# MODELING SUCCESSIVE FRAME DEPENDENCIES WITH HYBRID HMM/BN ACOUSTIC MODEL

*Konstantin Markov, Satoshi Nakamura*

Department of Acoustics and Speech Research,
Spoken Language Translation Research Labs,
Advanced Telecommunications Research Institute International,
Kyoto, Japan
{konstantin.markov,satoshi.nakamura}@atr.jp

## ABSTRACT

Most of the current state-of-the-art speech recognition systems use the Hidden Markov Model (HMM) for modeling acoustical characteristics of a speech signal. In the first-order HMM, speech data are assumed to be independently and identically distributed (i.i.d.), meaning that there is no dependency between neighboring feature vectors. Another assumption is that the current vector depends only on the current HMM state. In practice, however, these assumptions are not true. In this paper, we describe a hybrid HMM/BN (Bayesian Network) acoustic model, where the dependency of the current speech vector on the previous vector and on the previous state is also learned and used in speech recognition. This is possible because, the state probability distribution is modeled by a BN. Previous instances of the state and speech feature vector are represented by additional variables of the BN and the probabilistic dependencies between them, and their current instances are learned during the training. During recognition, the likelihood of the current feature vector is inferred from the BN where the previous state and previous feature vector are treated as hidden. We have evaluated this hybrid HMM/BN model with our LVCSR system by phoneme recognition and by large-vocabulary continuous word recognition tasks. In both cases, we observed improved performance over the conventional Gaussian mixture HMM.

## 1. INTRODUCTION

Since its introduction to automatic speech recognition about 20 years ago, the Hidden Markov Model (HMM) has become the dominant tool for speech signal modeling. In practice, the most widely used type is the first-order HMM, for which efficient learning and recognition algorithms are available. In this model, however, observation vectors are assumed to be independent and identically distributed (i.i.d.), given the HMM state. This is a serious drawback since in most cases speech feature vectors are highly correlated, thus making the stationarity assumption invalid. Researchers have been trying to overcome this problem for a long time, but an efficient solution has not been found.

Most of the reported studies have taken one of two main directions. The first approach attempts to extract or model the characteristics of a sequence of successive observations. Variable-length segments of frames are used in the Stochastic Segment Model [1]. In studies such as like [2], the frame correlation is accounted for by modeling the feature vectors' trajectories. Trended HMM [3] uses a parametric model to represent the dynamics of the observations within the state. In all of these cases, the frame correlation is implicitly accounted for by considering the data evolution over time. A drawback to this approach is that the resulting models are not compatible with the mainstream Continuous Density HMM (CDHMM) and are generally computationally expensive.

In contrast, the approach of the other direction attempts to express the dependency between successive observations directly in probabilistic form. This is done easily by conditioning the current vector distribution on the previous observation, i.e. using $p(x_t|x_{t-1}, s_t)$ as a state output likelihood. In addition, dependency on the previous state can also be included: $p(x_t|x_{t-1}, s_{t-1}, s_t)$. The theory of such an HMM was developed quite a long ago [4], but direct implementation leads to an excessive increase in the model parameter number [5], which is undesirable in practice because in most cases the training data are limited. Some research efforts have tried approximations of the conditional densities in order to reduce the model complexity. In the so-called Bigram-Constrained HMM (BC-HMM), the state output probability distribution is restricted by the observation symbol of the previous frame [6]. The correlation information between adjacent vectors is contained in the bigram probabilities. The BC-HMM approach has been further extended in [7], where the conditional probability distribution

0-7803-8874-7/05/$20.00 ©2005 IEEE

is approximated by an extended logarithmic pool of Gaussians. In these works, the current vector dependency on the previous state is not considered, and the frame correlation is assumed to be independent of the HMM state. An attempt to address these two issues has been made with the Frame Correlation HMM (FC-HMM) [8], where $p(x_t|x_{t-1}, s_{t-1}, s_t)$ is approximated by weighting the frame likelihood by a nonlinear function that depends on both the previous and current state output values. In [9], the state transition probabilities are conditioned on the previous observation as well. Recently, the Buried Markov model was proposed [10], where the correlation between the neighboring frames is considered at the vector component level.

