HYBRID MODELING OF PHMM AND HMM FOR SPEECH RECOGNITION

Tetsuji Ogawa and Tetsunori Kobayashi

Dept. EECE, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
{ogawa,koba}@tk.elec.waseda.ac.jp

ABSTRACT

A hybrid acoustic model of Partly Hidden Markov Model (PHMM) and HMM is proposed.

PHMM was proposed in our previous work to deal with the complicated temporal changes of acoustic features. It can realize the observation dependent behaviors in both observations and state transitions. It achieved good performance but some errors with different trend from HMM still remained.

In this paper, we designed a new acoustic model on the basis of PHMM, in which the observation and state transition probabilities are defined by the geometric means of PHMM-based ones and HMM-based ones. In this framework, if a word hypothesis is given a low score by either PHMM or HMM, it almost loses possibilities to be a probable candidate. Since many errors are due to the high-scores of incorrect categories rather than the low-score of the correct category, this property contributed to reduce errors. Moreover, the proposed model is more stable than PHMM because the higher order statistics of PHMM, which is generally accurate but sometimes less reliable, is smoothed by the lower order statistics of HMM, which is not so accurate but robust.

Experimental results showed the effectiveness of proposed model: it reduced the word errors by 25% compared with HMM.

1. INTRODUCTION

Many efforts have been made toward better acoustic models for speech recognition [1] [2] [3] [4] [5] [6] [7]. We also proposed Partly Hidden Markov Model (PHMM) aiming at treating the more complicated temporal changes of feature parameters. The previous papers revealed the effectiveness of PHMM for the modeling of speech and also gestures [8] [9]. In the framework of PHMM, the pair of the hidden state (H-state) and the observable state (O-state) determines the stochastic phenomena of not only current observations but also next state transitions. The existence of O-state realizes the observation-dependent behavior (not state-dependent piecewise-stationary behavior) in both observations and state transitions.

PHMM achieved better performance than HMM but some errors with different trend from HMM still remained. In this paper, we attempt to improve the performance of PHMM by utilizing this different trend of errors in PHMM and HMM.

Many errors in general recognizers, which usually adopt maximum likelihood (ML) estimation, are due to the high-scores of incorrect categories rather than the low-score of the correct category. The score of correct category is properly high even if the case of errors. The error arises when some scores of incorrect categories are unduly high. This is quite natural in ML estimation paradigm because model parameters are estimated so as to assure the high score for the data of the correct category.

The fact above implies that it is effective to use many recognizers with different trend of errors and regard the word hypothesis as the answer only if all recognizers give high scores to the hypothesis. In this paper, we propose a new hybrid acoustic model named Smoothed Partly Hidden Markov Models (SPHMM) based on PHMM and HMM. In the proposed model, the observation and the state transition probabilities are defined by geometric mean of PHMM-based ones and HMM-based ones.

This framework is also expected to have the following merit. PHMM is sometimes less reliable without sufficient training data, since the model uses higher order statistics than HMM. In this sense, the proposed model is more stable than PHMM because the higher order statistics of PHMM, which is generally accurate but sometimes less reliable, is smoothed by the lower order statistics of HMM, which is not so accurate but robust.

In the next section, PHMM is briefly surveyed as the base of the proposal in this paper. In section 3, the basic idea and the formulation of the new stochastic model, SPHMM, is proposed. Finally in section 4, the evaluation results for continuous speech of the proposed model are shown.

2. PARTLY HIDDEN MARKOV MODEL

In the proposed model, the observation probability and the state transition probability are conditioned by the two states s_t^h, s_t^o as follows:

observation probability : $P_r(x_t|s_t^h\, s_t^o)$
transition probability : $P_r(s_{t+1}^h|s_t^h\, s_t^o)$, $P_r(s_{t+1}^o|s_t^h\, s_t^o)$

We call s_t^h H-state (Hidden state). And we call s_t^o O-state (Observable state). We call this model "Partly Hidden Markov Model (PHMM)."

