way of acronyms like "IBM", "NBA" and "ISDN" often appears and only a small effort is needed to read English letters in Chinese TTS by just adding 26 waveform templates, the author therefore firstly concentrate on the problem of generating proper prosodic information for English letters embedded in Chinese Speech.

The paper is organized as follows. Section 2 discusses the proposed duration model in details. Section 3 describes the experimental results. Some conclusions and possible future works are given in the last section.

2 The duration model

Because of the mismatches between the phonetic structures of English letters and Mandarin, forcing the English letter to match with the Chinese initial-final structure doesn't make sense. Therefore, in this study, both Mandarin syllables and English letters will be used as the basic units, and lexical tone, base-syllable/letter and prosodic state are chosen as the relevant factors.

Mandarin Chinese is a tonal and syllable-based language. We therefore consider the 5 tones as a factor. For simplifying the modeling, except the tones of Chinese, one tone is left for English letters, though English is not tone language.

The prosodic state is conceptually defined as the state in a prosodic phrase. In continuous speech, speakers tend to group words into phrases whose boundaries are marked by durational and intonational cues. Those phrases are usually referred to as prosodic phrases. Many phonological rules limit their operation within prosodic phrases. While it is generally agreed that the prosodic structure of an utterance has some relationship with its syntactic structure, the two are not isomorphic. In the model, the prosodic state is used as a substitute for high-level linguistic information, like a word, phrase or syntactic boundaries.

Due to the fact that the prosodic state is not explicitly given, it has been treated as a hidden variable and expectation-maximization (EM) algorithms has been applied to estimate all the parameters based on training data. A by-product of the EM algorithm is the determination of the hidden prosodic states of all the units in the training set. This is an additional advantage because prosodic labeling has recently become an interesting research topic [10]. From the sequence of prosodic states, some high-level linguistic phenomenon could be observed, like the possible prosodic phrase boundaries.

By considering the factors, the additive duration model can be expressed by

$$Z_n = X_n + \gamma_{t_n} + \gamma_{y_n} + \gamma_{j_n}\,, \tag{1}$$

where Z_n and X_n are, respectively, the observed duration and the normalized duration of the *n*th unit; γ is factor; t_n, y_n and j_n represent respectively the lexical tone, prosodic state, and base-syllable/letter of the *n*th modeling unit; and X_n is modeled as a normal distribution with mean μ and variance ν.

To illustrate the EM algorithm, an auxiliary function is firstly defined in the expectation step as

$$Q(\lambda,\bar{\lambda}) = \sum_{n=1}^{N}\sum_{y_n=1}^{Y} p(y_n \mid Z_n,\lambda)\log p(Z_n,y_n \mid \bar{\lambda}), \quad (2)$$

where *N* is the total number of training samples, *Y* is the total number of prosodic states, $p(y_n \mid Z_n,\lambda)$ and $p(Z_n,y_n \middle| \bar{\lambda})$ are conditional probabilities which can be derived from the assumed model given in Eq.(1), and $\lambda = \{\mu,\nu,\gamma_t,\gamma_y,\gamma_j\}$ is the set of parameters to be estimated. Then, sequential optimizations of these parameters can be performed in the maximization step (M-step). A drawback of the above EM algorithm is that the non-uniqueness of the solution because of the use of additive factors. This is obvious because, if we scale up an factor and scale down another, the same objective value will be reached.

To cure the drawback, each optimization procedure is modified in the M-step to a constrained optimization via introducing a global duration constraint. The auxiliary function then changes to

$$\begin{aligned} Q(\bar{\lambda},\lambda) &= \sum_{n=1}^{N}\sum_{y_n=1}^{Y} p(y_n \mid Z_n,\bar{\lambda})\log p(Z_n,y_n \mid \lambda) \\ &+ \eta\Big(\sum_{n=1}^{N}(\mu+\gamma_{t_n}+\gamma_{y_n}+\gamma_{j_n}) - N\mu_z\Big) \end{aligned}, \quad (3)$$

where μ_z is the average of Z_n and η is a Lagrange multiplier. The constrained optimization is finally solved by the Newton-Raphson method.