In this paper, we describe an approach to modeling the successive frames dependency based on Bayesian Networks (BN) technology. The BN is able to represent complex joint probability distributions of many discrete and continuous variables and has great flexibility in modeling their dependencies. In our model, the HMM state probability distribution is described by a BN where the current and previous observation and state are represented by different nodes. Dependencies between them are expressed by directed arcs that connect the nodes. The power of the BN allows us to model not only the dependency on the previous frame but also the wider contexts of both the observation and state sequences. We implemented our approach using the hybrid HMM/BN model, which we have already applied successfully in various tasks [11, 12].

## 2. HYBRID HMM/BN MODEL

The HMM/BN model is a combination of an HMM and a Bayesian Network. The temporal characteristics of speech are modeled by the HMM state transitions, while the HMM states' probability distributions are represented by the BN. A block diagram of the HMM/BN is shown in Fig. 1. Structurally, the HMM/BN model is analogous to the hybrid HMM/NN model. The difference is that instead of a Neural Network, the HMM is coupled with a Bayesian Network. By definition, a BN represents a joint probability distribution of a set of random variables $Z_1, \ldots, Z_N$ and is expressed by a directed acyclic graph (DAG), where each node corresponds to a unique variable. Arcs between the nodes show the conditional dependencies of the BN variables. The immediate predecessors of variable $Z_i$ are called its *parents* and are referred to as $Pa(Z_i)$. The BN joint probability distribution function can be factored as

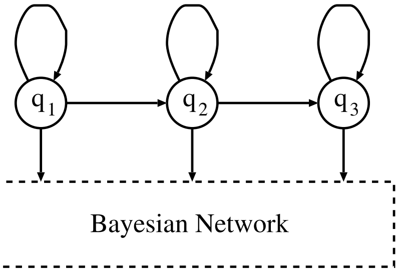
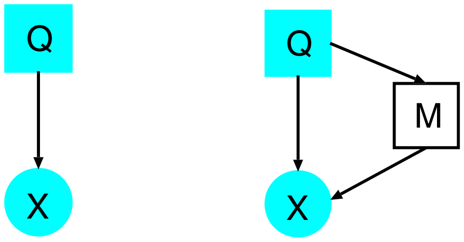


**Fig. 1**. HMM/BN model structure. HMM transitions model temporal characteristics of speech, and BN represents states' probability distributions.

$$P(Z_1, \ldots, Z_N) = \prod_{i=1}^{N} P(Z_i|Pa(Z_i)) \qquad (1)$$

The traditional HMM is a special case of HMM/BN; when the BN has one of the topologies shown in Fig. 2, the HMM and HMM/BN are equivalent. The variable $M$ of Fig. 2b represents the mixture index $(M = 1, \ldots, K)$. Since it is hidden, and according to Eq. (1) for $P(X|Q)$, we have



(a) Single Gaussian. (b) Mixture of Gaussians.

**Fig. 2**. BN topologies that make the HMM/BN model equivalent to the conventional HMM, with single (a) and mixture of Gaussians (b) state distributions.

$$\begin{aligned} P(X|Q) &= \\ &= \frac{P(X,Q)}{P(Q)} = \frac{\sum_{j=1}^{K} P(X, M=j, Q)}{P(Q)} \\ &= \frac{\sum_{j=1}^{K} P(X|M=j,Q)P(M=j|Q)P(Q)}{P(Q)} \\ &= \sum_{j=1}^{K} P(M=j|Q)P(X|M=j,Q) \qquad (2) \end{aligned}$$

Assuming that $P(X|M = j, Q)$ is a Gaussian, the above equation is simply a Gaussian mixture function, where $P(M = j|Q)$ is the $j^{th}$ mixture weight.

Parameter learning of the HMM/BN model is based on the Viterbi training algorithm, where each iteration consists of BN training and HMM transition probabilities update. More details can be found in [11]. Decoding with the HMM/BN model is essentially the same as with the HMM, where instead of calculating $p(x|q)$ as a Gaussian mixture it is inferred from the BN.

