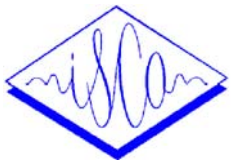

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Duration Modeling for Mandarin Speech Recognition Using Prosodic Information

Wern-Jun Wang[1] *and Chun-Jen Lee*[1,2]

[1] Internet & Multimedia Application Technology Laboratory, Chunghwa Telecom. Laboratories, Taiwan, R.O.C.
[2] Department of Computer Science, National Tsing Hua University, Taiwan, R.O.C.

{wernjun, cjlee}@cht.com.tw

Abstract

In this paper, a new duration modeling method for HMM-based Mandarin base-syllable recognition is proposed. It extends the conventional state duration method to further consider the speaking rate of utterance and add a syllable duration model to help the recognition search finding the best-recognized base-syllable string. Experimental results showed that the proposed method was effective on improving the recognition accuracy.

1. Introduction

Duration modeling has been widely employed in speech recognition to help confining the search process so as to improve the recognition accuracy. The most popular duration model used in HMM-based speech recognition is the state duration model, which explicitly models each state duration with a Gamma distribution [1]. Parameters of the model are usually estimated from a large training set containing utterances of different speaking rate. One main drawback of the approach is the inaccuracy of modeling the durations of speech patterns with both fast and slow speaking rates. Recognition performances are therefore suffered for testing utterances of these two extreme cases. Recently, speech recognition incorporated with high accurate duration models were proposed [2,3]. But, they usually used duration models in the post processing stage to provide penalties to candidates of abnormal duration for reordering the N-best hypotheses suggested by acoustic decoding. It remains unsolvable to directly incorporate such accurate duration models in acoustic decoding because some parameters, such as speaking rate, cannot be reliably estimated until the hypotheses are constructed.

In this paper we try to partially solve the problem via extending the conventional state duration method to further consider the speaking rate of the testing utterance and adding a new syllable duration model in an HMM-based Mandarin base-syllable recognizer for improving its performance. The proposed statistical duration modeling method has been tried on Min-Nan and Mandarin Chinese text-to-speech system and got some significant improvements on the duration prediction [4,5].

2. The proposed statistical duration model

The analysis unit used in duration modeling can be speech segment like HMM state, phone, *initial*/*final*, syllable or even word for Mandarin Speech. In this section, we start with introducing an estimation approach based on maximum likelihood (ML) criterion for constructing a syllable duration model. Then, the approach can be easily applied to the modeling of HMM state duration.

The syllable duration model, considering three affecting factors of speaking rate, prosodic state and lexical tone, is expressed by [4]

$$Z_n \cdot \beta_{t_n} \cdot \beta_{p_n} \cdot \beta_{l_n} = X_n \qquad (1)$$

where Z_n is the observed duration of the syllable $S_n=(j_n, t_n, p_n, l_n)$ with base-syllable j_n, lexical tone t_n, prosodic state p_n, and in utterance l_n; β_{t_n}, β_{p_n}, and β_{l_n} are the companding factors of the syllable duration due to the three affecting factors of t_n, p_n, and l_n, respectively; and X_n is the normalized syllable duration and is modeled as a Gaussian (normal) distribution. Here the speaking rate companding factor β_{l_n} of utterance l_n is simply defined as the mean duration of all syllables in the utterance. The prosodic state p_n is conceptually defined as the state of a syllable in a prosodic phrase and is labeled automatically by a vector quantization classifier using seven acoustic features extracted from the vicinity of the current syllable. These seven features include normalized syllable *final* duration, syllable pitch mean and its differences with the two nearest neighboring syllables, and syllable log-energy maximum and its differences with the two nearest neighboring syllables. Eight prosodic states are empirically set in our study and two special states are assigned to represent the beginning and ending syllables of utterance. Therefore the total number of prosodic states was ten. As shown in our previous study [6,7], the proposed prosodic modeling approach has functioned well on detecting prosodic states from acoustic cues [6] and has shown its effectiveness in Mandarin speech-to-text conversion process [7]. It is also noted that the numbers of lexical tone and base-syllable are, respectively, 5 and 411 for Mandarin speech. Finally, the ML estimation based on an iterative optimization procedure was used to sequentially estimate the parameters of the syllable duration model [4].

3. Incorporating the proposed syllable duration model into continuous Mandarin speech recognition

Since the three affecting factors of speaking rate, lexical tone, and prosodic state are not known in the acoustic decoding stage, we can not directly incorporate the above syllable duration model into the HMM-based speech recognizer. To solve the problem, we perform the following modification to the syllable duration model described in Section 2.

