DURATION MODELING USING CUMULATIVE DURATION PROBABILITY AND SPEAKING RATE COMPENSATION

Tae-Young Yang, Ji-Sung Kim, Chungyong Lee, Dae Hee Youn, and Il-Whan Cha

Dept. of Electronic Eng., Yonsei Univ.
134 Shinchon-dong, Sudaemoon-ku, Seoul, 120-749, Korea
e-mail) tyyang@ghost.yonsei.ac.kr

ABSTRACT

A duration modeling scheme and a speaking rate compensation technique are presented for the HMM based connected digit recognizer. The proposed duration modeling technique uses a cumulative duration probability. The cumulative duration probability also can be used to obtain the duration bounds for the bounded duration modeling. One of the advantages of proposed technique is that the cumulative duration probability can be applied directly to the Viterbi decoding procedure without additional postprocessing. Therefore, it rules the state and word transition at each frame. To alleviate the problems due to fast or slow speech, a modification to the bounded duration modeling which accounts for speaking rate is described. The experimental results on Korean connected digit recognition show the effectiveness of the proposed duration modeling scheme and the speaking rate compensation technique.

1. INTRODUCTION

In connected digit recognition task, the inserted words are observed for unrealistically short durations, while the previous or next word of the deleted words is observed for abnormally long durations. Such misalignments of word duration sequences can be reduced by modeling of the state and word durations. Speech rate is another problem in connected digit recognition task. It has been known to have a significant effect on recognition [1]. Furthermore, the duration modeling technique cannot play a proper role when the input speech is too fast or too slow.

Several approaches have been proposed to model the state and word durations explicitly or implicitly. In explicit duration modeling, the duration probabilities are measured from the training data or several distribution densities, which are usually incorporated in postprocessing as the weights of multiple candidates [2][3]; Parametric distribution [4][5] using Gaussian, Poisson or Gamma density has been suggested to model the state and word durations. In the bounded state duration modeling [6] the state duration is lower and upper bounded by two bounding parameters in the recognition phase.

Another approach to treat the duration problem is to model the state duration implicitly such as the Ferguson model [7], the expanded state HMM [8][9], the second-order HMM [10] and the inhomogeneous HMM (IHMM) [11].

In this paper, we propose a method using the cumulative duration probabilities to model the state and word durations explicitly and a speaking rate compensation technique. The cumulative duration probability is measured by the partial sum of the conventional explicit duration probabilities. Therefore, we call it "cumulative duration probability" in this paper.

The proposed duration modeling technique is described in section 2. In section 3 the modification to duration modeling which reduces the effect of speech rate is presented. The experimental settings and results are described in section 4.

2. DURATION MODELING USING CUMULATIVE DURATION PROBABILITY

2.1. Cumulative Duration Probability

In explicit duration modeling, the probability $P_s(w,i,\tau)$ which denotes a discrete distribution of the state duration in state i of word w for τ frames is defined as the following:

$$P_s(w,i,\tau) = Pr[s_t = (w,i) \text{ for } \tau \text{ frames} | s_t = i, s_{t+1} \neq i], \quad (1)$$

$$\sum_{\tau=1}^{D_s} P_s(w,i,\tau) \; = \; 1, \quad (2)$$

where, D_s is the largest state duration allowed. Since it should be added at the end of a state or a word, it cannot be applied to the Viterbi decoding procedure. Thus an additional postprocessor is needed. This implies that the state and word durations do not play a role in the forward

This work was supported by Korea Science and Engineering Foundation (KOSEF) 95-0100-22-01-3 in 1997.

recognition path and that they do not rule the state and word transition at each frame. To overcome such problems we propose a method using the cumulative duration probability that can be combined directly to the Viterbi decoding procedure.

The cumulative state duration probability $\hat{P}_s(w,i,\tau)$ is defined by

$$\hat{P}_s(w,i,\tau) \;=\; Pr[s_{t+1} = s_t = (w,i) \text{ for } \tau \text{ frames}]. \quad (3)$$

It is measured as the following:

$$\hat{P}_s(w,i,\tau) \;=\; Pr[\text{transition accurs after } \tau \text{ frames}] \quad (4)$$

$$= \sum_{d=\tau+1}^{D} P_s(w,i,d). \quad (5)$$

It is the partial sum of the explicit duration probabilities which can be calculated from the training speech data or estimated from the several parametric distributions. The cumulative word duration probability $\hat{P}(w,\tau)$ is obtained from the same way as we mentioned before.

