HMM Retraining Based on State Duration Alignment for Noisy Speech Recognition

Wei-Wen Hung and Hsiao-Chuan Wang
Department of Electrical Engineering, National Tsing Hua University,
Hsinchu, Taiwan, 30043, Republic of China
E-mail : hcwang@ee.nthu.edu.tw

Abstract

It is known that incorporating the temporal information of state durations into the HMM can achieve higher recognition performance. However, when a speech signal is contaminated by ambient noises, it is very possible for a state to stay too long or too short in decoding a state sequence even if state durations are adopted in the models. This phenomenon will severely reduce the efficiency of modeling techniques for state durations. To overcome this problem, a proportional alignment decoding **(PAD)** method combining with state duration statistics is proposed and proved experimentally to be effective when the speech signal is distorted by ambient noises. Instead of using Viterbi decoding algorithm, the PAD method is used for state decoding in the retraining phase of a conventional HMM and produce a new set of state duration statistics. This state duration alignment scheme is more efficient to prevent a state from occupying too long or too short in recognition phase.

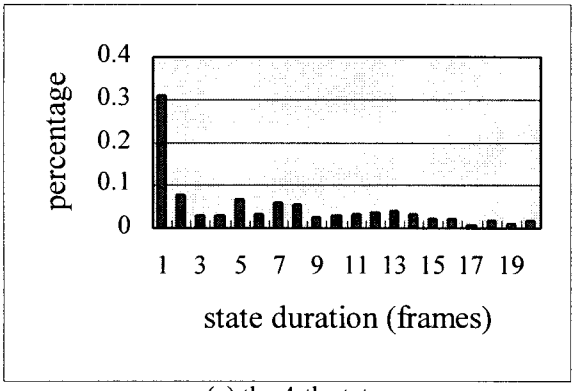
1. Introduction

For speech recognition based on the hidden Markov model, many researches have proved that explicitly modeling the probability density functions (pdf) of state durations can achieve higher recognition performance. Usually, the state durations are modeled in two ways, i.e., nonparametric [1]-[4] and parametric [5]-[7] methods. In nonparametric method, the probability density functions of state durations are estimated via a direct counting procedure on the training data. This approach requires an enormous amount of training utterances in order to reach to a desired degree of accuracy. On the other hand, in parametric method, some continuous probability density functions are used to model the state duration distributions explicitly, and by which only a few parameters are required to completely specify its probability density function. In the hidden semi-Markov model [5] proposed by M. J. Russell and R. K. Moore, they adopted Poisson function as the basic framework for modeling durational structures. Levinson [6] presented a continuously variable duration hidden Markov model (CVDHMM) in which the probability of occupying a state over a specific duration length is governed by the gamma density function. In addition, H. Y. Gu et al. [7] also proposed a hidden Markov model with bounded state durations (HMM/BSD) in which the allowable state durations are constrained by some boundaries. As compared to other approaches, the state durations of HMM/BSD are simply lower and upper bounded in the recognition phase, and can be estimated during the training phase.

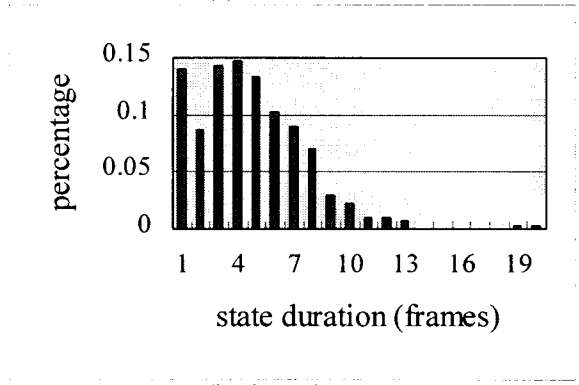
In fact, the distribution of state durations is different from state to state and not confined to a certain type of probability density function. This phenomenon can be verified by the histograms shown in Fig. 1. Fig. 1 shows the histograms of state duration distributions for the 4-th and 6-th states in a conventional HMM for isolated Mandarin digit '5'. From those histograms, we can find that any assumed pdf may not always fit to the practical statistical characteristics of each state in a hidden Markov model. Moreover, it is worth to note that the distributions of allowable state durations are widespread and smooth. This kind of pdf will make it very possible for a state to stay too long or too short in decoding a state sequence. When a speech signal is contaminated by ambient noises severely, an erroneous maximum likelihood state sequence may be obtained even if only a few of its states have extremely high likelihoods while the other states have very poor likelihoods. In this paper, a proportional alignment decoding **(PAD)** method combining with state duration statistics is proposed to retrain a conventional HMM and produce a new set of state duration statistics.

