ASYNCHRONOUS-TRANSITION HMM

Shigeki Matsuda, Mitsuru Nakai, Hiroshi Shimodaira and Shigeki Sagayama

Japan Advanced Institute of Science and Technology
Tatsu-no-Kuchi, Ishikawa, 923-1292 Japan
E-mail: `{matsuda,mit,sim,sagayama}@jaist.ac.jp`

ABSTRACT

We propose a new class of hidden Markov model (HMM) called asynchronous-transition HMM (AT-HMM). Opposed to conventional HMMs where hidden state transition occurs simultaneously to all features, the new class of HMM allows state transitions asynchronized between individual features to better model asynchronous timings of acoustic feature changes. In this paper, we focus on a particular class of AT-HMM with sequential constraints based on a novel concept of "state tying along time". To maximize the advantage of the new model, we also introduce feature-wise state tying technique. Speaker-dependent speech recognition experiments demonstrated error reduction rates more than 30% and 50% in phoneme and isolated word recognitions, respectively, compared with conventional HMMs.

1. INTRODUCTION

Conventional Hidden Markov Models (HMMs) for speech recognition implicitly assume that individual acoustic feature parameters change their statistical properties simultaneously as the result of treating the feature parameters as a vector sequence. This assumption seems over-simplified for modeling the temporal behavior of acoustic features. For example, cepstrum and its time-derivative (delta-cepstrum) can not synchronize with each other by definition, since a stationary value of time-derivative means a constant change in the cepstrum value. More in general, there is no guarantee that all feature parameters change at the same time; distinct features may have different timings of state transition. Moreover, they do not need to have the same number of hidden states. Such temporal behavior of multiple features have to be modeled by HMM with asynchronous state transition timings.

Recently, we proposed asynchronous-transition HMM (AT-HMM) to better model the asynchrony between features and discussed general classes of AT-HMMs [1]. The present paper focuses on a particular class of AT-HMM with sequential constraints in hidden state transition. The main idea here is "state tying along time" to implement the above idea still utilizing the conventional HMM structures and algorithms. This is yet another scheme of parameter tying to be added to existing various state tying techniques between allophones [3], between state output probabilities [2, 3], between mixture components [4], and between distribution parameters [5].

This paper consists of two major parts. In the first part, we introduce sequential AT-HMMs where state transition timings are asynchronous but constrained by a transition sequence. The state transition structures (topologies) representing phone context dependency are common throughout all features here. In the second part, in contrast, the structure is independently optimized for each of features to maximize the advantage of AT-HMM. This feature-wise state tying technique involves a new scheme of successive state splitting (SSS) algorithm. In both parts, AT-HMM is evaluated through phoneme and isolated word recognition experiments.

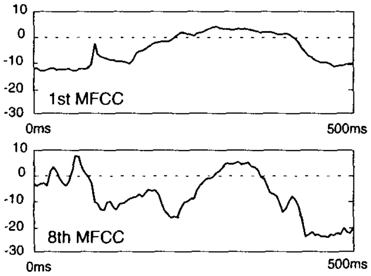


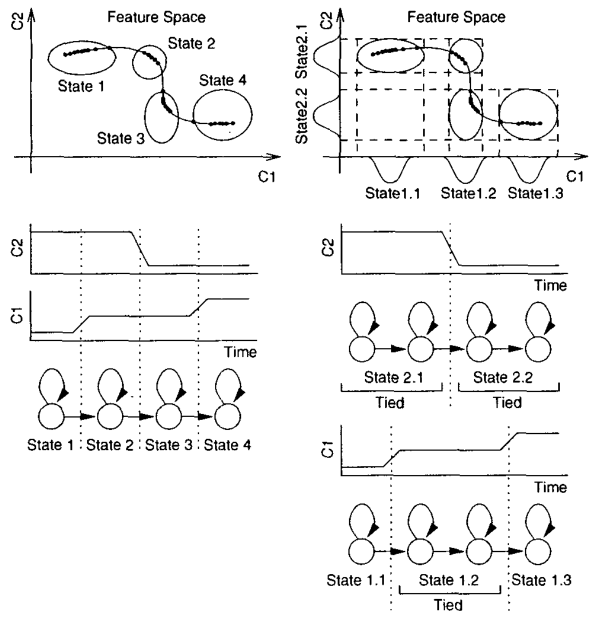
Figure 1: Asynchronous trajectories of 1st and 8th MFCCs in word /aoi/.

2. ASYNCHRONOUS-TRANSITION HMM

It is often observed that the dynamic patterns of individual feature sequences (vector components of acoustic feature vectors) have different timings of changing their values. Theoretically, cepstrum and its time-derivative have different timings. Fig. 1 contrasts the 1st and 8th MFCCs (mel-scaled cepstrum coefficients) in a word utterance, where these distinct features change their values in different timings. Depending on their statistical properties, they may require different numbers of hidden states for describing their trajectories along time. This fact may have increased the required number of hidden states in conventional HMMs for modeling speech signals. To represent such asynchrony between features, we introduced Asynchronous-Transition HMM (AT-HMM) [1] as a novel framework of HMM.

