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
{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 asynchronous 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 introducing a concept of "state tying across time". To maximize the advantage of the new model, we also introduce feature-wise state tying technique. Speaker-dependent speech recognition experiments demonstrated that reduced error rates more than 30% and 50% in phoneme and isolated word recognition, 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 by treating acoustic features of input speech as a vector sequence. This assumption seems over-simplified to model asynchronous changes of acoustic features. For example, cepstrum and its time-derivative (delta-cepstrum) can not synchronize with each other in principle, because a stationary value of time-derivative means a constant change in the cepstrum value. Intuitively, these features seem to be better modeled by HMM with different state transition timings. More in general, there is no guarantee that all feature parameters change at the same time; different features may have state transition of different timing.

We proposed asynchronous-transition HMM (AT-HMM) to better model asynchronize vector sequence and discussed general classes of AT-HMMs [8]. This paper focuses on a particular class of AT-HMM with sequential constraints in hidden state transition. The main idea here is "state tying across time" to implement the above idea still utilizing the conventional HMM structures and algorithms. This is yet another scheme of parameter tying in addition to existing various state tying techniques between allophones [4], state output probabilities [3, 4], mixture components [2], and distribution parameters [5].

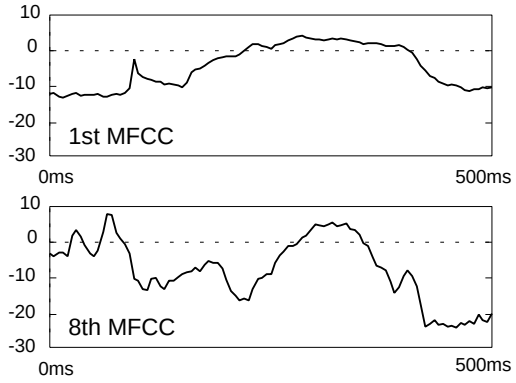


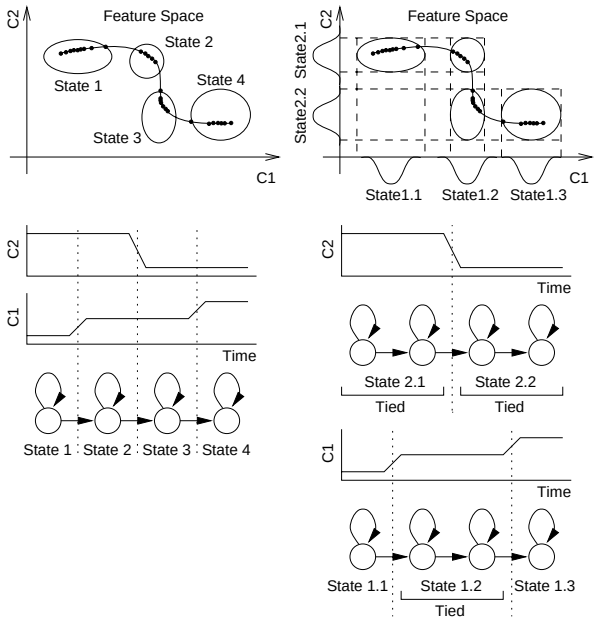
Figure 1: Asynchronous trajectories of 1st and 8th MFCCs in word /aoi/.

This paper consists of two major parts. In the first part, we introduce sequential AT-HMMs where state transition timings are asynchonous but constrained by a transition sequence. The state transition structures (topologies) to represent phone context dependency are common throughout all features here. In the second part, however, 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.

2. ASYNCHRONOUS-TRANSITION HMM

As shown in Fig. 1, 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. This fact may have increased the required number of hidden states in conventional HMMs. To enable representing such asynchrony between features, we introduced Asynchronous-Transition HMM (AT-HMM) as a new framework of HMM [8].

Fig. 2-(a) conceptually illustrates a trajectory in



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

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

a 2-dimensional feature space represented by conventional HMM where two distinct features have different timings of changing their values. In this case, four hidden states contain redundancy that feature C_1 does not change between states 1 and 2 and between 3 and 4, and feature C_2 does not change beteen states 2 and 3. To reduce ths redundancy and 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.

In this implementation of AT-HMM, transition timings are asynchronized while transitions are sequentially constrained in a certain order. This implementation of AT-HMM has two significant advantages. First, the sequential constraint may reduce excess freedom in a simple asynchronous scheme without any constraint. Second, since the structure is substantially same as conventional HMMs except for tying across time, the AT-HMM is easily adapted to most HMM-based speech recognition systems without any major modification.