2. HMM with Variable State Duration

When the statistics of state durations are incorporated into a conventional hidden Markov model (**HMM**) [6], the likelihood function of this variable duration HMM (**HMM/DUR**) [6][8] can be defined with the aids of forward and backward likelihoods. Let the forward likelihood $\alpha_t(w,j)$ and backward likelihood



(a) the 4-th state



(b) the 6-th state

Fig. 1. The state duration distributions of the 4-th and 6-th states of HMM for isolated Mandarin digit '5'.

$\beta_t(w,j)$ are defined as

$$\alpha_t(w,j) = p(o_1,o_2,...,o_t,q_t = j|M_w)$$
$$= \sum_{d=1}^{\min\{D(w,j),t\}} \sum_{i=1,i\neq j}^{S_w} \alpha_{t-d}(w,i)\cdot a_{ij}(w)$$
$$\cdot p_{w,j}(d)\cdot \prod_{\tau=1}^{d} b_{w,j}(o_{t-d+\tau}) \quad (1)$$

and

$$\beta_t(w,i) = p(o_{t+1},o_{t+2},...,o_T,q_t = i|M_w)$$
$$= \sum_{d=1}^{\min\{D(w,j),T-t\}} \sum_{j=1,j\neq i}^{S_w} a_{ij}(w)\cdot p_{w,j}(d)$$
$$\cdot \prod_{\tau=1}^{d} b_{w,j}(o_{t+\tau})\cdot \beta_{t+d}(w,j). \quad (2)$$

Then, given a hidden Markov model, the likelihood function of an observation sequence can be modeled as

$$p(O|M_w) =$$
$$\sum_{i=1}^{S_w} \sum_{j=1,j\neq i}^{S_w} \sum_{d=1}^{D(w,j)} \alpha_{t-d}(w,i)\cdot a_{ij}(w)\cdot p_{w,j}(d)$$
$$\cdot \prod_{\tau=1}^{d} b_{w,j}(o_{t-d+\tau})\cdot \beta_t(w,j), \quad (3)$$

where $O = (o_1,o_2,...,o_T)$ denotes the observation sequence, M_w the hidden Markov model for word w with S_w states, $D(w,j)$ the maximum duration within the j-th state of word model M_w, q_t the present state at time t, $a_{ij}(w)$ the state-transition probability from state i to state j of word model M_w, $b_{w,j}(o_t)$ the symbol distribution of o_t in the j-th state of word model M_w, and $p_{w,j}(d)$ the j-th state duration pdf of word model M_w with duration length of d frames. Based on above definition, the derivation of reestimation formulas for the variable duration HMM is formally identical to those for the conventional HMM [8]. Moreover, for a left to right HMM without jumps, the recursive equations listed in [7] are modified to calculate the likelihood, $p(O|M_w)$, and can be summaried as following :

$$for\ d = 1$$
$$\psi_t(w,j,1) =$$
$$\max_{d=1}^{\min\{D(w,j-1),t-1\}} \{\psi_{t-1}(w,j-1,d) + \log[p_{w,j-1}(d)]\}$$
$$+ \log[a_{(j-1)j}(w)] + \log[b_{w,j}(o_t)], \quad (4)$$

$$for\ d \geq 2$$
$$\psi_t(w,j,d) = \psi_{t-1}(w,j,d-1) + \log[b_{w,j}(o_t)], \quad (5)$$

$$p(O|M_w) =$$
$$\max_{d=1}^{T} \{\psi_T(w,S_w,d) + \log[p_{w,S_w}(d)]\}, \quad (6)$$

where $\psi_t(w,j,d)$ represents the maximum likelihood of proceeding from state 1 to state $j-1$ along a state sequence of duration length $(t-d)$ frames and producing the observations $o_1,o_2,...o_{t-d}$, and then staying at the state j and producing the observations $o_{t-d+1},...,o_{t-1},o_t$ at this state. From above recursive equations, we can find that the likelihood function $p(O|M_w)$ may be dominated by the term "$\log[b_{w,j}(o_t)]$" even though the probability density function $p_{w,j}(d)$ for some allowable duration lengths are very low.