If both of these states, s_t^h, s_t^o, are uniquely determined from past observations, this model is equivalent to Markov Model. If both of them are stochastically determined, it is equivalent to Hidden Markov Model.

In PHMM, different combination of H-state and O-state are used to determine the observation probability and the state transition probability. In order to realize this framework, we introduce two kinds of observable states, OO-State (s_t^{oo}) and OS-State (s_t^{os}). Here, s_t^{oo} is used to determine the observation probability, and s_t^{os} is used to determine the state transition one. The state-observation dependency graph in this framework is shown in Fig. 1. In PHMM, s_t^{oo} and s_t^{os} are observable from any past observations.

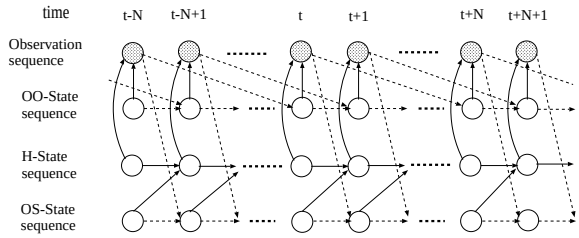

Fig. 1. Dependency of the observation sequence and the state transition sequence in PHMM. Solid lines express probabilistic dependencies and dotted lines express deterministic dependencies. s_t^{oo} is observable only from N frame previous observation and s_t^{os} is observable only from last observation.

In this paper, we adopt the framework, in which s_t^{oo} is observable only from N frame previous observation x_{t-N} and s_t^{os} is observable only from last observation x_{t-1}. The simplified dependency graph of PHMM using the above relation $s_t^{oo} = x_{t-N}$, $s_t^{os} = x_{t-1}$ is shown in Fig. 2.

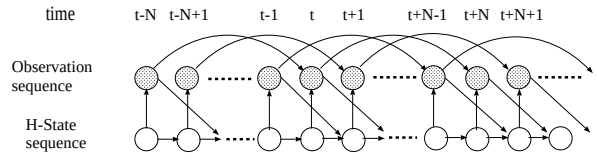

Fig. 2. The simplified dependency graph of PHMM using the relation $s_t^{oo} = x_{t-N}$, $s_t^{os} = x_{t-1}$.

HMM can deal with only the state-dependent piecewise-stationary behavior in both observations and state transitions. While, in PHMM, the pair of the hidden state and the observable state determines the stochastic phenomena of not only current observations but also next state transitions. By introducing O-state, the observation-dependent behavior is realized in both observations and state transitions. Therefore, the model can deal with more complicated process than piecewise stationary. In the PHMM, the probability that the observation sequence $x_1, x_2, \cdots, x_T$ comes from the model with the H-state transition $s_1^h, s_2^h, \cdots, s_T^h$ is defined by the following equation.

$$\begin{aligned} P_s &= P_r(x_1, x_2, \cdots, x_T, s_1^h, s_2^h, \cdots, s_T^h, s_1^o, s_2^o, \cdots, s_T^o) \\ &= P_r(s_1^h, x_{-N+1}) P_r(x_1|s_1^h, x_{-N+1}) \\ &\quad \cdot \prod_{t=1}^{T-1} P_r(s_{t+1}^h|s_t^h, x_{t-1}) P_r(x_{t+1}|s_{t+1}^h, x_{t-N+1}) \\ &= P_r(s_1^h) P_r(x_{-N+1}, x_1|s_1^h) \\ &\quad \cdot \prod_{t=1}^{T-1} \frac{P_r(s_{t+1}^h|s_t^h) P_r(x_{t-1}|s_t^h, s_{t+1}^h)}{P_r(x_{t-1}|s_t^h)} \\ &\quad \cdot \frac{P_r(x_{t-N+1}, x_{t+1}|s_{t+1}^h)}{P_r(x_{t-N+1}|s_{t+1}^h)} \end{aligned} \tag{1}$$

$P_r(x_1, x_2, \cdots, x_T)$ can be obtained by summing up Eq.(1) for all possible combinations of H-state transition $s_1^h, s_2^h, \cdots, s_T^h$.