2.1. Algorithm for Generating AT-HMMs

There are more than one possible algorithm for obtaining the AT-HMM tying structure for phones with time resolution of N points such as:

- Clustering N hidden states in a normal HMM to find appropriate tying

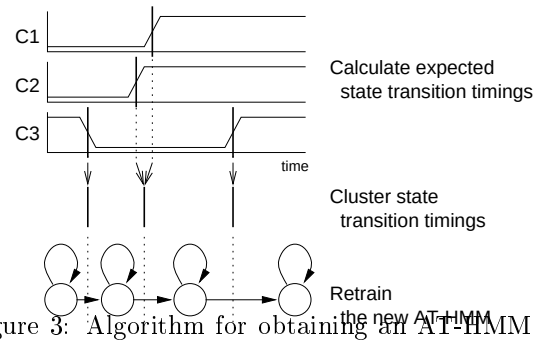


Figure 3: Algorithm for obtaining an AT-HMM temporally tied structure.

- Clustering transition timings in scalar output HMMs for all features into N points

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

Step 1: Given a conventional phone HMM, retrain the model for each of individual features, i.e., retrain 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 all expected transition timings for all features (utilizing that E[state duration] = 1/(state transtion probability)), cluster them into a given resolution N of timings, and determine the temporal tying structure for each phone model.

Step 3: Retrain the new AT-HMM with temporal tying structure generated in Step 2.

Vector sequence with asynchronous temporal structure more better model by using many clusters than using few clusters in Step 2. *** cannot understand ***

2.2. Phone Recognition Experiments

AT-HMM was evaluated in speaker-dependent phoneme recognition experiments and compared with synchronous HMM (conventional HMM). The both context-dependent HMM topologies were common and generated by the ML-SSS algorithm [7].

Speech data from ATR A-set data consists of four (2 male + 2 female) speakers sampled at 12kHz. 12th-order MFCCs, ΔMFCCs, log-power and Δlog-power were extracted with a 5ms frame period and a 25ms frame length. Hand-segmented phoneme data from odd numbered words out of 5240 Japanese common words and 516 phonetically balanced words were used for model training and phoneme data from even numbered words out of 5240 words were used for evaluation.

Fig. 3 shows the speaker-dependent phoneme recognition results. AT-HMM with five states per model reduced error rate by more than 20%. The number

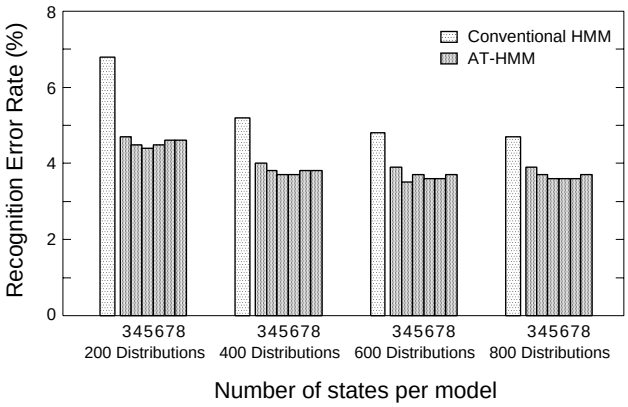


Figure 4: Phone recognition results of AT-HMM compared with conventional HMM (by ML-SSS)

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

of hidden states provides the time resolution in representing the asynchronous structure. The larger number allows the more precise modeling of sequential structure while the minimum phone duration is constrained by the number. Actually, in the experimental results, the AT-HMM with five states per model gave slightly higher recognition rates than ones with other numbers of states from 3 to 8.

2.3. Isolated Word Recognition Experiments

AT-HMM with five states per model was chosen as the best performing model in the phoneme recognition experiment, and was evaluated with isolated word speech recognition using 2620-word speech data and a 2620-word lexicon. Table 4 shows the experimental results of isolated word recognition. The AT-HMM reduced more than 40% of recognition errors by conventional HMM.

