A NOVEL SYLLABLE DURATION MODELING APPROACH FOR MANDARIN SPEECH

Wen-Hsing Lai[1,2] and Sin-Horng Chen[1]

[1] Dept. of Communication Engineering, National Chiao Tung University, Taiwan

[2] Chunghwa Telecommunication Laboratories, Taiwan

ABSTRACT

In this paper, a novel syllable duration modeling approach for Mandarin speech is proposed. It explicitly takes several main affecting factors as multiplicative companding parameters and estimates all model parameters by an EM algorithm. Experimental results showed that the variance of the observed syllable duration was greatly reduced from 183.4 frame^2 (1 frame = 5 ms) to 18.5 frame^2 by eliminating effects from these affecting factors. Besides, the estimated companding values of these affecting factors agreed well to our prior linguistic knowledge. A preliminary study of applying the proposed model to predict syllable duration for TTS is also performed. Experimental results showed that it outperformed the conventional regressive prediction method. Lastly, an extension of the approach to incorporate initial and final duration modeling is presented. This leads to a better understanding of the relation between the companding factors of initial and final duration models and those of syllable duration model.

1. INTRODUCTION

Prosody refers to aspects of the speech signal other than the actual words spoken, such as timing and fundamental frequency pattern, and plays an important role in the disambiguation of discourse structure. Speakers use prosody to convey emphasis, intent, attitude, and to provide cues to aid the listener in the interpretation of their speech. Researchers have noticed that fluent spoken speech is not produced in a smooth, unvarying stream. Rather, speech has perceptible breaks, relatively stronger and weaker, as well as longer and shorter syllables. The principal acoustic cues of prosody and the relationships between prosodic units are not totally understood. Indeed, the lack of a general consensus in these areas is the main reason why prosody was under-utilized in spoken language systems. Prosodic modeling is therefore important and urgent in speech processing.

In this paper, we concentrate our study on one important issue of prosodic modeling — duration modeling. Duration modeling is important in both automatic speech recognition (ASR) [1-3] and text-to-speech (TTS) [4,5]. In ASR, state duration models are usually constructed to assist in the HMM-based speech recognition. In TTS, synthesis of proper duration information is essential for generating a highly natural synthetic speech. A precise duration model is surely helpful to improve the performance of ASR as well as to make synthesized speech more natural in TTS. Our goal is thus to develop a precise syllable duration model for Mandarin speech. The approach we adopt is statistical-based. It employs a statistical model to describe the syllable duration while considering some major affecting factors that control its variation. These affecting factors include speaker, utterance, prosodic state, tone, and syllable. The first two factors account for two different levels of speaking rate. The third one, prosodic state, is conceptually defined as the state in a prosodic phrase. The well-known lengthening effect for the last syllable of a prosodic phrase belongs to this affecting factor. Due to the fact that prosodic state is not explicitly given, an expectation-maximization (EM) algorithm is derived to estimate all parameters of the model from a large training set. A by-product of the EM algorithm is the determination of the hidden prosodic states of all utterances in the training set. This is an additional advantage because prosodic labeling has become a popular research topic recently [6].

The paper is organized as follows. Section 2 discusses the proposed syllable duration model of Mandarin speech in details. Section 3 describes the experimental results of the syllable duration modeling study. An application of the model to syllable duration prediction for Mandarin TTS is presented in Section 4. Extension of the syllable duration modeling to incorporate initial and final duration modeling is discussed in Section 5. Some conclusions and possible future works are given in the last section.

2. THE PROPOSED SYLLABLE DURATION MODEL

In duration modeling, the desired modeling units can be speech segments like HMM states, phones, initials, finals, syllables or even words and they are affected by many different factors. In this study, the syllable duration of Mandarin speech is considered as the modeling unit, and lexical tone, base-syllable, utterance-level speaking rate, speaker-level speaking rate, and prosodic state are chosen as the relevant affecting factors. The model is expressed by

$$Z_n = X_n g_{t_n} g_{y_n} g_{j_n} g_{l_n} g_{s_n}, \quad (1)$$

where Z_n and X_n are, respectively, the observed duration and the normalized duration of the *n*th syllable; g is an affecting factor; t_n, y_n, j_n, l_n and s_n represent respectively the lexical tone, prosodic state, base-syllable, utterance, and speaker of the *n*th syllable; and X_n is modeled as a normal distribution with mean *u* and variance *v*.

