〔招待講演〕

HMM/BN 音響モデルの設計と実装

マルコフ・コンスタンティン　　中村　哲

ATR 音声言語コミュニケーション研究所　音声音響処理研究室
619-0288　京都府「けいはんな学園都市」光台 2-2-2
E-mail: {konstantin.markov,satoshi.nakamura}@atr.jp

あらまし　近年、従来の HMM を超えた音声モデリングの進むべき方向に関する研究が盛んに行われている。その一つの有効な手法として、ベイズネットワーク (BN) による音声モデルがあり、最近、ダイナミック BN を基とした認識システムと BN による音響モデルが提案された。我々ATR は、混合ガウス分布にかえ、BN によって HMM の状態確率分布をモデリングするハイブリット HMM/BN モデルを提案した。本稿では、ハイブリット HMM/BN 音響モデルのフレームワークに関し、特に設計と実装の問題に焦点をあてて、記述する。HMM/BN 学習は、ビタビ学習に基づく方法によって、BN 学習と HMM 状態遷移確率の更新という 2 つのステップを交互に行う。認識において、BN 推論はある特定の場合、混合ガウスの計算と同等となり、既存のデコーダに修正せずに、HMM/BN モデルの使用可能である。本稿で、音声認識システムへの HMM/BN モデルの適用例を 2 つ示す。様々な条件で、複数種類のタスクによる実験を行った結果、HMM/BN モデルが従来の混合ガウス HMM に比べ、常に良い性能が得られた。

キーワード　HMM/BN、音響モデル、ベイズネットワーク

【Invited】

Design and Implementation of HMM/BN Acoustic Models

Konstantin MARKOV and Satoshi NAKAMURA

Department of Acoustics and Speech Research, Spoken Language Translation Research Labs, ATR
Hikaridai 2-2-2, Keihanna Science City, Kyoto, 619-0288
E-mail: {konstantin.markov,satoshi.nakamura}@atr.jp

Abstract　In recent years, the number of studies investigating new directions in speech modeling that goes beyond the conventional HMM has increased considerably. One promising approach is to use Bayesian Networks (BN) as speech model. Full recognition systems based on Dynamic BN as well as acoustic models using BN have been proposed lately. Our group at ATR has been developing the hybrid HMM/BN model which is a HMM where the state probability distribution is modeled by a BN, instead of commonly used mixture of Gaussian functions. In this paper, we describe the hybrid HMM/BN acoustic modeling framework especially emphasizing on some model design and implementation issues. The HMM/BN training is based on the Viterbi training paradigm and consists of two alternating steps - BN training and HMM transitions update. For recognition, in some cases, BN inference is computationally equivalent to mixture of Gaussians which allows HMM/BN model to be used in existing decoders without any modification. We present two examples of HMM/BN model application in speech recognition systems. Evaluations under various conditions and for different tasks showed that the HMM/BN model gives consistently better performance than the standard mixture of Gaussians HMM.

Key words　HMM/BN, acoustic model, Bayesian network

1. Introduction

For many years, since the introduction of the HMM for speech recognition [1], [2], observations conditional distributions $P(x|q)$ for each state have been modeled most often by mixture of parametric probability density functions (pdf). Gaussian as well as Laplacian pdfs are commonly used for this purpose. Later, a hybrid HMM/NN systems were proposed [3] where Neural Networks are used to estimate HMM state likelihoods given input observation. In most of the cases, features extracted from speech spectrum form these observations. However, research in speech recognition has shown that using only these features is not enough to achieve high system performance. Thus, many researchers have tried to include additional features representing some other knowledge into their HMM systems. For example, in [4] multi-space probability distribution is proposed for modeling additional pitch information. But, in almost each case, different approach is taken depending on the properties of the additional feature.

Recently, the Bayesian Networks (BN) have attracted researchers attention as an alternative to the HMM. They can model complex joint probability distributions of many different (discrete and/or continuous) random variables in well structured and easy to represent way. Especially suitable for modeling temporal speech characteristics is the Dynamic BN (DBN) [5]~[7]. DBN is regarded as generalization of the HMM, which in addition to speech spectral information can easily incorporate additional knowledge, such as articulatory features, sub-band correlation, speaking style, etc. In [8], acoustic features are easily supplemented with pitch information within the framework of DBN. Another advantage of the Bayesian Networks is that additional features which are difficult to estimate reliably during recognition may be left hidden, i.e. unobservable. Despite these attractive properties of BN, their application in speech recognition is still limited to small recognition tasks like digit recognition [9]. The reason is that the existing algorithms for BN parameter learning and inference are not efficient enough and become computationally prohibitive for large vocabulary continuous recognition tasks.

The model described in this paper aims at utilizing advantages of both HMM and BN while being free from their drawbacks described above. In the HMM/BN, temporal characteristics of speech signal are modeled by HMM state transitions and the BN is used to model HMM state distributions. The advantage of this is that the existing methods for HMM design, training and recognition can be used without significant modifications since the HMM/BN behaves essentially as a conventional HMM.

2. Hybrid HMM/BN Model

In this section, we give a brief description of the hybrid HMM/BN model and provide details about its design, training and implementation.