From the above discussion, it is found that PHMM can be expressed by following 6 parameters.

$\pi_i = P_r(s_1^h = S_i^h)$:
the probability that the initial H-state is S_i^h.

$a_{ij} = P_r(s_{t+1}^h = S_j^h|s_t^h = S_i^h)$:
the probability that the next H-state is S_j^h in the case that the current H-state is S_i^h.

$b_i(x_{t-1}) = P_r(x_{t-1}|s_t^h = S_i^h)$:
the probability that the last observation is x_{t-1} in the case that the current H-state is S_i^h.

$c_{ij}(x_{t-1}) = P_r(x_{t-1}|s_t^h = S_i^h, s_{t+1}^h = S_j^h)$:
the probability that the last observation is x_{t-1} in the case that the current H-state is S_i^h and the next H-state is S_j^h.

$d_i(x_{t-N}) = P_r(x_{t-N}|s_t^h = S_i^h)$:
the probability that the N frames previous observation is x_{t-N} in the case that the current H-state is S_i^h.

$e_j(x_{t-N}, x_t) = P_r(x_{t-N}, x_t|s_t^h = S_j^h)$:
the probability that the current observation is x_t and the N frames previous observation is x_{t-N} in the case that the current H-state is S_j^h.

3. SPHMM: A HYBRID MODEL OF PHMM AND HMM

In the framework of maximum likelihood estimation, the model parameters are estimated so as to assure the high score for all data of the correct category. Since preciseness of the model expression cannot be perfect, a slight difference occurs between the area of feature parameter where the model gives high score (high-scored-area) and the area where the data of correct category is actually distributed (correct-category-area). So, the model gives unfairly high score to some data of incorrect category also. This is one of the substantial causes of recognition errors.

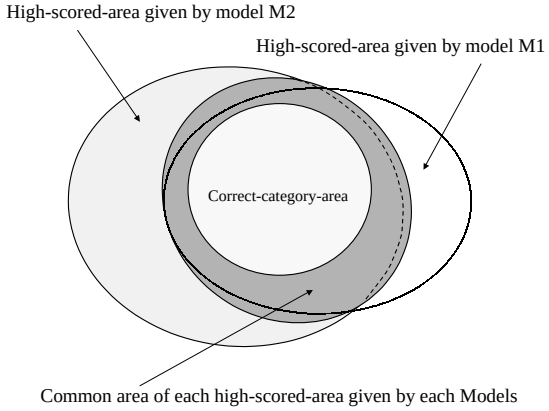

Fig. 3. The conceptual image of effectiveness of Hybrid Model.

However, the areas with unfairly high score (unfairly-high-scored-area) are dependent on the recognizers. Thus, as shown in Fig.3, it is thought that we can remove unfairly-high-scored-areas and extract the closer area to the correct-category-area by extracting the common area of each high-scored-area given by each recognizer.

From above point of view, we propose integration of the state transition and the observation probabilities of PHMM and those of HMM.

We utilize geometric mean to extract the common area of high-scored-areas. The value of geometric mean remains high only if the all values are high. By replacing the observation and the state transition probabilities of PHMM by the geometric mean of those of PHMM and those of HMM, Eq.(1) becomes,

$$
\begin{aligned}
P_s &= P_r(s_1^h)\cdot P_r(x_{-N+1}, x_1|s_1^h) \\
&\cdot \prod_{t=1}^{T-1}\left\{\left(\frac{P_r(s_{t+1}^h|s_t^h)P_r(x_{t-1}|s_t^h, s_{t+1}^h)}{P_r(x_{t-1}|s_t^h)}\right)^{w_t}\right. \\
&\left.\cdot P_r(s_{t+1}^h|s_t^h)^{(1-w_t)}\right\}\cdot\left\{\left(\frac{P_r(x_{t-N+1}, x_{t+1}|s_{t+1}^h)}{P_r(x_{t-N+1}|s_{t+1}^h)}\right)^{w_o}\right. \\
&\left.\cdot P_r(x_{t+1}|s_{t+1}^h)^{(1-w_o)}\right\} \qquad (2)
\end{aligned}
$$

We call this model "Smoothed Partly-Hidden Markov Model." In Eq.(2), w_t is a smoothing weight of the state transition probability and w_o is that of the observation probability. By changing these smoothing weights w_t and w_o, various models can be represented. Those are shown in Table 1.