Fig. 2-(a) conceptually illustrates how the conventional HMM models a trajectory in a two-dimensional feature space. This two-dimensional trajectory consists of two one-dimensional trajectories shown at the middle of the figure where the two distinct features, C_1 and C_2, have different timings of changing their values. In representing these

0-7803-6293-4/00/$10.00 ©2000 IEEE.



(a) Modeling by conventional HMM (b) Modeling by AT-HMM

Figure 2: Conventional and AT- HMMs representing a 2-dimensional trajectory.

trajectories, for simplicity, by an HMM with a single mixture diagonal-covariance Gaussian distribution per state, we can point out some modeling redundancy between the four hidden states because feature C_1 does not change through states 1 to 2 and through 3 to 4, and feature C_2 does not change through states 2 to 3. This model contains an excessive number of model parameters and thus requires excessive amount of training data to train it.

To reduce such redundancy and to better model the trajectory, we can tie states 1 and 2, and 3 and 4 for feature C_1, and tie states 2 and 3 for feature C_2 as shown in Fig. 2-(b). Consequently, the model contains a smaller number of independent parameters in state output probabilities. Once the state-tying structure is established along time for each feature, it can be trained by the model training procedure with multiple-mixture Gaussian output probabilities and even with full-covariance Gaussian components. This means that AT-HMM is a general concept applicable to most types of HMMs.

In this implementation of AT-HMM, transition timings are asynchronized though sequentially constrained in a certain order. This implementation of AT-HMM has two significant advantages. First, the sequential constraint can reduce excess freedom in a simple asynchronous scheme without any constraints. Second, since the structure is substantially same as conventional HMMs except for sequential tying, the AT-HMM is easily adopted in most HMM-based speech recognition systems without any major modification. Apart from asynchrony of feature parameters, this novel idea of hidden state tying along the time axis enriches available choices in acoustic modeling together with existing tying techniques.

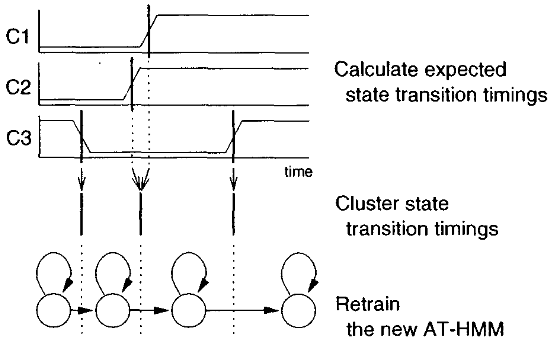


Figure 3: Algorithm for obtaining an AT-HMM temporally tied structure.

2.1. Algorithm for Designing AT-HMMs

There are more than one possible algorithm for obtaining the AT-HMM tying structure for phones with time resolution of N points, though they are approximations to the strictly optimal structure. Some of them are as follows:

Method 1: Train an N-state, left-to-right, diagonal-covariance single Gaussian HMM. Cluster adjacent hidden state outputs for individual features to find appropriate tying.

Method 2: Train an left-to-right, scalar-output HMM with appropriate number of states for each of all features. Cluster the expected transition timings into N-point time resolution.

The latter method is simple as described below and depicted in Fig. 3:

Step 1: Given a conventional phone HMM, re-train the model for each of the individual features, i.e., re-train 1-dimensional (scalar-output) phone HMM for each feature (vector component of acoustic feature sequence) to obtain state transition probabilities for individual features.

Step 2: Calculate expected transition timings for all features and states (utilizing that E[state duration] = 1/(state transition probability)), cluster them into a given resolution N of timings to obtain the temporal tying structure for each phone model.

Step 3: Re-train the new AT-HMM to update the model parameters under the obtained structure.

Note that the number of hidden states, N, provides the time resolution in representing the asynchronous structure. There is a trade-off: the larger number allows the more precise modeling of sequential structure, while it constrains the minimum phone duration.