The Viterbi decoding algorithm is modified to utilize both the cumulative state and word duration probabilities. When only two state transition paths are considered for convenience, a modified Viterbi decoding algorithm within a word is given by

$$\begin{aligned}\delta_t(w,i) \;=\; & \\ & \max[\delta_{t-1}(w,i)\cdot a^w_{i,i}\cdot \hat{P}_s(w,i,\tau_i)\cdot \hat{P_w}(w,\tau_w), \\ & \delta_{t-1}(w,i-1)\cdot a^w_{i-1,i}\cdot(1-\hat{P}_s(w,i-1,\tau_{i-1}))\cdot \\ & \hat{P_w}(w,\tau_w)]\cdot b^w_i(O_t), \end{aligned} \quad (6)$$

where, $\delta_t(w,i)$ is the Viterbi score in state i of word w at time t, $a^w_{i,j}$ denotes the state transition probability from state i to state j of word w, $b^w_i(O_t)$ is the observation probability in state i of word w at time t, and τ_i and τ_w represent the duration in state i and in word w, respectively. For the first state of each word where the word transition is considered, a modified Viterbi decoding algorithm is of the form:

$$\begin{aligned}\delta_t(w,1) \;=\; & \\ & \max_{1\le v\le W}[\delta_{t-1}(w,1)\cdot a^w_{1,1}\cdot \hat{P}_s(w,1,\tau_1)\cdot \hat{P}_w(w,\tau_w), \\ & \delta_{t-1}(v,N)\cdot a^v_{N,N+1}\cdot(1-\hat{P}_s(v,N,\tau_N))\cdot \\ & (1-\hat{P_w}(v,\tau_v))]\cdot b^w_1(O_t), \end{aligned} \quad (7)$$

where, W denotes the total number of words, N is the number of states in word w and $a^w_{N,N+1}$ represents the transition probability from the last state N to the virtual state $N+1$, which means the word transition occurs.

2.2. Bounded Duration Model

The cumulative state and word duration probabilities can be used to estimate the lower and upper bounds for the bounded duration modeling. The cumulative duration probability is a monotonically decreasing probability. Therefore, the lower and upper bounds are easily estimated by applying proper thresholds. Let $DL_s(w,i)$ and $DU_s(w,i)$ be the lower and upper duration bounds for state i of word w, and $DL_w(w)$ and $DU_w(w)$ be the lower and upper duration bounds for word w. They are estimated by applying the thresholds TH_{SL}, TH_{SU}, TH_{WL} and TH_{WU} to the state and word cumulative duration probabilities, where TH_{SL} and TH_{SU} denote the lower and upper thresholds for the state duration bounds, TH_{WL} and TH_{WU} represent the lower and upper thresholds for the word duration bounds, respectively.

3. SPEAKING RATE COMPENSATION

In this section, we present a speaking rate compensation technique for the bounded duration modeling. The proposed technique performs the recognition process twice. From the recognition results obtained in the first recognition process, the speaking rate is estimated. Then, the lower and upper duration bounds are adjusted to the speed of the input speech and the second recognition process is performed.

The average duration of each word $\bar{D}(w)$ and the average duration of the total words $\bar{D}$ are estimated from the training data.

$$\bar{D}(w) \;=\; \sum_{\tau=1}^{D_w} P_w(w,\tau)\cdot\tau, \quad 1\le w\le W, \quad (8)$$

$$\bar{D} \;=\; \frac{1}{W}\sum_{w=1}^{W}\bar{D}(w). \quad (9)$$

The duration rate of each word $DR(w)$ is defined by

$$DR(w) \;=\; \frac{\bar{D}(w)}{\bar{D}}. \quad (10)$$

The value of $DR(w)$ is greater than 1 for long words and less than 1 for short words. Using the duration $D(w)$ of each recognized word in the first recognition process, the expected average duration $\hat{D}$ of the input speech and the expected duration $\hat{D}(w)$ of each word is estimated as the following:

$$\hat{D} \;=\; \text{Median}[D(w)/DR(w)], \quad w\in\{\text{recognized words}\}, \quad (11)$$

$$\hat{D}(w) \;=\; \hat{D}\cdot DR(w), \quad 1\le w\le W, \quad (12)$$

where, Median$[\cdot]$ denotes the Median average. We use the Median average, rather than normal average, so that the expected average duration $\hat{D}$ is not affected by wrong words which usually appear for abnormally long or short frames. For the second recognition process, we choose narrower duration bounds than those for the first recognition process.