To estimate the parameters of the model, the widely-used approach based on the maximum likelihood (ML) criterion can be adopted. But the closed-form solution of the ML estimation is difficult to obtain. We therefore solve the problem using the EM algorithm [7]. The EM algorithm is derived based on incomplete training data with prosodic state being treated as hidden or unknown. In the following, we discuss it in more detail.

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

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

where N is the total number of training samples, Y is the total number of prosodic states, $p(y_n \mid Z_n, l)$ and $p(Z_n, y_n \mid \bar{l})$ are conditional probabilities which can be derived from the assumed model given in Eq.(1), and $l = \{u, v, g_t, g_y, g_j, g_l, g_s\}$ 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 multiplicative affecting factors. This is obvious because, if we scale up an affecting factor and scale down another, the same objective value will be reached.

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

$$Q(l,\bar{l}) = \sum_{n=1}^{N}\sum_{y_n=1}^{Y} p(y_n \mid Z_n, l)\log p(Z_n, y_n \mid \bar{l}) + h\left(\sum_{n=1}^{N} u g_{t_n} g_{y_n} g_{j_n} g_{l_n} g_{s_n} - N u_z\right), \quad (3)$$

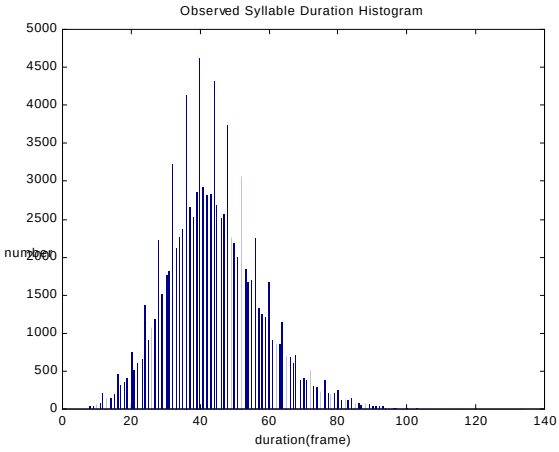
where u_z is the average of Z_n and h is a Lagrange multiplier. The constrained optimization is finally solved by the Newton-Raphson method.

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

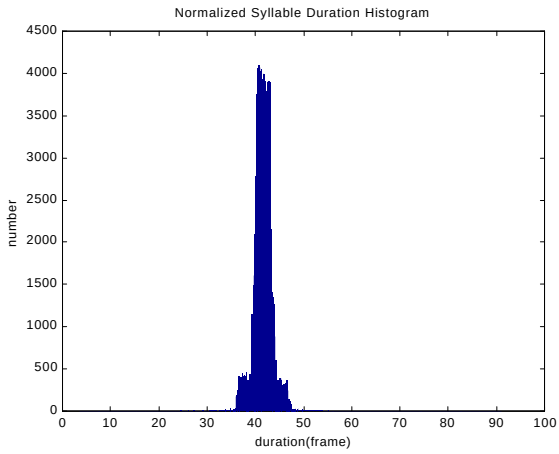
$$y_n = \max_{y} p(y \mid Z_n, l) \quad (4)$$

3. EXPERIMENTAL RESULTS

Effectiveness of the proposed method was examined by simulation using two data sets, one for training and the other for testing. The training set contained utterances of 455 sentences and 200 long paragraphs uttered by four speakers including two males and two females. Each speaker read these texts once. The



(a)



(b)

Figure 1: The histograms of (a) the observed syllable duration and (b) the normalized version in the training set.

data set contained, in total, 101432 syllables. The test set contained utterances of different texts of 100 paragraphs generated by another female. The total number of syllables was 22132. Texts of these two sets covered a wide range, such as news, primary school textbooks, literature, and even some phonetic balanced sentences. All speech signals were manually pre-segmented into sub-syllables of initial and final. The transcription and tone labeling were performed automatically by a linguistic processor with an 80000-word lexicon.

First, initial values of all parameters were independently estimated from the training set. Here, prosodic states of syllables were labeled by a vector quantizer with 8 codewords. Then, the EM algorithm was performed to update all parameters until convergence. The variances of the observed syllable duration were 183.4 and 139.6 frame^2 for the training and testing data sets. Here one frame equals 5 ms. The resulting variances of the normalized syllable duration reduced to 18.5 and 31.4 frame^2 for the closed and open tests. The corresponding root mean squared

Table 1: The estimated companding factors for five lexical tones.

