ASYNCHRONOUS-TRANSITION HMM FOR ACOUSTIC MODELING

Shigeki Sagayama, Shigeki Matsuda, Mitsuru Nakai and *Hiroshi Shimodaira*

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

ABSTRACT

We propose a new class of hidden Markov model (HMM) which we call Asynchronous-Transition HMM (AT-HMM) to model asynchronous temporal structure of acoustic feature sequences. Conventional HMM models a sequence of feature vectors, while temporally changing patterns of acoustic features do not necessarily synchronize with each other. In this paper, AT-HMMs with and without sequential constraints are discussed. Algorithms for generating context-dependent AT-HMM and for deriving sequentially constrained AT-HMM are provided. A new concept of "state tying across time" is also introduced. Speaker-dependent speech recognition experiments demonstrated error reduction rates of more than 30% and 40% in phoneme and isolated word recognition, respectively, compared with conventional HMMs.

1. INTRODUCTION

In conventional Hidden Markov Models (HMMs) for speech recognition, acoustic features of input speech are usually treated as a vector sequence (or, equivalently, vector-quantized code sequences). Since the output probability of the model changes with hidden state transition, this modeling implicitly assumes that individual acoustic feature parameters change their statistical properties simultaneously.

This assumption is not true. Different features may have different timings. For example, as shown in Fig. 1, cepstrum and its differential feature (delta-cepstrum) can not synchronize by definition. If a delta-cepstrum takes relatively stable non-zero values at a certain interval of the input speech and is assigned to a hidden state, it means the cepstrum is having a dynamic change and is possibly having a hidden state transition. Also, if cepstrum takes relatively stable values, delta-cepstrum must have near-zero values. Intuitively, these features seem to be better modeled by HMM with different state transition timings. The same situation seems to be found in all different features such as MFCCs (mel-frequency cepstrum coefficients). Moreover, different features may need different numbers of hidden states to describe their trajectories.

Until today, however, research effort on HMM-based acoustic modeling has been much focused on HMM output probability design. A number of models such as continuous-mixture HMM [1], tied mixture [2], state tying [3, 4], parameter tying [5] and discrete mixture [6] have been proposed to better represent state output probabilities with smaller numbers of parameters.

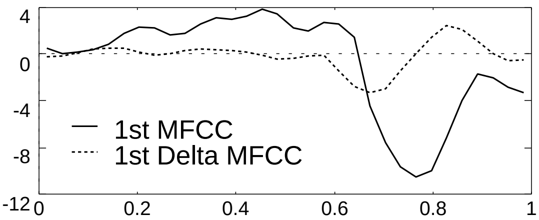

Figure 1: An example of phoneme /k/ where MFCC and ΔMFCC have asynchronized trajectories

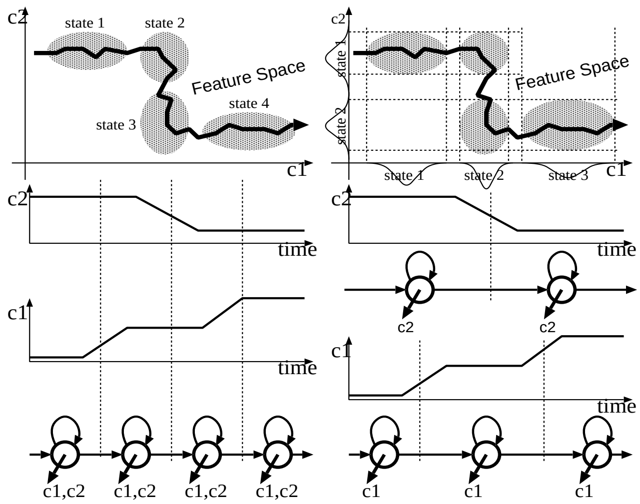

Figure 2: Representations of a 2-dimentional trajectory with conventional and AT- HMMs

On the other hand, relatively little attention has been paid to the HMM temporal structure. A few examples of asynchronous handling of features are found as a result of taking a multiple feature streams approach [7, 8] though their main interest stays within the robust speech recognition against noise. This paper discusses a new type of HMM with state transitions not necessarily synchronized between features.