2.2. Phone Recognition Experiments

To compare the performances of AT-HMM and conventional HMMs, speaker-dependent phone recognition experiments were conducted. For training HMMs and AT-HMMs, odd-numbered words out of 5240 common Japanese words

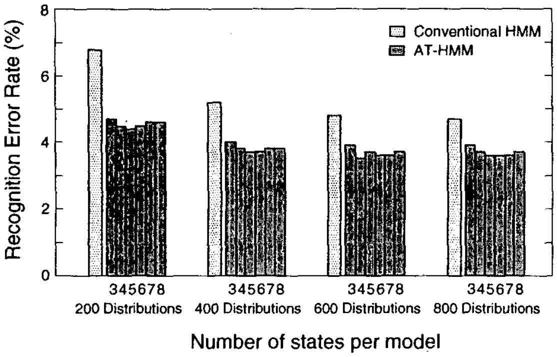


Figure 4: Phone recognition results of AT-HMM compared with conventional HMM (by ML-SSS)

Table 1: Isolated word recognition results by AT-HMM compared with conventional HMM (generated by ML-SSS)

Method	#distributions	%errors	%reduction
HMM	200	8.1	—
AT-HMM	200	4.3	46.9
HMM	400	6.2	—
AT-HMM	400	3.8	38.7

and 516 phonetically balanced words uttered by four (2 male + 2 female) speakers and sampled at 12kHz (ATR's set-A) were used. Acoustic features consisted of 12 MFCCs, 12 ΔMFCCs, log-power and Δlog-power obtained with a frame length of 25ms and period of 5ms. The same training algorithm called Maximum-Likelihood Successive State Splitting (ML-SSS) [6] was applied to generate context-dependent phone HMMs and AT-HMMs. As the result, both models have the same state transition topologies representing phonetical context dependencies.

Even-numbered word utterances in the 5240-word set were used for evaluation. Phone recognition was done for hand-segmented and labeled data. Fig. 4 shows the results of the speaker-dependent phoneme recognition task. AT-HMM with five states per model reduced error rate by more than 20%. As already discussed, the number of hidden states is slightly related to the performance; the AT-HMM with five states per model gave a little higher recognition rates than ones with other numbers of states from 3 to 8.

It should be noted, in the overall comparison between HMM and AT-HMM in this figure, that AT-HMMs with sequential constraints achieved higher performances with lower model complexities (fewer model parameters) than conventional HMMs.

2.3. Isolated Word Recognition Experiments

Table 1 shows the experimental results of speaker-dependent subword-based isolated word recognition task, where 2620-word lexicon was used to recognize the same size of speech data. In the experiment, the number of states for each AT-HMM was fixed to five. As is seen in the table, the AT-HMM successfully reduced the recognition error rates of the conventional HMM by approximately 40%.

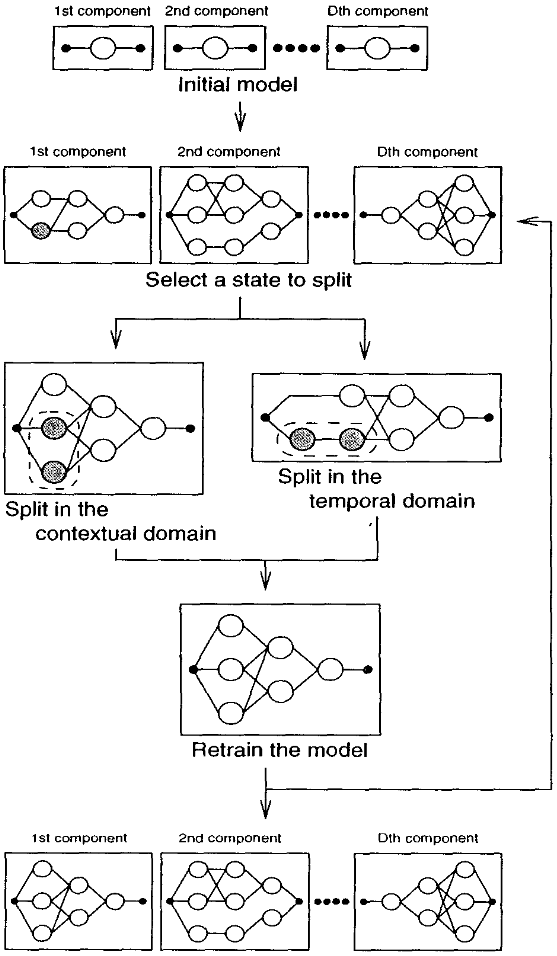


Figure 5: FW-SSS algorithm for designing the AT-HMM topology with optimized hidden-state allocation over features and allophones

3. FEATURE-WISE ALLOPHONE CLUSTERING

In the previous section, as for allophone (context-dependent phone) clusters, all feature share the same state-clustered structure (allophone network topology). The optimal allophone clusters, however, may differ among individual features. In other words, optimal allocation of hidden states may depend on allophones and features. To obtain feature-dependent allophone clusters, we propose a feature-wise state tying technique called Feature-Wise Successive State Splitting (FW-SSS). FW-SSS is a scalar version of ML-SSS [6] for each feature: the state splitting operation for each runs almost in parallel except that the splitting state is chosen among all states and features. The algorithm is shown in Fig. 5 and outlined as follows:

Step 1: Train a single state HMM for each feature with all phone samples, i.e., the output probability for each feature is represented by a single Gaussian with a

