HIDDEN SEMI-MARKOV MODEL BASED SPEECH RECOGNITION SYSTEM USING WEIGHTED FINITE-STATE TRANSDUCER

Keiichiro Oura, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda

Nagoya Institute of Technology
Department of Computer Science and Engineering,
Gokiso-cho, Showa-ku, Nagoya, 466-8555 Japan
{ uratec, zen, nankaku, ri, tokuda } @ics.nitech.ac.jp

ABSTRACT

In hidden Markov models (HMMs), state duration probabilities decrease exponentially with time. It would be inappropriate representation of temporal structure of speech. One of the solutions for this problem is integrating state duration probability distributions explicitly into the HMM. This form is known as a hidden semi-Markov model (HSMM) [1]. Although a number of attempts to use explicit duration models in speech recognition systems have been proposed, they are not consistent because various approximations were used in both training and decoding.

In the present paper, a fully consistent speech recognition system based on the HSMM framework is proposed. In a speaker-dependent continuous speech recognition experiment, HSMM-based speech recognition system achieved about 5.9% relative error reduction over the corresponding HMM-based one.

1. INTRODUCTION

Hidden Markov models (Fig. 1(a)) have formed the basis for many speech recognition systems since 1970's. The advantages of using the HMM are that i) it can represent speech as probability distributions, ii) it is robust, iii) efficient algorithms for estimating its model parameters are provided.

However, a number of limitations of the HMM for modeling speech have been reported [2]. One of the major limitations is its duration modeling. In the HMM, state duration probabilities are implicitly modeled by its state transition probabilities: state duration probabilities decrease exponentially with time. This would be inappropriate representation of temporal structure of speech.

One of the solutions for this problem is to integrate state duration probability distributions explicitly into the HMM. This model is known as a hidden semi-Markov model (HSMM) which is illustrated in Fig. 1(b). Although a variety of attempts to include the explicit duration models in speech recognition system have been reported [3, 4], they are not consistent because various approximations listed below were used both in training and decoding:

a) State duration probability distributions were estimated from statistics calculated by the forward-backward algorithm of the HMM, not of the HSMM [5].

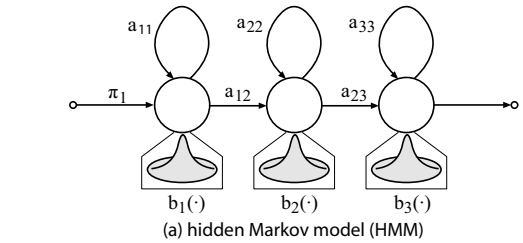

Fig. 1. Examples of an HMM and HSMM with 3-state left-to-right with no skip structures.

b) Context-independent model or context dependent parameter tying structure of state output probability distributions was used [3].

c) State duration models were not used in decoding process. Rescoring of N-best hypotheses generated by the HMMs using the HSMM likelihood was performed [4].

In the present paper, we avoid the above approximations and construct a fully consistent HSMM-based speech recognition system. For approximation ***a)***, both state output and duration probability distributions are estimated based on the HSMM statistics calculated by the generalized forward-backward algorithm [1, 2]. For ***b)***, state output and duration probability distributions are individually clustered by phonetic decision trees [6]. For ***c)***, a speech decoder for the HSMM is constructed using Weighted Finite-State Transducers (WFSTs).

The rest of the present paper is organized as follows. Section 2 describes training algorithms for the HSMM. Section 3 shows context clustering for state duration probability distributions. Section 4 presents a speech decoder for the HSMM using the WFSTs. Results of a speech recognition experiment

is shown in Section 5. Finally, concluding remarks and future plans are presented in Section 6.

2. HIDDEN SEMI MARKOV MODEL

2.1. Generalized forward-backward algorithm

The output probability of an observation vector sequence $\boldsymbol{o}$ from an HSMM Λ can be computed efficiently using the generalized forward-backward algorithm [1, 2]. The partial forward probabilities $\alpha_t(\cdot)$ and partial backward probabilities $\beta_t(\cdot)$ are defined as follows:

$$\alpha_0(j) = \pi_j, \tag{1}$$

$$\begin{aligned}\alpha_t(j) &= P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_t, q_t = j \mid q_{t+1} \neq j, \Lambda) \\ &= \sum_{d=1}^{t} \sum_{\substack{i=1, \\ i \neq j}}^{N} \alpha_{t-d}(i) a_{ij} p_j(d) \prod_{s=t-d+1}^{t} b_j(\boldsymbol{o}_s),\end{aligned} \tag{2}$$