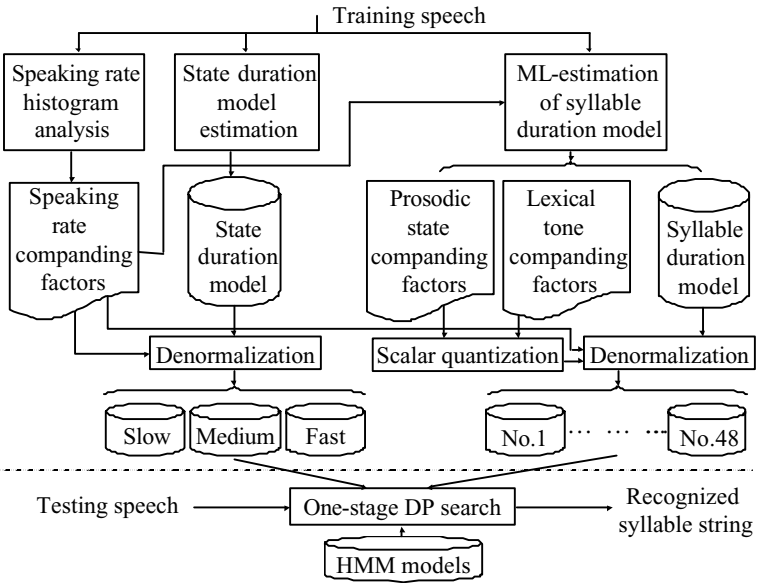

Figure 1: *A block diagram of the proposed duration modeling method for Mandarin base-syllable recognition.*

Fig. 1 shows a block diagram of the duration modeling part of the proposed method. The duration modeling consists of a training phase (the upper part of Fig. 1) and a testing phase (the lower part of Fig. 1). In the training phase, it first generates a state duration model and then performs a speaking rate histogram analysis to partition all training utterances into three classes of slow, medium and fast rates. The mean speaking rate of each class is then calculated and converted into the speaking rate companding factor by divide it to the average speaking rate of the whole training set. It then uses these three companding factors to de-normalize the state duration model into three models of slow, medium and fast rates.

Meanwhile, a syllable duration model is ML estimated based on Eq. (1). This syllable duration model also considers three affecting factors of speaking rate, prosodic state and lexical tone. But now the speaking rate companding factor in Eq. (1) is quantized to three values corresponding to slow, medium and fast rates. Besides, the numbers of lexical tone, base-syllable, and prosodic state are, respectively, set to 5, 411, and 10, respectively for Mandarin as described in Section 2. Lastly, the ML iterative optimization procedure is performed to sequentially estimate parameters of the syllable duration model.

After obtaining the companding factors of prosodic state and lexical tone, it then uses a scalar quantization to reduce the combinations of them from 50 (i.e. *number of lexical tone* multiplied by *number of prosodic state*) to a smaller number (16 in this study). It lastly uses these 16 values and the 3 speaking rate companding factors to de-normalize the syllable duration model into 48 models of various rates.

In the testing phase of Fig. 1, these three state duration models and 48 syllable duration models are used in the one-stage dynamic programming (DP) search to help finding the best base-syllable sequence. Based on theses duration models, the duration penalties are applied for both state and syllable transitions in the DP search. The reasons why we adopted such schemes are according to the pilot tests (section *4.2.1*).

4. Experimental results

A large multi-speaker database recorded by 50 female and 50 male speakers was used to test the proposed method. The text materials of this database were randomly extracted from Sinica Corpus [8] that was established by Academia Sinica in Taiwan. Each speaker read several short paragraphic texts. Each text contains about 150 characters. All paragraphic utterances were manually divided into short sentential utterances. The database was divided into two parts, one for training and one for testing. The training set contained 124403 syllables uttered by 40 female and 40 male speakers. The test set contained 23965 syllables uttered by other 10 female and 10 male speakers.

The recognition features used in this study included 12 Mel-frequency cepstral coefficients (MFCCs), 12 delta MFCCs, and a delta log-energy. An HMM recognizer was trained from the training set by the segmental k-means training algorithm. It used 100 3-state right-*final*-dependent (RFD) syllable *initial* models and 39 5-state context-independent (CI) syllable *final* models to form 411 8-state base-syllable models. For silence, a single-state model was used. Observation features in each HMM state were modeled by a mixture Gaussian distribution. The number of mixture components used in each state depended on the number of training data and was set in the range from 1 to 15.