To execute the EM algorithm, initializations of these parameters are needed. This can be done by estimating each parameter independently. After initialization, all parameters are sequentially updated in each iterative step. Iterations are continued until a convergence is reached. The prosodic state can finally be assigned by

$$y_n = \max_{y} p(y \mid Z_n,\lambda) \quad (4)$$

We then consider the effect of the three Tone 3 patterns of falling-rising (full tone), middle-rising (sandhi tone) and low-falling (half tone). These three patterns are simply denoted as Tone 3_f, Tone 3_s and Tone 3_h. The EM algorithm is then modified for parameter estimation. In initialization, we first assign all lexical Tone 3 to Tone 3_s when they precedes other lexical Tone 3, and then use VQ to divide all others lexical Tone 3 into two clusters of Tone 3_f and Tone 3_h. Besides, at the end of each iteration, syllables with lexical Tone 3 are re-assigned to one of these three patterns by

$$t_n^* = \arg\max_{t_n} p(t_n \mid Z_n,\lambda), \quad (5)$$

for t_n =3_f, 3_s, 3_h, where $p(t_n \mid Z_n,\lambda)$ is the conditional probability of a Tone 3 pattern.

3 Experimental Results

An English-Mandarin bilingual speech database was used in the experiment. All the English words are in spelling style, that is, are read letter by letter, like acronyms "WTO", "DDT" and "CPU". The database consists of 539 sentential utterances. All utterances were generated by a female speaker, who is a native speaker of Mandarin Chinese. They were all spoken naturally at an average speed of 3.5 syllables/s. There are, in total, 13540 characters including 1872 English letters and 11668 Chinese characters. The longest English word contains 6 letters. The shortest contains 1 letter. The average length of English word is 2.996 letters/word. The database was divided into two parts: a training set and an open test set. Training set contains 8607 Chinese syllables and 1413 English letters. Test set contains 3061 Chinese syllables and 459 English letters. All speech signals were digitally recorded at a 20-kHz sampling rate. They were manually segmented into Chinese syllable and English letter sequences.

First, initial values of all parameters were independently estimated from the training set. Prosodic states were labeled by a vector quantizer with 8 codewords. Then, the EM algorithm was performed to update all parameters until convergence. As shown in Table 1, the standard deviations of the observed duration were 64.55 and 59.83 ms for the training and testing data sets. The resulting standard deviations of the normalized duration reduced to 8.97 and 9.37 ms for the closed and open tests. The standard deviations were greatly reduced after compensating the effects of the factors. The corresponding root mean squared errors between the original and estimated durations were 8.95 and 9.37 ms for the closed and open tests.

Tables 2 and 3 show the values of the lexical tone and prosodic state factors, respectively. For simplifying the problem, except the 7 tones of Chinese, one tone is left for English letters, though English is not tone language. Can be seen from Table 2 that Tone 5 has relatively smaller value so as to make the associated syllable duration much shorter than those of the other tones. This agrees to the prior linguistic knowledge. Besides, when it's English letter, the value 21.74 is the largest. It may be because it's an alien language for the speaker, and the speaker will automatically slow down when pronouncing English letters. Besides, many English letters are not monosyllable.

Table 3 shows the values of 8 prosodic state factors. It can be found from Table 3 that State 7 has the largest value while State 0 has the smallest. Fig. 1 shows an example of prosodic state labeling by the EM training algorithm. From Fig. 1, we find that states with larger index, that is, with larger value, usually associates with the ending syllables of sentences or phrases and states with smaller index, that is, with smaller value, always associate with intermediate syllables of polysyllabic words. The finding also complies with the prior knowledge of the lengthening effect for the last syllable of a prosodic phrase or sentence.

Table 4 is the estimated values for 26 English letters in the duration model. Letters with more phonemes like {W, X} are obviously much more longer. Single vowel or single vowel with a very short consonant like {E, O, B, D} are shorter.

Training set		Testing set	
mean	standard deviation	mean	standard deviation
216.08	64.55	213.00	59.83

(a)

Training set			Testing set		
mean	standard deviation	RMSE	mean	standard deviation	RMSE
213.83	8.97	8.95	213.51	9.37	9.37

(b)

Table 1: Statistics of (a) the observed durations, and (b) the normalized durations obtained in the additive models with 8 prosodic states. (units: ms)

Tone	1	2	3_f	4
	3.11	10.79	11.95	0.74
Tone	5	3_s	3_h	E
	-40.77	-14.70	-38.88	21.74

Table 2: The estimated values for tone factor in the duration model. E is for English letters.

State	0	1	2	3
	-75.67	-46.49	-29.40	-15.02
State	4	5	6	7
	5.75	26.72	61.97	124.05

Table 3: The estimated values for prosodic state factor in the duration model.