$$\beta_T(i) = 1, \tag{3}$$

$$\begin{aligned}\beta_t(i) &= P(\boldsymbol{o}_{t+1}, \ldots, \boldsymbol{o}_T, q_t = i \mid q_{t+1} \neq i, \Lambda) \\ &= \sum_{d=1}^{T-t} \sum_{\substack{j=1, \\ j \neq i}}^{N} a_{ij} p_j(d) \prod_{s=t+1}^{t+d} b_j(\boldsymbol{o}_s) \beta_{t+d}(j),\end{aligned} \tag{4}$$

where a_{ij}, $b_j(\boldsymbol{o}_t)$, N, $p_j(d)$, and π_j are a state transition probability from the i-th state to the j-th state, a state output probability of an observation vector $\boldsymbol{o}_t$ from the j-th state, the total number of states, a state duration probability of the j-th state, and an initial state probability of the j-th state, respectively. From the above equations, the output probability of the observation vector sequence $\boldsymbol{o} = \{\boldsymbol{o}_1, ..., \boldsymbol{o}_T\}$ from the HSMM Λ is given by

$$P(\boldsymbol{o} \mid \Lambda) = \sum_{i=1}^{N} \sum_{\substack{j=1, \\ j \neq i}}^{N} \sum_{d=1}^{t} \alpha_{t-d}(i) a_{ij} p_j(d) \prod_{s=t-d+1}^{t} b_j(\boldsymbol{o}_s) \beta_t(j). \tag{5}$$

2.2. Parameter reestimation formulas

In the present paper, we assume that each state output probability $b(\cdot)$ is represented by a mixture of Gaussian distributions. Parameter reestimation formulas of the mixture weight w_{jg}, mean vector $\boldsymbol{\mu}_{jg}$ and covariance matrix $\boldsymbol{\Sigma}_{jg}$ of the g-th mixture of the j-th state are given by

$$\overline{w}_{jg} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j, g)}{\sum_{h=1}^{G} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j, h)}, \tag{6}$$

$$\overline{\boldsymbol{\mu}}_{jg} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \zeta_t^d(j, g)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j, g)}, \tag{7}$$

$$\overline{\boldsymbol{\Sigma}}_{jg} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \eta_t^d(j, g)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j, g)}, \tag{8}$$

respectively, where G is the number of Gaussian distributions, $\gamma_t^d(j, g)$, $\zeta_t^d(j, g)$ and $\eta_t^d(j, g)$ are occupancy probability, first and second order statistics, which are calculates as

$$\begin{aligned}\gamma_t^d(j, g) = \frac{1}{P(\boldsymbol{o} \mid \Lambda)} &\sum_{i=1}^{N} \alpha_{t-d}(i) a_{ij} p_j(d) \beta_t(j) \\ &\sum_{s=t-d+1}^{t} w_{jg} \mathcal{N}(\boldsymbol{o}_s \mid \boldsymbol{\mu}_{jg}, \boldsymbol{\Sigma}_{jg}) \prod_{\substack{k=t-d+1, \\ k \neq s}}^{t} b_j(\boldsymbol{o}_k),\end{aligned} \tag{9}$$

$$\begin{aligned}\zeta_t^d(j, g) = \frac{1}{P(\boldsymbol{o} \mid \Lambda)} &\sum_{i=1}^{N} \alpha_{t-d}(i) a_{ij} p_j(d) \beta_t(j) \\ &\sum_{s=t-d+1}^{t} w_{jg} \mathcal{N}(\boldsymbol{o}_s \mid \boldsymbol{\mu}_{jg}, \boldsymbol{\Sigma}_{jg}) \prod_{\substack{k=t-d+1, \\ k \neq s}}^{t} b_j(\boldsymbol{o}_k) \boldsymbol{o}_s,\end{aligned} \tag{10}$$

$$\begin{aligned}\eta_t^d(j, g) = \frac{1}{P(\boldsymbol{o} \mid \Lambda)} &\sum_{i=1}^{N} \alpha_{t-d}(i) a_{ij} p_j(d) \beta_t(j) \\ &\sum_{s=t-d+1}^{t} w_{jg} \mathcal{N}(\boldsymbol{o}_s \mid \boldsymbol{\mu}_{jg}, \boldsymbol{\Sigma}_{jg}) \\ &\prod_{\substack{k=t-d+1, \\ k \neq s}}^{t} b_j(\boldsymbol{o}_k) [\boldsymbol{o}_s - \mu_{jg}][\boldsymbol{o}_s - \mu_{jg}]^\top,\end{aligned} \tag{11}$$

respectively.