3. FEATURE-WISE ALLOPHONE CLUSTERING

The optimal allophone (context-dependent phone) clusters may differ among individual features. To obtain feature-dependent allophone clusters, we propose a feature-wise state tying technique called Feature-Wise

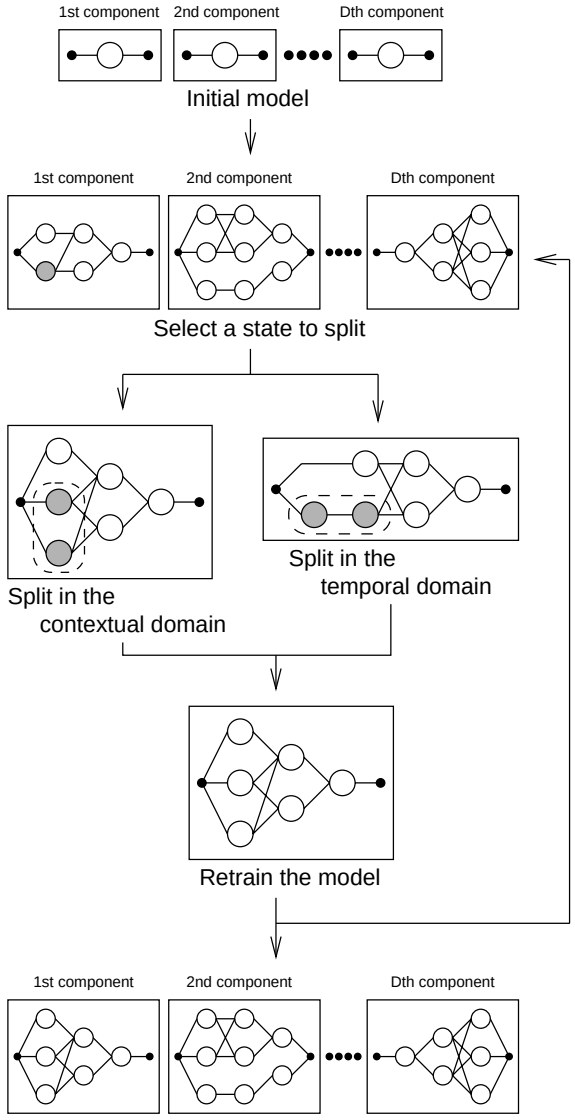


Figure 5: FW-SSS algorithm for generating AT-HMMs

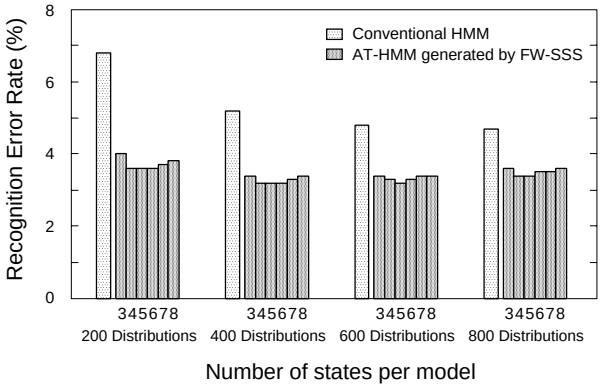


Figure 6: Phone recognition performance of AT-HMM generated by FW-SSS

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

Successive State Splitting (FW-SSS). FW-SSS is an extension of the Maximum Likelihood (ML)-SSS algorithm. The main difference is that FW-SSS is scalar version of ML-SSS for all features in parallel and that the state to be split next is selected from all states of all features. The outline of the algorithm is as follows and shown in Fig. 4:

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 mean and a variance.

Step 2: Find the best state of all states that will earn the largest likelihood gain by splitting it into two states with a single Gaussian distribution for each. State splitting is examined both in contextual and temporal domains.

Step 3: Retrain states affected by the split using the corresponding data subsets.

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

Step 5: Finally, utilize the algorithm in subsection 2.1 to obtain AT-HMMs.

Through the FW-SSS algorithm, a hidden Markov network is obtained with 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 over 4 speakers using AT-HMMs generated for several different numbers of distributions.

Fig. 5 shows the performance of AT-HMM for six 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 for isolated word speech recognition. Phone models as same as the models for phoneme recognition were evaluated using 2620-word speech data and a 2620-word lexicon.

Table 4 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 some new concepts such as asynchronous transition HMM (AT-HMM), tying across/along time, and FW-SSS algorithm for the optimal context-dependent structure of AT-HMM. In phoneme and isolated word recognition experiments, 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 mixture density speaker-independent models and more experimental evaluation in continuous speech recognition.

5. REFERENCES

[1] B. H. Juang: "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains," AT&T Tech. J., 64, 6, pp. 1234–1249, 1985.

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

[3] 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.

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

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

[6] S. Takahashi, S. Sagayama: "Discrete Mixture HMM for Speech Recognition," Proc. ICASSP97, vol. 2, pp. 971–974, 1997.

[7] M. Ostendorf, H. Singer: "HMM Topology Design Using Maximum Likelihood Successive State Split-

ting," Computer Speech and Language, 11(1), pp. 17–41, 1997.

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