The purpose of the present paper is to provide a new model structure from a temporal properties point of view. To maximize the advantage of the new model, feature-wise state tying (i.e., allophone clustering) technique will be introduced. Also, a new type of state tying across time will be introduced, primarily as an implementation of a new model.

2. ASYNCHRONOUS-TRANSITION HMM

As stated above, 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. (This fact may have increased the required number of hidden states in conventional HMMs.) We introduce Asynchronous-Transition HMM (AT-HMM) as a new framework of HMM that enables representing such asynchrony between features. Fig. 2 conceptually illustrates how a two-dimensional vector trajectory is modeled by conventional and AT- HMMs. In the latter case, the vector is decomposed into features, each modeled basically by a scalar-HMM independently as shown in Fig. 3.

Liberated from the synchrony constraints, AT-HMM has much more freedom in state transition timings than conventional HMMs. This means that, when used in recognition, it will yield higher likelihood than that from conventional HMMs not only for the matched category but also for unmatched categories. A mixed use of asynchrony as a freedom and synchrony as a constraint may be effective for the recognition purpose.

Fig. 4 shows some of various types of mixtures of asynchrony and synchrony of feature parameters. Type A (synchronous HMM) is conventional vector-based HMM where all features share the same state transition timings in contrast to type C (asynchronous HMM) where all state transitions are asynchronous. We can consider type B (partially asynchronous HMM) as a mixture between these extremes. This type includes the following categories. Type B-1 has feature-grouped asynchronous transitions where features are grouped to have synchronized transitions, each group having asynchronous transitions. Type B-2 has model-synchronized asynchronous transitions to constrain the model to synchronize all features at the phone boundary. Type B-3 illustrates the most general type of mixed synchronous and asynchronous transitions. Transitions are either asynchronous or synchronized with one or more other features. In this sense, synchronization can be interpreted as tying state transition timings.

Furthermore, we can introduce sequential constraints for asynchronous transitions to better model asynchronous dynamic characteristics. In Fig. 4, to maintain the "z-shaped" two-dimensional trajectory, it is essential to constrain transition timings in a sequence such as first, from state 1 to state 2 for feature 1, second, from state 1 to state 2 for feature 2, and third, from state 2 to state 3 for feature 1. Thus, we have a wide variety of new HMMs with mixed synchronous and asynchronous transitions with/without sequential constraints from which we can select one to best model speech.

3. NON-SEQUENTIAL AT-HMM

First, we discuss the first category, i.e., AT-HMM without sequential constraints where the state transition of the individual features are assumed to occur independently. This category can be implemented by representing each feature by a scalar-output HMM as depicted in Fig. 3. The overall likelihood is obtained by summing up all likelihoods of individual features.

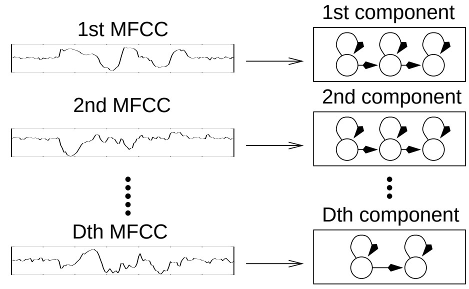

Figure 3: The structure of the non-sequential AT-HMM

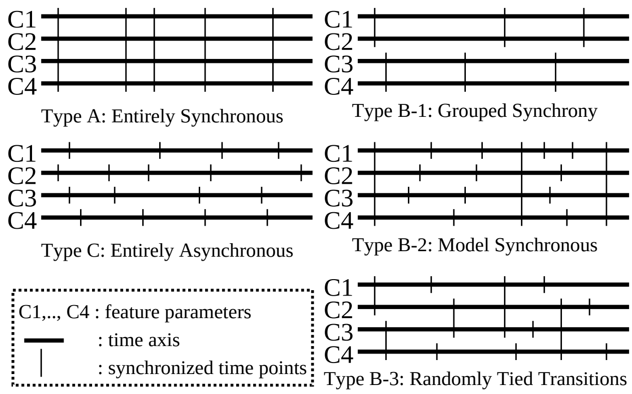

Figure 4: Classes of asynchrony between features (conceptual diagrams)

3.1. Generating CD AT-HMM Phone Models

To automatically design context-dependent phone models in the AT-HMM framework, we propose a new algorithm called "Feature-wise Successive State Splitting" (FW-SSS). This is an extension of the Maximum Likelihood (ML)-SSS algorithm [4, 9]. The main difference is that FW-SSS does scalar version of 'ML-SSS' for all features in parallel and the state to be split next is selected from all states of all features. Given a number of phone samples from a single speaker, the outline of the algorithm shown in Fig. 5 is as follows:

Step 1: Train a single state HMM for each feature by 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.