The state duration probability distribution of the j-th state is modeled by a single Gaussian distribution with mean ξ_j and variance σ_j^2 i.e. $p_j(d_j) = \mathcal{N}(d_j \mid \xi_j, \sigma_j^2)$. Although a gamma distribution could be used , a Gaussian distribution is used in the present paper. The reestimation formulas of these parameters are given as follows:

$$\overline{\xi}_j = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(j)(t_1 - t_0 + 1)}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(j)}, \tag{12}$$

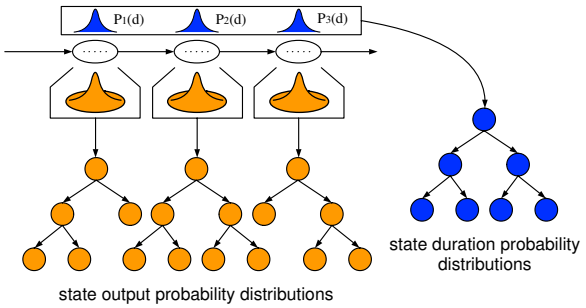

Fig. 2. Context clustering that separate state duration probability distributions from HMM parameters.

$$\overline{\sigma}_j^2 = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(j)(t_1-t_0+1)^2}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(j)} - (\overline{\xi}_j)^2, \quad (13)$$

where

$$\chi_{t_0,t_1}(j) = \frac{1}{P(\boldsymbol{o}\mid\Lambda)}\sum_{i\neq j}\alpha_{t_0-1}(i)a_{ij}\prod_{s=t_0}^{t_1}b_j(\boldsymbol{o}_s)p_j(t_1-t_0+1)\beta_{t_1}(j). \quad (14)$$

3. CONTEXT DEPENDENT DURATION MODELING

In the conventional speech recognition system using explicit duration models, context-independent duration model or the same parameter tying structure that of state output probability distributions was used [3]. However, it is generally considered that state output and duration probability distributions have different context-dependency. In the present paper, we adopt context-dependent duration modeling technique used in the HMM-based speech synthesis [5]. The state duration probabilities of the each HSMM are modeled by single multivariate Gaussian distributions whose dimensionality is equal to the number of states of the HSMM. State output and duration probability distributions are context-dependent and they are clustered separately by the phonetic dicision trees [6] (Fig. 2).

4. WEIGHTED FINITE-STATE TRANSDUCERS FOR SPEECH RECOGNITION

Finite-state machines have been used in many areas of computational linguistics. These transducers appear as very interesting in speech processing. Weighted finite-state transducers associate weights such as probabilities, duration, penalties, or any other quantity that accumulates linearly along paths, to each pair of input and output symbol sequences. It offers a unified framework representing various model used in speech and language processing [7, 8]. An integrated WFST for speech recognition can be represented as

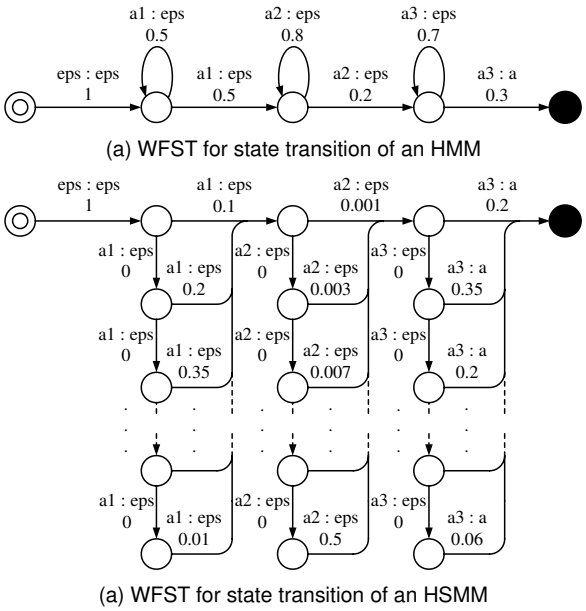

Fig. 3. WFST for state transition of HMM and HSMM.

$$H \circ C \circ L \circ G, \quad (15)$$

where H, C, L, and G are WFSTs for a state transitions network, a context-dependent model mapping, a pronunciation lexicon, and a language model, respectively.

Advantages of using WFSTs for speech decoder are as follows:

- It offers combining component individually designed.
- Each component can be individually optimized.
- The decoder offer easy managing, because network and decoder are constructed individually.

However, in the case of using a large model or an huge transducer is usually generated by composing all the components. Accordingly, both the amount of computation and the memory usage in decoding increase even if the WFST is optimized. To avoid this problem, the on-the-fly composition [9, 10] is applied. In the on-the-fly composition, the set of WFSTs are separated into two or more groups, and in each groups on WFST is composed and optimized. Composition between the groups is performed during decoding if necessary. For example, Fig. 3 represent state transition of an HMM and an HSMM represented by WFSTs. All arcs of Fig. 3(a), and Fig. 3(b) are weighted by state transition probabilities, and state duration probabilities. The state duration in Fig. 3(b) is limited, because infinite state duration are not allowed in WFSTs.