## 3. MODELING SUCCESSIVE FRAME DEPENDENCIES

The advantage of using the BN as a state distribution model is that it is very easy to add additional variables. In order to model the dependency between the current and previous observations, we add one additional variable representing $X_{t-1}$ as shown in Fig. 3. Since $X_t$ and $X_{t-1}$ are real valued variables, their conditional distribution can only be approximated. The Linear Regression (LR) model is one popular choice [10]. However, when the number of states is too big, as in the context-dependent acoustic models, having an LR approximation for each state excessively increases the complexity of the entire model. The approach we took is to approximate $X_{t-1}$ by a discrete variable, i.e. by VQ labels. This makes the BN computationally very simple in both the learning and inference stages. Since all BN variables are observable during training[1], the standard Maximum Likelihood (ML) parameter estimation is used. During recognition, the $X_{t-1}$ variable can be either observable (labels can be obtained by VQ) or hidden. The latter case is especially interesting because inferring the $P(X_t|Q_t)$ becomes a Gaussian mixture calculation. Similarly to Eq. (2), we have

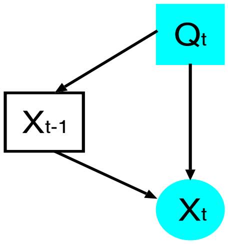


**Fig. 3**. BN topology for modeling dependency on the previous observation.

$$P(X_t|Q_t) = \sum_{j=1}^{K} P(X_{t-1} = j|Q_t)P(X_t|X_{t-1} = j, Q_t) \tag{3}$$

where $K$ is the VQ codebook size. Another advantage is that during recognition the quantization of $X_{t-1}$ is not necessary.

Extending this model to include the current observation dependency on the previous state is as easy as adding another discrete variable representing $Q_{t-1}$. The BN topology in this case is shown in Fig. 4. Assuming again that $X_{t-1}$ and $Q_{t-1}$ are hidden during recognition, for $P(X_t|Q_t)$ we get

$$\begin{aligned} P(X_t|Q_t) &= \\ &= \sum_{i=1}^{S}\sum_{j=1}^{K} P(Q_{t-1} = i|Q_t)P(X_{t-1} = j|Q_{t-1} = i, Q_t) \\ &\quad P(X_t|X_{t-1} = j, Q_{t-1} = i, Q_t) \end{aligned} \tag{4}$$

where $S$ is the number of states in the model. It is easy to see that the above equation is again a Gaussian mixture expression with weights calculated as $P(Q_{t-1} = i|Q_t)P(X_{t-1} = j|Q_{t-1} = i, Q_t)$.

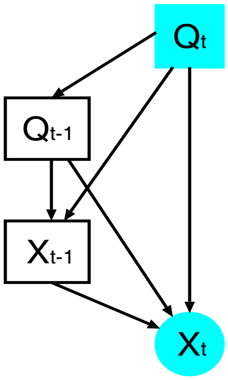


**Fig. 4**. BN topology for modeling dependency on both previous observation and previous state.

## 4. EXPERIMENTS AND RESULTS

For acoustic model training, we used the official Wall Street Journal (WSJ) training corpus WSJ-284. It includes about 60 hours of read speech by 284 speakers. The test data consisted of 200 utterances selected randomly from a set of 4000 read speech utterances spoken by 40 speakers. The speech material of the test data consists of travel-related expressions and is quite different from that of the training data. All speech utterances were collected in quiet environments. Here, 25-dimensional (12MFCC + 12DeltaMFCC + Pow) feature vectors are extracted with a 20 ms sliding window at a 10 ms frame rate.

Our baseline acoustic model is an HMnet obtained by a successive state splitting algorithm with an MDL stopping criterion [13]. The total number of states is 2009. Four versions with 5, 10, 15 and 20 Gaussian components per state were trained in order to compare the models with different parameter numbers. The HMM/BN models were initialized using the baseline HMnet, meaning that they have the same number of states and the same state topology. Only the probability distribution model of the individual states is different: a fixed number of Gaussian components in the baseline case and BN in the HMM/BN case.

[1] The state labels are obtained from the Viterbi alignment, and the $X_{t-1}$ labels from a VQ codebook trained in advance.

As can be seen from Eqs. (3) and (4), the number of Gaussian components of the HMM/BN model depends on the VQ codebook size, and this number can become quite large. Indeed, using the BN topology of Fig. 3 with a VQ size of 128 resulted in more than 100,000 Gaussians. To reduce the number of mixtures and to avoid badly estimated parameters, we adopted a tied mixture model structure, where Gaussians are first clustered and then the components belonging to the same cluster are tied. In this way, for each VQ codebook size of 32, 64 and 128, we made four models, with the total number of Gaussians corresponding to that of the baseline models.