If parameters of PHMM and HMM are estimated independently before they are integrated, it means that the parameter estimation of SPHMM is performed based on the assumption that the state transitions of PHMM and HMM are differ. The formulation of Eq.(2) is requiring to integrate the observation and the state transition probability of PHMM and those of HMM mutually. With such a method, the optimality as a parameter of the stochastic model formulized by the Eq.(2) is not assured. Thus, we let smoothing weight w_t and w_o in Eq.(2) be a constant, and training of SPHMM is performed by applying EM algorithm for every smoothing weight.

Table 1. The relation between smoothing weight and models.

w_t	w_o	Model
0	0	HMM
0	1	conditional HMM
1	1	PHMM
$0 \le w_t \le 1$	$0 \le w_o \le 1$	SPHMM

This integration is also interesting from the view point of gaining the reliability of model. In HMM, the observation probability and the state transition one in each state are independent of the previous observations. While in PHMM, those are represented by conditional probability of the previous observations. Because of this complexity of the structure, PHMM is sometimes less reliable without sufficient training data. The hybrid structure proposed here can be regarded as the smoothing of higher statistics in PHMM with the lower order statistics in HMM. This is similar to the well known example that the trigram language model smoothed with bigram performs better than simple trigram.

4. CONTINUOUS SPEECH RECOGNITION EXPERIMENT

In order to evaluate the effectiveness of SPHMM for speech recognition, continuous speech recognition experiment was done.

4.1. Experimental Setup

Training data and test data are represented by 12 MFCCs, power, delta MFCCs and delta power, sampled every 10ms.

The acoustic models we used are trained with 20414 sentences from the ASJ speech database of phonetically balanced sentences (ASJ-PB) and newspaper article sentences (ASJ-JNAS)[11]. We adopted the demi-syllable models, which consist of the on-glide and the stationary parts (no off glide), because PHMM is excellent at expressing the transitional part. The distribution function of each state in models is represented by a normal distribution with full covariance.

We use 3-gram language models, which were constructed using the lexicon of 20K vocabulary size. The vocabulary set consists of the most frequent words in Mainichi newspaper articles from Jan. 1991 to Sep. 1994.

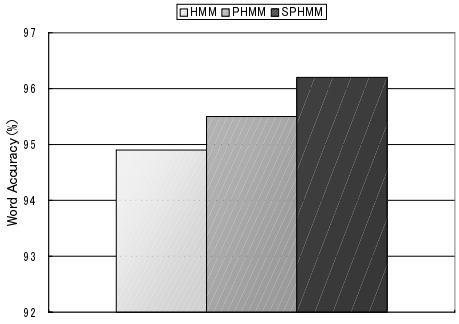

Fig. 4. Word accuracy of HMM, PHMM and SPHMM.

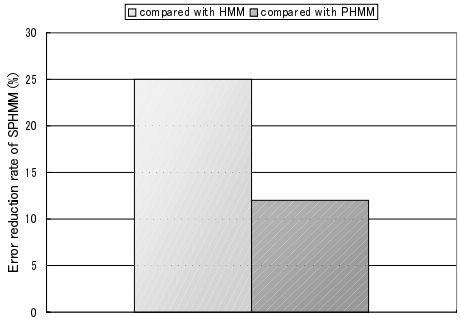

Fig. 5. Error reduction rate of SPHMM.

For the evaluation, we used 100 utterances of ASJ-JNAS speech database which are not used for training of the acoustic models. We picked up 20 speakers and 5 utterances per speaker.