3. PAD Method for HMM Retraining

A hidden Markov model which has a more concentrated pdf of state durations may be more efficient

to inhibit a state occupying too long or too short signal frames. In this paper, a proportional alignment decoding (PAD) method instead of the Viterbi decoding algorithm is proposed for state decoding and produce a new set of state duration statistics whose probability density functions are more concentrated.

The PAD method is proceeded as follows. At first, using the segmental k-means algorithm, all of training utterances can be trained into an initial set of word models. Based on those word models and Viterbi decoding algorithm, we can decode each training utterance into state sequence. By using the state sequences decoded for training utterances, we can calculate the state duration mean for each state. For the case of state i in word w, its state duration mean is

$$\bar{d}(w,i) = \frac{1}{N_w}\sum_{j=1}^{N_w} d(w,i,j), \qquad (7)$$

where N_w is the number of training utterances for word w, and $d(w,i,j)$ is the duration of state i in word w for the j-th training utterance. The word duration mean is defined as the accumulation of all the state duration means in a word, and expressed as

$$\bar{d}(w) = \sum_{i=1}^{S_w} \bar{d}(w,i), \qquad (8)$$

where S_w is the number of states in the hidden Markov model of word w. Then the ratio of a state duration in a word is calculated by

$$r(w,i) = \frac{\bar{d}(w,i)}{\bar{d}(w)}. \qquad (9)$$

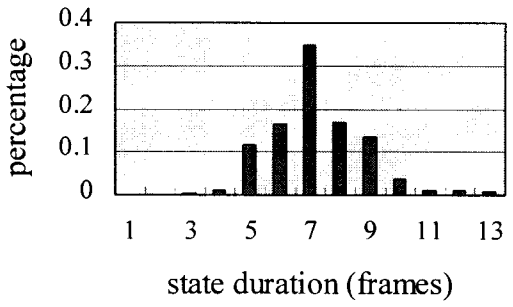
Once we obtain $r(w,i)$ for all word model, the state decoding procedure can be proceeded in a simple way. For example, a training utterance j of word w has duration of $d(w,j)$ frames. We align the duration for each state in this utterance by the following equation

$$\tilde{d}(w,i,j) = \lceil r(w,i) \times d(w,j) - 0.5 \rceil, \qquad (10)$$

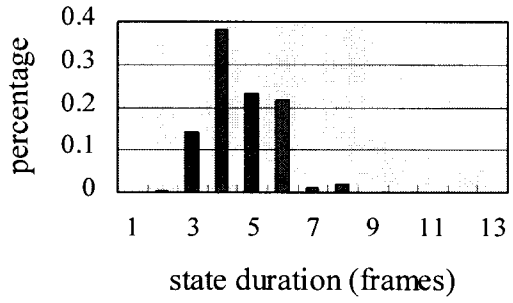
where $\lceil x \rceil$ denotes the smallest integer which is greater than or equal to x.

The proposed retraining scheme has four stages : (1) Use the segmental k-means algorithm and Viterbi decoding method to train a set of conventional HMMs which are used as the initial set of word models. (2) Based on those initial word models, estimate the parameter $r(w,i)$ for every word model. (3) Realign all the training utterances by the PAD method and produce a new set of word models. (4) Collect the state duration statistics $p_{w,j}(d)$ from the new set of word models and use Eqs. (1)-(3) to retrain a final set of word models. The resulted hidden Markov model is denoted as HMM/PAD.

In Fig. 2, the histograms of state duration distributions for the 4-th and 6-th states in the HMM/PAD for isolated Mandarin digit '5' are illustrated. Making a comparison between Fig. 1 and Fig. 2, we can find that the state duration distributions of HMM/PAD are more concentrated than those of the conventional HMM. In addition, the relative probability is raised from 0.15 in Fig. 1(b) to 0.4 in Fig. 2(b).



(a) the 4-th state



(b) the 6-th state

Fig. 2. The state duration distributions of the 4-th and 6-th states of HMM/PAD for isolated Mandarin digit '5'.