5. EXPERIMENTS

To evaluate the peformance of HSMM-based ASR system, a speaker-dependent continuous speech recognition experi-

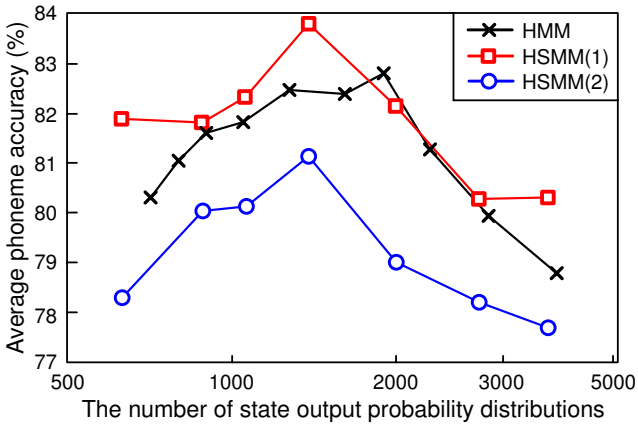

Fig. 4. Experimental results.

ment was conducted on the ATR Japanese speech database (phonetically-balanced sentences). Four handred and fifty sentences spoken by a male speaker MHT were used for training. The test data consists of a total of 53 utterances from the same speaker that are not included in the training data.

The speech data was down-sampled from 20kHz to 16kHz, windowed at a 5-ms frame rate using a 25-ms Blackman window, and parameterized into 24 mel-cepstral coefficients with a mel-cepstral analysis technique. Static coefficients including the zero-th coefficients and their first and second derivatives were used as feature parameters. Three-state left-to-right structures were used and 118 questions about left and right phonetic contexts were prepared in decision tree construction. Each state output probability distribution was modeled by a single Gaussian distribution with a diagonal covariance matrix.

The phonetic decision tree-based context clustering [6] was applied for state output and duration probability distributions separately. The MDL criterion was used for stopping tree growth [11]. We changed the weight for penalty term of c Eq. (9) in [11] and constructed acoustic models with the various number of parameters. The same weight c was used for clustering both state output and duration probability distributions. Thus, the number of state duration probability distributions has changed according to the number of state output probability distributions. To evaluate the effect of context-dependency of the state duration probability distributions, we also constructed HSMMs with context-independent state duration probability distributions. Figure 4 shows the experimental results. Horizontal axis presents the number of state output probability distributions, and vertical axis shows average phoneme accuracy. In this figure, HMM, HSMM(1), HSMM(2) show HMM-based system, HSMM-based system with triphone state duration probability distributions, and HSMM-based system with monophone state duration probability distributions, respectively. The HSMM(1) achieved about 5.9% error reduction over the HMM. Comparing HSMM(1) and HSMM(2), Context dependency of state duration probability distribution can be confirmed.

6. CONCLUSIONS

In the present paper, we constructed a fully consistent HSMM-based speech recognition system , and evaluted its performance while avoiding several approximations. As the result, obvious improvement of phoneme accuracy was confirmed by modeling state duration probability distribution with context dependence. Future work includes evaluations on large vocabulary continuous speech recognition tasks

7. REFERENCES

[1] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, pp. 29-45, 1986.

[2] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMMs to segment models," *IEEE Transactions on Speech & Audio Processing*, vol. 4, no. 5, pp. 360-378, 1996.

[3] Myoung-Wann Koo, Sung-Joon Park, Dan-Young Son, "Context dependent phoneme duration modeling with tree-based state tying," *Proc. INTERSPEECH2004*, vol. 1, pp. 721-724, 2004.

[4] V. R. R. Gaddem, "Modeling Word Duration," *Proc. ICSLP2000*, vol. 1, pp. 601-604, 2000.

[5] Yoshimura, Tokuda, Masuko, Kobayashi, Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Proc. EUROSPEECH*, vol. 5, pp. 2347-2350, 1999.

[6] J. Odell, "The use of context in large vocabulary speech recognition," *Ph. D. thesis*, Cambridge University, 1995.

[7] M. Mohri, F. Pereira,M. Riley, "Weighted Finite-State Transducers In Speech Recognition," *Proc. ASR2000*, pp. 97-106, 2000.

[8] C. Allauzen, M. Mohri, "Generalized Optimization Algorithm For Speech Recognition Transducers," *Proc. ICASSP2003*, vol. 1, pp. 352-355, 2003.

[9] H. J. G. A. Dolfing, I. L. Hetherington, "Incremental language models for speech recognition using finite-state transducers," *Proc. ASRU2001*, 2001.

[10] D. Willett, S. Katagiri, "Recent advances in efficient decoding combining on-line transducer composition and smoothed language model incorporation," *Proc. ICASSP2002*, vol. 1, pp. 713-716, 2002.

[11] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," *Proc. EUROSPEECH*, vol. 1, pp. 99-102, 1997.