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

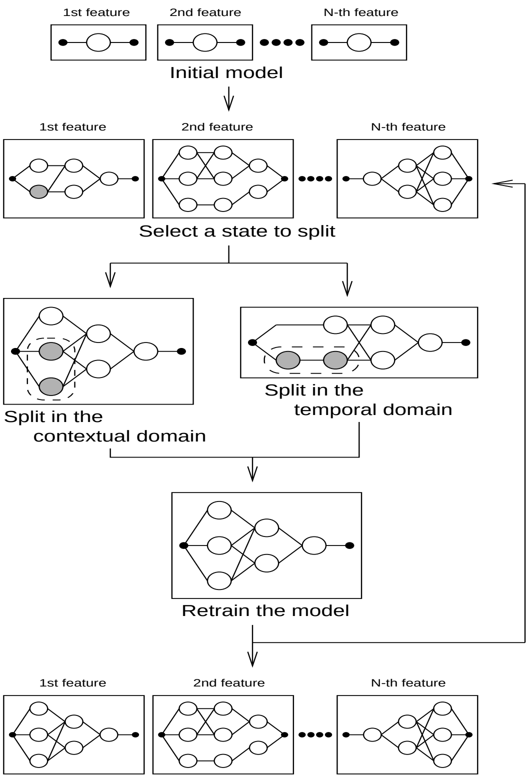

Figure 5: FW-SSS algorithm to generate AT-HMMs

may have different allophone clusters and network topologies. The numbers of allocated hidden states to individual features may be different from each other.

3.2. Phoneme Recognition by AT-HMM

AT-HMM was experimentally evaluated in speaker-dependent phoneme recognition experiments and compared with synchronous HMM generated by the ML-SSS algorithm.

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

Table 1 shows the speaker-dependent phoneme recognition results using transition-asynchronous HMM phone models generated by FW-SSS compared with those using transition-synchronous HMM generated by ML-SSS, which is considered as one of best models among conventional HMMs. In FW-SSS, an hidden Markov network topology was automatically generated so as to provide each feature with the appropriate number of states and complexity.

3.3. Isolated Word Recognition by AT-HMM

For isolated word speech recognition, AT-HMM phone models were constrained so that state transitions synchronize at phoneme boundaries while they were left asynchronous within phonemes as shown as Type B-2 in Fig. 4. Since likelihood computation is not achieved by a simple Viterbi algorithm for this case, the forward algorithm is used within phonemes and an algorithm similar to the two-stage DP algorithm is used to compute the entire word likelihood.

Table 1: Phoneme recognition results of AT-HMM (generated by FW-SSS, non-sequential) compared with conventional HMM (by ML-SSS)

method	#parameters	%errors	%reduction
ML-SSS	10600	7.9	-
FW-SSS	10608	6.2	21.5
ML-SSS	21200	5.8	-
FW-SSS	21216	3.8	34.5
ML-SSS	31800	5.3	-
FW-SSS	31824	3.3	37.3
ML-SSS	42400	5.2	-
FW-SSS	42432	3.2	38.5

Table 2: Isolated word recognition results by (non-sequential) AT-HMM

method	#parameters	%errors	%reduction
ML-SSS	10600	5.4	-
FW-SSS	10608	2.9	46.6
ML-SSS	21200	3.2	-
FW-SSS	21216	1.7	47.0

The same context-dependent phone models as the models for phoneme recognition were evaluated using 1310-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 achieved an error reduction rate more than 40% compared with conventional (transition-synchronous) HMM with approximately same number of model parameters generated by ML-SSS. It is concluded that the AT-HMM generated by FW-SSS better represents the acoustic patterns of speech.