First, we evaluated the HMM/BN models in an LVCSR task. We used a 20k-word dictionary and a bi-gram language model trained on about 150,000 travel-related sentences. There were about 1.5% out-of-vocabulary words in the test data. Recognition results of the best performing HMM/BN models and the baseline MDL-SSS HMnet are shown in Fig. 5. In this figure, the HMM/BN with the BN topologies from Fig. 3 and Fig. 4 are denoted as BN1 and BN2, respectively. The numbers after the name indicate the size of the VQ codebook used. By using forced Viterbi alignment we obtained reference phoneme transcriptions of the test data used in the phoneme recognition experiments. Figure 6 shows the results. In both the LVCSR and phoneme recognition tasks, the HMM/BN model performed better than the baseline HMM.

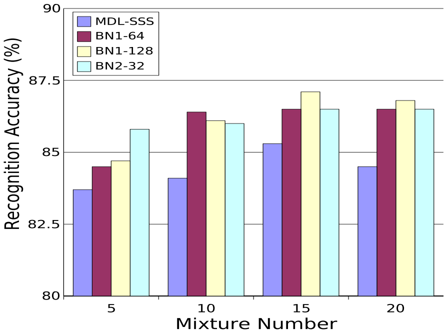


**Fig. 5**. Results of LVCSR experiments.

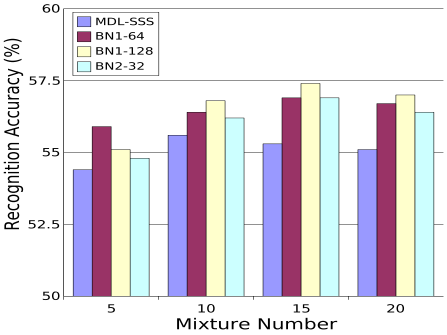


**Fig. 6**. Results of phoneme recognition experiments.

## 5. CONCLUSION

In this paper, we presented a method for modeling the successive frames dependency based on the BN framework. this method was implemented using the hybrid HMM/BN model, and it thus improved ASR system performance at no additional computational cost, since the resulting acoustic models have the same number of Gaussian components as the baseline HMM.

## 6. ACKNOWLEDGMENT

The research reported here was supported in part by the Ministry of Public Management, Home Affairs, Posts and Telecommunications of Japan.

## 7. REFERENCES

[1] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 35, pp. 1857–1869, Dec. 1989.

[2] V. Digalakis, J. Rohlicek, and M. Ostendorf, "A dynamical system approach to continuous speech recognition," in *Proc. ICASSP*, 1991, pp. 289–292.

[3] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, vol. 27, no. 1, pp. 65–78, Apr. 1992.

[4] C. J. Wellekens, "Explicit correlation in hidden Markov model for speech recognition," in *Proc. ICASSP*, 1987, pp. 384–387.

[5] K. Paliwal, "Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer," in *Proc. ICASSP*, 1993, vol. II, pp. 215–218.

[6] S. Takahashi, T. Matsuoka, Y. Minami, and K. Shikano, "Phoneme HMMs constrained by frame correlations," in *Proc. ICASSP*, 1993, vol. II, pp. 219–222.

[7] N. Kim and C. Un, "Frame-Correlated Hiddem Markov Model based on extended logarithmic pool," *IEEE SAP*, vol. 5, no. 2, pp. 149–160, mar 1997.

[8] G. Qing, Z. Fang, W. Jian, and W. Wenhu, "A new method used in HMM for modeling frame correlation," in *Proc. ICASSP*, 1999, pp. 169–172.

[9] T. Ogawa and T. Kobayashi, "Generalization of state-observation-dependency in partly hidden Markov models," in *Proc. ICSLP*, 2002, pp. 2673–2676.

[10] J. Bilmes, "Buried markov models: a graphical-modeling approach to automatic speech recognition," *Computer Speech and Language*, vol. 17, no. 2-3, pp. 213–231, 2003.

[11] K. Markov and S. Nakamura, "A hybrid HMM/BN acoustic model for automatic speech recognition," *IEICE Trans. Inf. & Syst.*, vol. E86-D, no. 3, pp. 438–445, 2003.

[12] K. Markov and S. Nakamura, "Hybrid HMM/BN LVCSR system integrating multiple acoustic features," in *Proc. ICASSP*, 2003, vol. I, pp. 888–891.

[13] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 8, 2004.