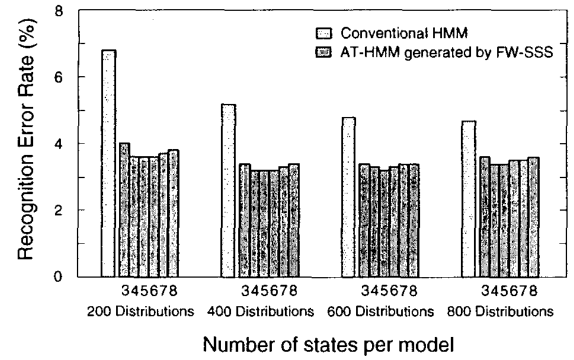


Figure 6: Phone recognition performance of AT-HMM generated by FW-SSS

Table 2: Isolated word recognition result by AT-HMM generated by FW-SSS

Method	#distributions	%errors	%reduction
HMM	200	8.1	—
AT-HMM	200	3.2	60.5
HMM	400	6.2	—
AT-HMM	400	3.0	51.6

mean and a variance.

Step 2: Among all states, find the state that will earn the largest likelihood gain when split into two states with single Gaussian distributions. State splitting gains are examined both in contextual and temporal domains.

Step 3: Re-train all states affected by the split using the corresponding data subsets.

Step 4: Repeat steps 2 and 3 until the number of all states reaches a preset number.

Step 5: Apply the algorithm described in 2.1 to obtain AT-HMMs.

Through the FW-SSS algorithm, a hidden Markov network is obtained wlth sub-optimized combination of numbers of hidden states for features reflecting the dynamic properties of distinct features. As the result, individual features have different allophone clusters and network topologies. The number of allocated hidden states to individual features differ from each other.

3.1. Phone Recognition Experiments

For evaluation of this type of AT-HMM generated by FW-SSS algorithm, speaker-dependent phoneme recognition was performed using the same data as used in the previous section.

Fig. 6 shows the performance of AT-HMM for four different model complexities. In comparison with conventional HMM, more than 30% of error reduction was obtained. AT-HMM generated from FW-SSS include asynchrony between features and feature-wise allophone clusters generated by the FW-SSS.

3.2. Isolated Word Recognition Experiments

AT-HMM generated by FW-SSS algorithm was evaluated in subword-based isolated word speech recognition. Phone models, same as the models for phoneme recognition, were evaluated using 2620-word speech data and a 2620-word lexicon.

Table 2 shows the experimental results of isolated word recognition. The acoustic model generated by the FW-SSS algorithm lowered the error rates by more than 50% compared with conventional HMM. AT-HMM generated by FW-SSS gave higher recognition rate than AT-HMM without being considered state sharing structure for each features.

4. CONCLUSION

Focusing on asynchrony between acoustic features for HMM-based speech recognition, we introduced novel concepts such as asynchronous transition HMM (AT-HMM), tying along time, and FW-SSS algorithm for the optimal context-dependent structure of AT-HMM. With these ideas combined together, the proposed model is a highly sophisticated and quite general model containing feature-wise state tying along time and across allophones. This class of a new HMM can be regarded as a yet further generalization to existing HMM classes with tied structures such as tied-mixture, allophone clusters, and parameter tying.

In experimental performance evaluation of phoneme and isolated word recognition, AT-HMMs gave more than 20% and 40% lower error rates compared with conventional HMMs. Furthermore, the FW-SSS algorithm gave an AT-HMM reducing more than 30% and 50% errors.

Future works will include evaluation of mixture-density speaker-independent AT-HMMs and experimental evaluation in continuous speech recognition.

REFERENCES

[1] S. Sagayama, S. Matsuda, M. Nakai and H. Shimodaira, "Asynchronous Transition HMM for Acoustic Modeling," Proc. 1999 IEEE Workshop on Speech Recognition and Understanding, to appear in Dec. 1999.

[2] X.D. Huang, K.F Lee, H.W. Hon, M.Y. Hwang: "Improved Acoustic Modeling with the SPHINX Speech Recognition System," Proc. ICASSP91, pp. 345–348, 1991.

[3] J. Takami, S. Sagayama: "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. ICASSP92, pp. I-573–576, 1992.

[4] J. Bellegarda, D. Nahamoo: "Tied mixture continuous parameter models for large vocabulary isolated speech recognition," Proc. ICASSP89, pp.13–16, 1989.

[5] S. Takahashi, S. Sagayama: "Four-level Tied Structure for Efficient Representation of Acoustic Modeling," Proc. ICASSP95, pp. 520–523, 1995.

[6] M. Ostendorf, H. Singer: "HMM Topology Design Using Maximum Likelihood Successive State Splitting," Computer Speech and Language, 11(1), pp. 17–41, 1997.