在3美5國7，
並2稱6五2大5職1業5運5動0的2是7，
N2F7L7－
美4式5足5球7、
M5L3B4－
大3聯2盟0棒4球7、
N4B3A6－
職4業1籃5球7、
N6H4L6－
冰1上3曲4棍5球7以1及6N3C5A0A6－
大5學5籃2球6。

Figure 1: An example of prosodic state labeling. The number indicates the state number of each Mandarin syllable or English letter.

Letter	A	B	C	D	E
	-17.51	-50.32	32.69	-44.92	-37.86
Letter	F	G	H	I	J
	-42.93	-3.40	3.31	-13.35	41.29
Letter	K	L	M	N	O
	-8.80	54.92	24.17	5.20	-34.58
Letter	P	Q	R	S	T
	3.79	81.16	17.34	44.68	-1.69
Letter	U	V	W	X	Y
	23.97	6.18	149.45	104.66	45.55
Letter	Z				
	102.86				

Table 4: The estimated values for 26 English letters in the duration model.

5 Conclusions and future works

The paper presents duration modeling approach for mixed language. Experimental results have confirmed its effectiveness on isolating several main factors that seriously affects the duration. Aside from greatly reducing the standard deviation of the modeled duration, the estimated factors conformed well to the prior linguistic knowledge. Besides, the prosodic-state labels produced by the EM algorithm were linguistically meaningful. So it is a promising duration modeling approach for English letters embedded in Mandarin speech.

Further studies to tackle the more difficult task of reading English words embedded in Chinese text will be done in the future. Besides, some future works are worthwhile doing. Firstly, the duration model can be further improved via considering more factors. This needs the help of a more sophisticated text analyzer. Secondly, the applications of the model to both ASR and TTS are worth further studying. Lastly, the approach may be extended to the modeling of other prosodic features such as pitch, energy, and pause duration.

Acknowledgments

This work was supported by NSC, Taiwan under Contract NSC 96-2628-E-327-032-MY2. The author thanks Dept. of Communication Engineering, National Chiao Tung University, Taiwan for supplying the speech databases.

References

[1] Anastasios Anastasakos, Richard Schwartz, Han Shu, "Duration Modeling in Large Vocabulary Speech Recognition," *ICASSP*, vol. 1, pp. 628 – 631, 1995

[2] Haiping Li, Fangxin Chen, Li Qin Shen, Xi Jun Ma, "Trainable Cantonese/English dual language speech synthesis system," *Proc. of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, Volume 1, Page(s):I-508 - I-511, 6-10 April 2003

[3] Badino Leonardo, Barolo Claudia, Quazza Silvia, "A General Approach to TTS Reading of Mixed-Language Texts", *Proc. of 5th ISCA Tutorial and Research Workshop on Speech Synthesis*, Pittsburgh, PA, 2004

[4] Badino Leonardo, Barolo Claudia, Quazza Silvia, *"*Language Independent Phoneme Mapping For Foreign TTS,*" Proc. of 5th ISCA Tutorial and Research Workshop on Speech Synthesis*, Pittsburgh, PA, 2004

[5] Wei-Chih Kuo, Yih-Ru Wang, Hung-Mao Lu, and Sin-Horng Chen, "An NN-based Approach to Prosody Generation for English Word Spelling in English-Chinese Bilingual TTS," *Proc. of International Conf. on Chinese Spoken Language Processing* 2002, Taipei, ROC, pp. 29-32, Oct. 2002

[6] Helen M. Meng, Chi Kin Keung, Kai Chung Siu, Tien Ying Fung and P. C. Ching, "CU VOCAL: corpus-based syllable concatenation for Chinese speech synthesis across domains and dialects," In *ICSLP2002*, 2373-2376

[7] Chung-Hsien Wu, Yu-Hsien Chiu, Chi-Jiun Shia, Chun-Yu Lin, "Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs," *IEEE Transactions on Audio, Speech and Langauge Processing*, Volume 14, Issue 1, Jan. 2006 Page(s): 266 – 276

[8] Dau-cheng Lyu, Ren-yuan Lyu, Yuang-chin Chiang, Chun-nan Hsu, "Speech Recognition on Code-Switching Among the Chinese Dialects" *ICASSP*2006, Volume 1, 14-19 May 2006, Page(s):I-1105 - I-1108

[9] Sin-Horng Chen, Wen-Hsing Lai, Yih-Ru Wang, "A New Duration Modeling Approach for Mandarin Speech," *IEEE Trans. Speech and Audio processing*, vol. 11, no.4, pp.308-320, July 2003

[10] Colin W. Wightman, Mari Ostendorf, "Automatic Labeling of Prosodic Patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, October 1994, pp. 469 – 481