4.1. Results of duration modeling

We first examined the experimental results of the two duration modeling studies for HMM state duration and syllable duration. Tables 1 and 2 presented the results of companding factors for five lexical tones and for 10 prosodic states, respectively. It can be found from Table 1 that the duration shortening phenomena is very clear for tone 5 because most of its companding factors are greater than one. As for the other four tones, the shortening and lengthening effects are not significant. From Table 2, we find that duration lengthening is

effective for prosodic states 0, 3 and 9. Obviously, these three states correspond to the ending parts of prosodic phrases. This was confirmed by checking the associated text to find that syllables with these three states always located at the endings of words, phrases, or sentences. Finally, an unusual value of 0.76 is found as the companding factor of prosodic state 8 for the duration of HMM state 0. Since prosodic state 8 corresponds to the beginning syllable of an utterance, this may be result from the segmentation error of the HMM recognizer caused by the silence resided in the beginning parts of the training utterances.

4.2. Results of continuous speech recognition

The performance of incorporating the proposed duration modeling method in speech recognition was then evaluated. Fig. 2 shows the histogram of speaking rate for all 6920 utterances of the training set. Two thresholds were used to partition the training set into three classes of slow, medium and fast rates. They contained, respectively, 25%, 50% and 25% utterances of the training set.

Table 1: *Experimental results of the companding factors for five lexical tones.*

HMM State	Tone				
	1	2	3	4	5
0	1.00	0.98	1.00	1.00	1.26
1	1.01	0.98	1.02	1.00	1.06
2	0.98	1.01	1.06	0.95	1.26
3	0.98	0.96	1.08	0.97	1.40
4	0.97	0.99	1.08	0.97	1.24
5	0.99	1.01	1.01	1.00	0.99
6	0.99	1.00	1.02	1.00	1.00
7	1.02	1.03	1.02	0.97	0.91
Syllable	0.99	0.99	1.05	0.98	1.13

Table 2: *Experimental results of the companding factors for 10 prosodic states.*

HMM State	Prosodic State									
	0	1	2	3	4	5	6	7	8	9
0	1.00	0.95	0.99	0.93	1.07	1.00	1.07	1.05	0.76	0.90
1	0.98	1.00	0.99	0.98	1.03	1.00	1.02	1.02	0.99	0.87
2	1.01	1.01	0.99	0.96	1.02	1.00	1.03	1.01	0.98	0.83
3	0.91	1.11	1.16	0.89	0.98	1.10	1.05	1.03	1.04	0.80
4	0.90	1.13	1.21	0.88	0.96	1.12	1.06	1.03	0.97	0.83
5	0.86	1.14	1.16	0.90	0.98	1.14	1.04	1.04	1.04	0.68
6	0.89	1.11	1.12	0.91	1.00	1.11	1.02	1.06	1.00	0.61
7	0.80	1.20	1.07	0.84	1.02	1.29	1.03	1.19	1.19	0.47
Syllable	0.92	1.07	1.06	0.92	1.01	1.09	1.03	1.06	0.97	0.74

4.2.1. The pilot tests

Before continuing our following experiments, we first want to decide the suitable duration model to be incorporated into speech recognizer. We have considered the following two issues, the improvement on accuracy rate and the complexity introduced from the specific duration model, and tried to find a compromise between these two issues. Two pilot tests were therefore performed to compare the performance of the following two schemes using different duration models:

P-test 1: The HMM-based recognizer incorporating with state-based duration models are de-normalized by three speaking rates, five lexical tones, and 10 prosodic states.

P-test 2: The HMM-based recognizer incorporating with state-based duration models are de-normalized by three speaking rates and syllable-based duration models are de-normalized by three speaking rates, five lexical tones, and 10 prosodic states.

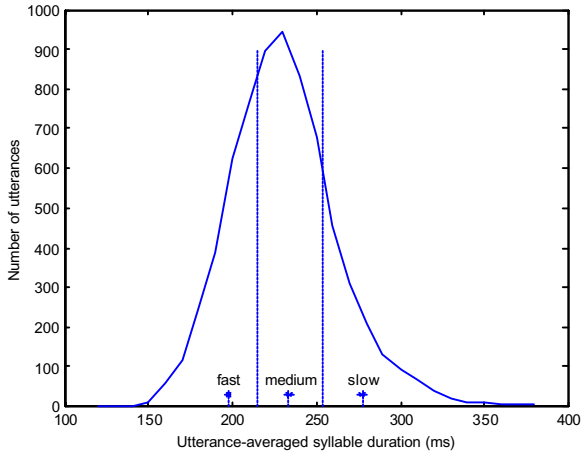

Figure 2: *The histogram of speaking rate for all 6920 training utterances. Three short lines with star marks indicate the mean speaking rates of the three classes of slow, medium and fast rates.*

A subset of the testing set containing utterances of 2 male and 2 female speakers was used in these pilot tests. To evaluate the complexity introduced by these two schemes, it was easy to find that the search space of the one-stage DP search would be enlarged by 50 times if the affecting factors of lexical tone and prosodic state were considered for all state transitions. The computation time for pilot test 1 would thus increase tremendously. On the other hand, if these two affecting factors were considered for syllable transitions only, the increase of the computation time for pilot test 2 would be moderate.