4. Experimental Results and Discussions

A multispeaker (50 males and 50 females) isolated Mandarin digit recognition [9] was conducted to verify the effectiveness of the proposed HMM retraining method using the PAD method. There were three sessions of data collection and for each session every speaker uttered a set of 10 Mandarin digits. The first two sessions are used for training the word models and the other for testing. Endpoints are not detected so that each utterance still contains about 0.1~0.5 seconds of pre-silence and post-silence. Each digit is modeled as a left-to-right HMM without jumps in which the output of each state is a Gaussian distribution of feature vectors. Each feature vector consists of 12 LPC derived cepstral coefficients, 12 delta cepstral coefficients, and one delta log-energy. The additive white noise was added to clean speech with predetermined SNRs to generate various noisy speech signals.

Table 1. Comparison of recognition rates

training method	recognition method	clean	20dB	15dB	10dB	5dB	0dB
HMM	Viterbi	97.2	48.8	30.8	19.2	11.2	10.0
HMM	Dur	97.6	62.0	42.8	26.8	20.8	17.6
HMM/DUR	Dur	97.6	67.6	49.2	31.2	24.0	18.4
HMM/PAD	Dur	96.8	72.4	60.0	44.0	29.6	24.8

In our experiments, except the conventional HMM (HMM), two kinds of hidden Markov models are also investigated. The first one is the HMM with variable state duration (HMM/DUR), i.e., using Eq. (1)-(3) to retrain a conventional HMM. The second one is the HMM trained by the proposed retraining scheme (HMM/PAD). Moreover, in the recognition phase, two kinds of state decoding methods are implemented, i.e., Viterbi decoding algorithm (Viterbi) and state decoding using Eq. (4)-(6) (denoted as 'Dur'). The preliminary results of our experiments are shown in Table 1. From this table we can find the following facts : (1) Even without incorporating the statistics of state durations into the training phase of a conventional HMM, state decoding method based on pdf of state durations in the recognition phase can make further improvement in recognition rates. (2) When the information of state durations is taken into account in both training and recognition phases, it can reduce the mismatch between the reference models and testing speech, and obtains higher recognition performance. (3) By means of the PAD method, the HMM/PAD model is more robust to noisy environment and the improvement is remarkable.

5. Conclusions

In this paper, we first demonstrated the practical distributions of state durations in a conventional HMM and pointed out its impacts on speech recognition performance. And then, a new method based on the proportional alignment decoding is proposed to train a new set of hidden Markov models which are more robust to noisy environment. The PAD method enables us to make the pdf of state durations more concentrated, and thus it will be able to prevent a state from staying too long or too short in decoding a state sequence. Experimental results shows that the HMM/PAD model is more robust to ambient noises than those hidden Markov models with or without incorporating durational model.

References

[1] L. R. Rabiner, B. H. Juang, S. E. Levinson, and M. M. Sondhi, "" Recognition of isolated digits using hidden Markov models with continuous mixture densities, "" AT&T Tech. J. , vol. 64, no. 6, pp. 1211-1234, July-Aug. 1985.

[2] B. H. Juang and L. R. Rabiner,"" Mixture autoregressive hidden Markov models for speech signals, "" IEEE Trans. Acoust., Speech, Signal Processing, vol. 33, no. 5, pp. 1404-1413, 1985.

[3] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "" High performance connected digit recognition, using hidden Markov models, "" in Proc. ICASSP (New York), 1988, pp. 119-122.

[4] Kevin, Power. ""Durational Modelling for Improved Connected Digit Recognition"", Int. Conf. Spoken Language Processing, pp. 885-888, 1996.

[5] M. J. Russell and R. K. Moore, "" Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition, "" in Proc. ICASSP, pp. 5-8, 1985.

[6] S. E. Levinson, ""Continuously variable duration hidden Markov models for speech analysis, "" in Proc. ICASSP, pp. 1241-1244, 1986.

[7] H. Y. Gu, C. Y. Tseng and L. S. Lee, "" Isolated - Utterance Speech Recognition Using Hidden Markov Models with Bounded State Durations, "" IEEE Trans. on Signal Processing, vol. 39, no.8 , pp. 1743-1752, August 1991.

[8] L. Rabiner, B. H. Juang, : Fundamentals of Speech Recognition, Prentice Hall International Editions, 1993, pp. 342-361.

[9] L. M. Lee and H. C. Wang, "" A study on adaptations of cepstral and delta cepstral coefficients for noisy speech recognition, "" Proc. ICSLP, 1994, pp. 1011-1014.