4. SEQUENTIAL AT-HMM

The second category of AT-HMM is sequential AT-HMM in which state transitions are still asynchronous between features while they satisfy a certain sequential constraint. We have seen that the synchronous temporal structure of conventional HMMs is too much constrained to model asynchronous features such as cepstra and delta-cepstra. AT-HMM, however, may be too free from temporal constraints. Acoustical events in speech signals may not be synchronous but seem sequentially ordered to some extent. From this idea, sequential AT-HMM is considered as a the second category of AT-HMM to incorporate asynchrony between features.

4.1. State Tying across Time

One of possible implementations of sequential AT-HMM is state tying across time. In contrast with existing state tying techniques between allophones [4], state output probabilities [3, 4], mixture components [2], and distribution parameters [5] for robust modeling, state tying across time is another scheme of utilizing tied structures. As shown in

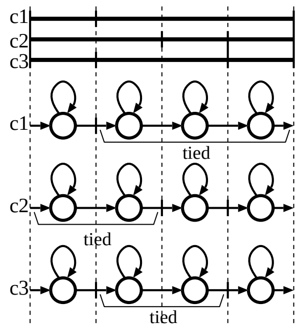

Figure 6: The structure of sequential AT-HMM

Fig. 6, the synchrony and asynchrony specified by the upper diagram is implemented by the lower HMMs where features c_1, c_2, and c_3 have 2, 3, and 3 distinct hidden states, respectively, all represented by 4 hidden states whose output distributions are tied across time according to the synchrony diagram.

This implementation has a significant advantage. Since the structure is substantially same as conventional HMMs except for tying across time, the AT-HMM is easily devised in most HMM-based speech recognition systems without any major revision.

4.2. Generation of AT-HMM Phone Models

One simple algorithm to obtain a sequential AT-HMM is to convert a non-sequential AT-HMM into a sequential one is as follows:

Step 1. Use the FW-SSS algorithm to generate a non-sequential AT-HMM models.

Step 2. Retrain state transition probabilities for each of all context-dependent classes (triphones in this case) to obtain expected state transition timings.

Step 3. Cluster all expected transition timings into a given number of timings and determine the temporal tying structure for each triphone.

Step 4. If adjacent transition timings are clustered into the same cluster, merge the respective state output probabilities. (The number of model parameters may decrease.)

The advantage of this algorithm is that it inherits feature-wise allophone clusters (context dependency) generated by the FW-SSS algorithm to represent speech with fewer parameters.

4.3. Phoneme Recognition Experiment

For initial evaluation of sequential AT-HMMs, speaker-dependent phoneme recognition was performed over 4 speakers using single-Gaussian output density sequential AT-HMMs generated for several different numbers of states. Table 3 shows the performance of sequential AT-HMM for two different model complexities. In comparison with ML-SSS, more than 30% of error reduction was obtained with less numbers of model parameters. The source of this improvement may include sequential asynchrony between features and feature-wise allophone clusters generated by the FW-SSS algorithm.

Table 3: Phone recognition results of sequential AT-HMM (compared with conventional HMM)

method	#states	#parameters	%errors	%reduction
ML-SSS	3	10400	6.6	-
FW-SSS	3	7072	4.7	28.8
	6	7072	4.5	31.8
	8	7072	4.9	25.8
ML-SSS	3	20800	4.6	-
FW-SSS	3	12792	3.7	20.0
	6	12792	2.9	37.0
	8	12792	3.1	32.6

5. CONCLUSION

Focusing on asynchrony between acoustic feature sequences, we introduced some new ideas such as asynchronous-transition HMM (AT-HMM), FW-SSS algorithm for generating context-dependent AT-HMMs, and "state tying across time" for deriving sequentially constrained AT-HMMs. Initial evaluation through phoneme recognition experiments, both non-sequential and sequential AT-HMMs gave more than 30% of error reduction rates compared with conventional HMMs. Future works will include mixture density speaker-independent models and more experimental evaluation in continuous speech recognition.

6. 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] H. Bourlard, S. Dupont: "A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands," Proc. ICSLP96, pp. 426–429, 1996.

[8] S. Tibrewala, H. Hermansky: "Sub-Band-Based Recognition of Noisy Speech," Proc. ICASSP97, pp. 1255–1258, 1997.

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