Since the ASJ speech database has no phonetic labels, we automatically labeled ASJ data using HMM and used them for the training of initial parameters. Then, the final parameters of models are obtained by repeating re-estimation by EM algorithm. The recognition engine we used for the evaluation performs one-pass search using word 3-gram. A frame synchronous beam search is performed in search algorithm using the tree-structured lexicon which makes a phoneme the unit of the node.

4.2. Experimental Results

We evaluate the effectiveness of smoothing only of the observation probability of PHMM, because the performance of the state transition probability of PHMM is sufficient and effectiveness of smoothing is not appeared. Here, frame intervals of inter-frame observation correlation (given by N in Eq.(1)) of PHMM are fixed to the value with which the model gave the best score. From the preliminary experiments, PHMM gives the best score (95.5%) when the frame interval equals 7 frames.

Fig. 4 shows the word accuracy of HMM, PHMM, and SPHMM. Fig. 5 shows the error reduction rate of SPHMM compared with HMM and PHMM.

Here, SPHMM gives the best score (96.2%) when smoothing weight w_o equals 0.2. This is 25% reduction in error rate compared with HMM (94.9%). This improvement represents the effectiveness of conditioning the state transition and the observation probabilities by the previous feature observations. SPHMM reduced errors by 12% compared with PHMM (95.5%). This improvement represents the effectiveness of integrating the observation probability of PHMM and those of HMM.

5. CONCLUSION

We proposed Smoothed Partly-Hidden Markov Model (SPHMM), in which probabilities of PHMM are smoothed with those of HMM, and evaluated effectiveness of SPHMM using continuous speech recognition. The performance was improved by smoothing of observation probabilities of PHMM and those of HMM. SPHMM reduced word error by 25% compared with HMM.

In the next stage, we would like to apply SPHMM to the spontaneous speech.

6. REFERENCES

[1] L. Deng, M. Aksmanovic, “Speaker-Independent phonetic classification using Hidden Markov Models with mixture of trend functions,” IEEE trans on Speech and Audio Processing, vol. 5, pp.319-324, JULY. 1997.

[2] J-F. Mari, J-P. Haton, “Automatic word recognition based on second-oder hidden Moarkov models,” IEEE Trans. on Speech and Audio Process, vol.5, n.1, Jan. 1997.

[3] Y. Ariki, “Mixture density HMMs with two-level transition,” Journal of Acoustic Society Japan(E), vol.14, no.4, pp.279-280, Sep. 1993.

[4] C.J. Wellekens, “Explicit correlation in Hidden Markov Model with optimal inter-frame dependence,” Proc. ICASSP87, pp.383-386, 1987.

[5] S.Takahashi, T.Matsuoka, Y.Minami and K.Shikano, “Phoneme HMMs constrained by frame correlations,” Proc. ICASSP93, pp.219-222, 1993.

[6] V. Digalakis, M. Ostendorf and J.R. Roglicek, “Improvement of the stochastic segment model for phoneme recognition,” Proc. DARPA Workshop on Speech and Natural Language, pp.332-338, 1989.

[7] J.A.Bilmes, “Buried Markov Models for speech recognition,” Proc. ICASSP99, pp713-716, Mar. 1999.

[8] Tetsuji Ogawa, Tetsunori Kobayashi, “Generalization of State-Observation-Dependency in Partly Hidden Markov Models,” Proc. ICSLP2002, pp.2673-2676, Sep. 2002.

[9] T.Kobayashi, S.Haruyama, “Partly Hidden Markov Model and its Application to Gesture Recognition,” Proc. ICASSP97, pp.3081-3084, April. 1997.

[10] T.Kawahara et al., “Sharable software repository for Japanese large vocabulary continuous speech recognition,” Proc. ICSLP98, pp.3257-3260, Sep. 1998.

[11] K.Itou et al., “The Design of the Newspaper-Based Japanese Large Vocabulary Continuous Speech Recognition Corpus,” Proc. ICSLP98, pp.3261-3264, Nov. 1998.