The experimental results indicated that the syllable accuracy rates are 66.1% and 67.0% for pilot test 1 and 2, respectively. The results of this pilot test did not imply that syllable-based duration model outperformed the state-based duration model. Actually, duration penalty provided by the state-based duration model have been applied for both schemes in this test. The conclusion we can make is that lexical tone and prosodic state are more effective to be considered in syllable transition than in state transition. Therefore, syllable-based companding factors of lexical tone and prosodic state were used in the following test.

4.2.2. The main test

In this test, we have the following four schemes:

Baseline: The HMM method without using any explicit duration model.

Scheme 1: The HMM method incorporating with the conventional state duration model.

Scheme 2: The HMM method incorporating with 3 state duration models of slow, medium and fast rates.

Scheme 3: The HMM method incorporating with 3 state duration models of slow, medium and fast rates and 48 syllable duration models.

Table 3 displays the experimental results. It can be found from this table that the performances of these four schemes are in the order of Baseline, Scheme 1, Scheme 2 and Scheme 3, with Scheme 3 has the best recognition rate of 60.8%. This shows that the performance of the HMM recognizer was improved as we considered more affecting factors in duration modeling. The results also indicate that the performance improvements are consistent for all three testing classes of slow, medium and fast rates.

Table 3: *Base-syllable accuracy rates (%) of the four base-syllable recognition schemes for different testing classes of slow, medium and fast speaking rates.*

Speaking Rate	Slow	Medium	Fast	Total
Baseline	51.8	56.6	50.9	55.3
Scheme 1	56.6	60.1	53.4	58.9
Scheme 2	58.3	60.7	54.0	59.8
Scheme 3	59.8	61.7	55.1	60.8

5. Conclusions

A new duration modeling method for syllable duration and HMM state duration has been discussed in this paper. Experimental results confirmed that the method could isolate the effects of the three major affecting factors of speaking rate, prosodic state and lexical tone. Incorporation of the proposed duration modeling method in continuous Mandarin speech recognition has also been studied. The encouraging results have been showed form the experiments. To further improve the effectiveness of our proposed duration modeling in speech recognition, we have to consider more affecting factors and choose the effective ones via conducting more detailed analysis processes.

6. References

[1] Burshtein, D., 1995. Robust Parametric Modeling of Durations in Hidden Markov Models. *Proc. IEEE Intern. Conf. on Acoust., Speech, Signal Process (ICASSP)*, vol. 1, 548-551.

[2] Chung, G. Y. and Seneff, S., 1999. A Hierarchical Duration Model for Speech Recognition Based on the ANGIE Framework. *Speech Communication*, vol. 27, 113-134.

[3] Pols, L. C. W., Wang, X., and ten Bosch, L. F. M., 1996. Modeling of Phone Duration (Using the TIMIT Database) and Its Potential Benefit for ASR. *Speech Communication*, vol. 19, 161-176.

[4] Ho, C. C., and Chen, S. H., 2000. A Maximum Likelihood Estimation of Duration Models for Taiwanese Speech. *Proc. 4th World multiconference on Systemics Cybernetics and Informatics,* vol. VI, 395-399.

[5] Lai, W. H. and Chen, S. H., 2002. Analysis of Syllable Duration Models for Mandarin Speech. *Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process. (ICASSP),* vol. I, 497-500.

[6] Wang, W. J. and Chen, S. H., 2002. The Study of Prosodic Modeling for Mandarin Speech. *Proc. Int. Computer Symposium (ICS)*, vol. 2, 1777-1784.

[7] Wang, W. J., Liao, Y. F. and Chen, S. H., 2002. RNN-based Prosodic Modeling of Mandarin Speech and Its Application to Speech-to-Text Conversion. *Speech Communication*, vol.36, no.3-4, 247-265.

[8] Chinese Knowledge Information Processing Group, 1995. The contents and descriptions of Sinica Corpus. Technical Report no. 95-02, Academia Sinica.