Tone	1	2	3	4	5
g	1.01	1.03	0.98	1.02	0.84

Table 2: The estimated companding factors for eight prosodic states.

State	0	1	2	3	4	5	6	7
g	0.59	1.02	0.87	1.25	0.95	1.10	0.78	1.61

Yong 3 bai 3 zi 2 lian 7, lei 6 si 2 hua 7, ji 1 bai 1 he 7, long 5 dan 3, tu 6 er 0 qi 1 jie 5 geng 3 he 3 suan 3 xiang 4 teng 6 wei 2 cai 7, yi 6 wei 0 na 5 si 3 zhi 5 hu 1 de 7 shi 1 gao 1 hua 2 qi 3 hong 5 tuo 7, hao 6 yi 0 tang 5 chun 2 yu 1 meng 2 meng 4 de 7 jiao 1 wai 4 tian 1 ye 2 feng 1 guang 7.

Figure 2: An example of prosodic state labeling of Mandarin by the EM algorithm.

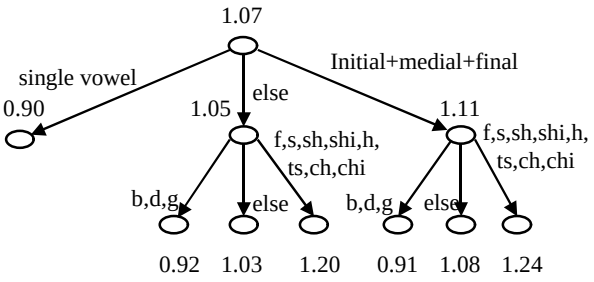


Figure 3: A tree analysis of the base-syllable companding factors. The number associated with a node is the mean of companding values of the base-syllables belonging to the class.

errors between the original and estimated syllable durations were 2.45 and 2.96 frame for the closed and open tests.

The histograms of the observed and normalized syllable durations of the training set are plotted in Fig. 1. Can be seen from these two figures that the assumptions of Gaussian distribution of these two duration units are reasonable. Besides, the histogram of final duration also shows a great match to a Gaussian distribution. For initial duration, the distribution looks similar without considering the very short initials {b, d, g} which are generally difficult to segment correctly from speech.

Tables 1 and 2 show the companding values of the two affecting factors for lexical tone and prosodic state, respectively. Can be seen from Table 1 that Tone 5 has relatively smaller companding value so as to make the associated syllable duration much shorter than those of the other four tones. This agrees to the prior linguistic knowledge. It can be found from Table 2 that the two prosodic states of 3 and 7 have relatively large companding

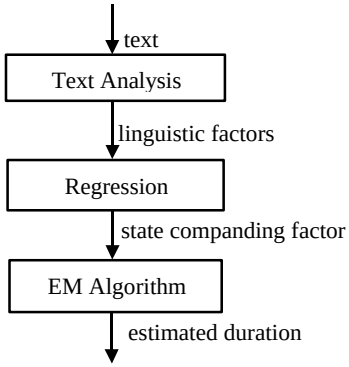


Figure 4: A hybrid statistical/regression approach for syllable duration prediction.

values while State 0 has smaller companding value. Fig. 2 shows an example of prosodic state labeling of Mandarin by the EM training algorithm. From Fig. 2, we find that States 3 and 7 usually associate with the ending syllables of prosodic phrases and State 0 always associates with intermediate syllables of ploy-syllabic words. The finding complies with the prior knowledge of the lengthening effect for the last syllable of a prosodic phrase. Fig. 3 illustrates the result of a tree analysis of the base-syllable companding factors based on the phonetic structure of base-syllable. It reveals that syllables of single vowel have shorter durations than those with initials or medials, syllables with initials belonging to {f, s, sh, shi, h, ts, ch, chi} are longer, and syllables with initials belonging to {b, d, g} are shorter. This observation matches with our general knowledge of the phonetic characteristics of Mandarin base-syllables.

4. AN APPLICATION TO DURATION PREDICTION FOR TTS

We now apply the above model to predict syllable duration for Mandarin TTS. A hybrid statistical/regression approach to synthesizing syllable duration is suggested. Fig. 4 shows the block diagram of the approach. Instead of direct predicting syllable duration from the input linguistic features, it first estimates the prosodic-state companding value from the linguistic features by the regression technique, and then predicts the syllable duration by the statistical model. Here linguistic features are extracted via analyzing the input text by an automatic word tokenization algorithm with an 80000-word lexicon. The linguistic features used include some sentence-level features, such as sentence length and position in sentence, some word-level features, such as word length and position in word, and punctuation-mark indicators.