These duration bounds $DL_s(w,i)$, $DU_s(w,i)$, $DL_w(w)$ and $DU_w(w)$ are adjusted using the expected duration $\hat{D}(w)$ of each word.

$$\hat{DL_s}(w,i) = DL_s(w,i) + (\hat{D}(w) - \bar{D}(w))/N, \quad (13)$$

$$\hat{DU_s}(w,i) = DU_s(w,i) + (\hat{D}(w) - \bar{D}(w))/N, \quad (14)$$

$$\hat{DL_w}(w) = DL_w(w) + (\hat{D}(w) - \bar{D}(w)), \quad (15)$$

$$\hat{DU_w}(w) = DU_w(w) + (\hat{D}(w) - \bar{D}(w)), \quad (16)$$

Then, the second recognition process is performed with the adjusted duration bounds $\hat{DL_s}(w,i)$, $\hat{DU_s}(w,i)$, $\hat{DL_w}(w)$ and $\hat{DU_w}(w)$ which account for the speaking rate.

4. EXPERIMENTAL SETTINGS AND RESULTS

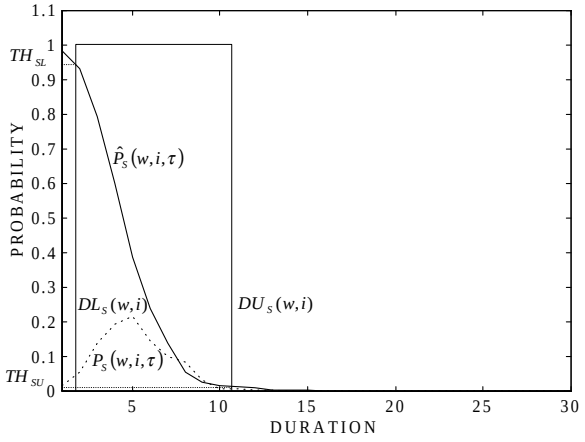
The experiments on Korean connected digit recognition was performed using the DigitDB[1] database which consists of 5,169 connected digit strings from 70 males and 50 females. The data was partitioned into 4,064 strings from 50 males and 40 females for training, and 1,105 strings from 20 males and 10 females for testing. The vocabulary was made up of 30 models; 29 models for 11 digits (the digit "0" is read in two ways in Korean) and 1 model for the silence. The 29 digit models were designed to cover the coarticulation effects due to previous and next digits. The transitions between the 29 digit models were ruled by the word pair grammar.

The speech signals were down-sampled from 16 KHz to 8 KHz and pre-emphasized by a factor of 0.95. Three feature vectors, 12 mel-frequency cepstral coefficients (MFCC), 12 delta MFCC, and delta energy with delta-delta energy were computed every 10 ms using a 20 ms Hamming window. The band pass lifter (BPL) was used in cepstral distance measure.

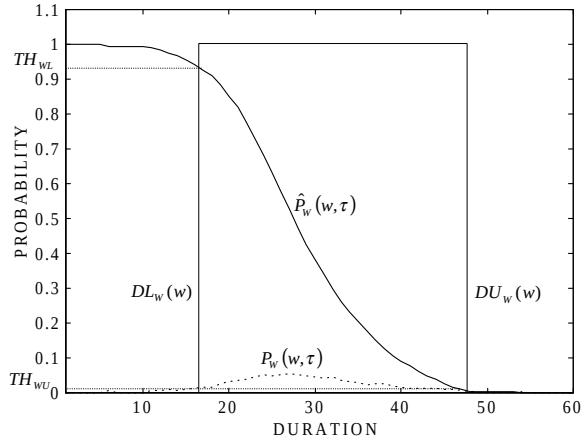
An example of the cumulative duration probability and the duration bounds obtained from it is shown in Figure 1. The cumulative state and word duration probabilities represent the continuity of corresponding state and word. In figure 1 the value of the cumulative duration probability is almost 1 for a low duration period, which makes a state and a word remain as they are. It decreases gradually and for a high duration period it is close to 0, which encourages the state and word transitions.

Several duration modeling schemes using the cumulative state and word duration probabilities and a speaking rate compensation technique were examined. The experimental results are shown in Table 1. In Table 1, BD1 denotes the bounded duration modeling with only one set of state and word duration bounds for all states and words. BD means the bounded duration modeling in which the duration

[1]The DigitDB database has been distributed by the Korea Advanced Institute of Science and Technology (KAIST).