RMSEs of 8.18 and 11.76 frame were obtained for the closed and open tests, respectively. The results are better than those of 9.36 and 14.73 frame achieved by the conventional regressive prediction method. It is noted that, in the latter case some additional linguistic features related to the tones and the phonetic features of the current syllable and its two neighbors were used.

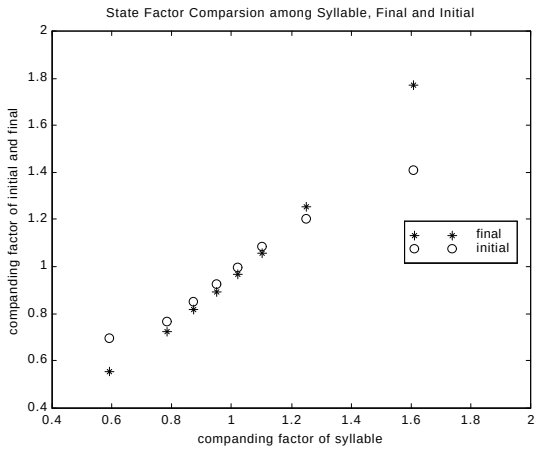


Figure 5: The relations between the prosodic-state companding factors of the initial and final duration models and those of the syllable duration model.

5. INCORPORATION OF INITIAL AND FINAL DURATION MODELING

Though many people are working on duration modeling, there are very few studies trying to exploit the relationship between the syllable duration and its constituent initial and final durations for Mandarin speech. By utilizing the distinct ability of the proposed syllable duration model to isolate the affections from several main affecting factors, we extend the study to incorporate initial and final duration modeling. In the study, both initial and final durations are modeled in the same way as the syllable duration with additional constraints of sharing the same prosodic states derived from the previous syllable duration modeling. Fig. 5 displays the relations between the prosodic-state companding factors of the initial and final duration models and those of the syllable duration model.

Can be seen from Fig. 5 that the companding factors matched well for the three models for all states except the two extreme cases of States 0 and 7 which have the smallest and largest companding values. By closer examination, we find that final duration is compressed and expanded more seriously than initial duration at these two extreme states.

6. CONCLUSIONS AND FUTURE WORKS

A new statistical-based duration modeling approach has been proposed in the paper. Experimental results have confirmed its effectiveness on isolating several main factors that seriously affects the syllable duration of Mandarin speech. Aside from greatly reducing the variance of the modeled syllable duration, the estimated companding factors conformed well to the prior linguistic knowledge of Mandarin speech. Besides, the prosodic-state labels produced by the EM algorithm were linguistically meaningful. So it is a promising syllable duration modeling approach for Mandarin speech.

Some future works are worthwhile doing. Firstly, the syllable duration model can be further improved via considering more affecting 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 inter-syllable pause duration.

7. REFERENCES

[1] C.Mitchell, M. Harper, L. Jamieson & R. Helzermam (1995), "A parallel implementation of a hidden Markov model with duration modeling for speech recognition," *Digital Signal Proc.* 5, pp.43-57

[2] X. Huang, H. Hon, M. Hwang & K. Lee (1993), "A comparative study of discrete, semicontinuous, and continuous hidden Markov models," *Computer, Speech & Lang.* 7, pp.359-368.

[3] S. Levinson (1986), "Continuously variable duration hidden Markov models for speech analysis," *Proc. IEEE ICASSP*, pp. 1241-1244

[4] D. H. Klatt (1987), "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.* 82, pp.137-181.

[5] S. H. Chen, S. H. Hwang and Y. R. Wang (1998), "An RNN-based prosodic information synthesizer for Mandarin text-to-speech," *IEEE Trans. Speech and Audio processing*, vol. 6, no.3, pp.226-239.

[6] 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.

[7] S. Young and G. Bloothooft, "Corpus-Based Methods in Language and Speech Processing". Kluwer Academic Publishers, 1997, pp. 1 – 26.

[8] C. C. Ho and S. H. Chen, "A Maximum Likelihood Estimation of Duration Models for Taiwanese Speech," *proc. of ISAS/SCI'2000*, Orlando, USA, July 2000.

[9] C. C. Ho and S. H. Chen, "A Hybrid Statistical/RNN Approach to Prosody synthesis for Taiwanese TTS," *proc. of ICSLP'2000*, Beijing, Oct. 2000.