(a)



(b)

Figure 1: An example of the state and word duration modelings: (a) the state duration of the 2nd state of the word "/i/" (2), (b) the word duration of the word "/gong/" (0)

bounds are obtained from the cumulative duration probabilities. CD represents that the cumulative state and word duration probabilities are applied to the Viterbi decoding procedure. CD+BD is the combined scheme of BD and CD. CD-A means that the state transition matrix A of HMM is not used in CD. CD+BD-A denotes that BD is added to CD-A. BD+SRC represents that the speaking rate compensation technique is applied to BD scheme. In BD+SRC scheme, the threshold probabilities for the duration bounds in the first recognition process were $TH_{SL} = 0.95$, $TH_{SU} = 0.001$, $TH_{WL} = 0.93$ and $TH_{WU} = 0.001$. For the second recognition process, they were $TH_{SL} = 0.95$, $TH_{SU} = 0.005$, $TH_{WL} = 0.8$ and $TH_{WU} = 0.01$.

The recognition results in Table 1 show that the proposed duration modeling techniques increase the recognition accuracy approximately by 9.5% compared to that of the conventional HMM, and approximately by 3.4% com-

Table 1: Experimental results for several duration modeling schemes using the cumulative duration probability and the speaking rate compensation technique

Duration modeling	Recognition accuracy[%]	Number of errors		
		Ins	Del	Sub
Baseline	83.60	354	64	228
BD1	89.76	99	164	319
BD	93.47	83	98	190
CD	93.10	73	101	218
CD+BD	93.01	67	107	223
CD-A	93.12	103	79	209
CD+BD-A	93.07	94	84	216
BD+SRC	94.28	51	85	189

pared to that of the BD1 scheme. Among the proposed duration modeling schemes, the best performance was achieved by the BD. In the BD, several sets of thresholds probabilities were used to measure the duration bounds and we carefully tuned them, while one set of thresholds $TH_{SL} = 0.95$, $TH_{SU} = 0.001$, $TH_{WL} = 0.93$ and $TH_{WU} = 0.001$ were used in other BD schemes. Although such a special tuning was not conducted in the proposed speaking rate compensation technique (BD+SRC), the BD+SRC technique achieved the 0.8% of further improvement compared to the BD scheme.

5. CONCLUSION

We have presented a duration modeling using a cumulative duration probability and a speaking rate compensation technique. Since the proposed duration modeling scheme is combined to the Viterbi decoding procedure, an additional postprocessor is not needed. The speaking rate compensation technique is applied to the bounded duration modeling and it reduces the errors due to fast or slow speech.

From the experimental results for the duration modelings, we can note two facts. First, when the cumulative duration probability is combined to the Viterbi decoding procedure, an additional duration modeling such as the bounded duration modeling cannot achieve further enhancement of recognition accuracy. Second, the performance of the duration modeling using the cumulative duration probability is better when the state transition matrix A of HMM is not applied.

6. REFERENCES

[1] Matthew A. Siegler and Richard M. Stern, "On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems," *Proc. ICASSP*, vol. 1, pp. 612-615, May 1995.

[2] Lalit R. Bahl, Frederick Jelinek and Robert L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 5, no. 2. pp. 179-190, Mar. 1983,

[3] Lawrence R. Rabiner, Jay G. Wilpon and Frank K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Models," *IEEE Trans. on ASSP*, vol. 37. no. 8, pp. 1214-1225, Aug. 1989.

[4] M. J. Russel and R. K. Moore, "Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition," *Proc. ICASSP*, vol. 1, pp. 5-8, Mar. 1985.

[5] David Burshtein, "Robust Parametric Modeling of Durations in Hidden Markov Models," *Proc. ICASSP*, vol. 1, pp. 548-551, May 1995.

[6] H. Y. Gu, C. Y. Tseng and L. S. Lee, "Isolated- Utterance Speech Recognition Using Hidden Markov Models with Bounded State Durations," *IEEE Trans. Signal Processing*, vol. 39, no. 8, pp. 1743-1751, August 1991.

[7] J. D. Ferguson, "Variable Duration Models for Speech," *Proc. Symp. on the Application of Hidden Markov Models to Text and Speech*, pp. 143-179, Oct. 1980.

[8] Antonio Bonafonte, Josep Vidal and Albino Noqueiras, "Duration Modeling with Expanded HMM Applied to Speech Recognition," *Proc. ICSLP*, vol. 2, pp. 1097-1100, Oct. 1996.

[9] Kevin Power, "Durational Modeling for Improved Connected Digit Recognition," *Proc. ICSLP*, vol. 2, pp. 885-888, Oct. 1996.

[10] Jean-Francois Mari and Jean-Paul Haton, "Automatic Word Recognition Based on Second-Order Hidden Markov Models," *Proc. ICSLP*, vol. 1, pp. 247-250, Sep. 1994.

[11] Padma Ramesh and Jay G. Wilpon, "Modeling State Durations in Hidden Markov Models for Automatic Speech Recognition," *Proc. ICASSP*, vol. 1, pp. 381-384, Mar. 1992.