2.1 Background

The HMM/BN model is a combination of an HMM and a Bayesian Network. Speech temporal characteristics 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.

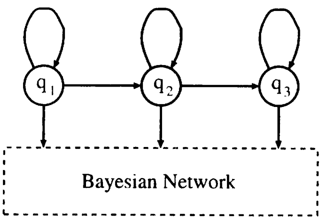

図 1 HMM/BN model structure. HMM transitions model speech temporal characteristics and BN represents states' probability distributions.

Structurally, the HMM/BN model is analogous to the hybrid HMM/NN model [10]. The difference is that instead of a Neural Network, the HMM is coupled with a BN.

By definition, a Bayesian Network 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. 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 [11]:

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

In practice, the HMM state distribution is often modeled with a mixture of Gaussian functions. It can be graphically represented by a BN with topology shown in Fig.2, where $M = \{m_j\}, j = 1, \ldots, K$ is a discrete variable representing mixture component index.

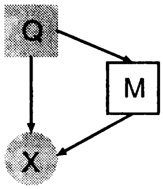

図 2 BN representing mixture of Gaussians.

The data likelihood $p(x_t|q_i)$ can be calculated using the BN joint probability function (Eq.(1)) as follows:

$$\begin{aligned} p(x_t|q_i) &= \\ &= \frac{P(x_t, q_i)}{P(q_i)} = \frac{\sum_{j=1}^{K} P(x_t, m_j, q_i)}{P(q_i)} \\ &= \frac{\sum_{j=1}^{K} P(x_t|m_j, q_i)P(m_j|q_i)P(q_i)}{P(q_i)} \\ &= \sum_{j=1}^{K} P(m_j|q_i)P(x_t|m_j, q_i) \qquad (2) \end{aligned}$$

If we replace $P(m_j|q_i)$ with w_{ji} and $P(x_t|m_j, q_i)$ with Gaussian function $N(x_t; \mu_{ji}, \Sigma_{ji})$, we get a standard mixture of Gaussians equation:

$$p(x_t|q_i) = \sum_{j=1}^{K} w_{ji} N(x_t; \mu_{ji}, \Sigma_{ji}) \qquad (3)$$

Figure 2 allows us to interpret the Gaussian mixture distribution in a different way. It shows that observation variable X depends not only on the state index but also on the variable M. However, M has no physical meaning. In this respect, Gaussian mixture learning is "blind" and does not reflect the way a speech signal is produced or at least does not account for the factors it depends on, such as speaker gender, environmental noises, communication channels, etc. Variable M, for example, could represent pitch value, articulatory configuration or some other parameter that effects the speech spectrum.

2.2 HMM/BN model design and training

The HMM/BN acoustic model design involves several main steps: choosing the speech unit to be modeled (phoneme, word, etc.); determining the the number of states per unit and the state topology; and choosing the BN structure. The first two steps are essentially the same as for the standard HMM. Therefore, the same methods and techniques are applicable in the HMM/BN case. Ideally, the BN structure should be learned automatically from the training data, but this is a very difficult task [12] and, usually, BN topology is chosen manually by taking into account the available data and the task at hand [13], [14]. The BN can have many variables corresponding to different speech features or variability factors. Dependencies are usually set according to prior knowledge or data correlation analysis. In this way, we can impose knowledge-based structure on the speech generation process and achieve a more precise speech model. Which BN variables should be hidden or observable depends on the available additional speech training data (pitch, articulatory observations, prosodic features, etc.) or high-level knowledge (speaker gender, environment factor, phoneme position, etc.). In case we don't have observations of some variable, we could assume it hidden. However, as in the Gaussian mixture example described above, in such cases, the training with the EM algorithm is "blind" and there is no guaranty that after the training this hidden variable represents the speech feature it is supposed to represent. So, it is better to avoid having hidden BN variables during training.

As in the case of the HMM/NN model, parameter learning of the HMM/BN is based on the Viterbi training paradigm and can be summarized in the following algorithm.

- Step 1. Initialization.
- Step 2. Viterbi alignment.
- Step 3. Update BN parameters.
- Step 4. Update HMM transition probabilities.
- Step 5. Stop or go to Step 2.

Although random initialization is possible, we first train a bootstrap HMM model and use its state structure and transition probabilities to initialize the HMM/BN. Thus, the main part of HMM/BN training becomes the BN parameter estimation. Since the state variable Q is observable, before BN training we need to obtain its values for each sample of X. This is done by the Viterbi alignment step. For BN parameter estimation, several methods are available. In the simplest case, when all variables are observable, maximum likelihood (ML) estimates can be computed in closed form[注1]. In a partially observed case, i.e. when some of the (discrete) variables are hidden, the Expectation-Maximization (EM) algorithm can be applied. After BN is trained and its parameters fixed, the HMM transition probabilities are re-estimated with a standard forward-backward algorithm. All of these steps are repeated until the convergence criterion is met. This can be an increase in data likelihood or simply a fixed number of iterations.

2.3 Implementation and decoding

The decoding in HMM based ASR systems is usually done in a frame synchronous manner using the Viterbi algorithm. As the difference between HMM/BN and HMM is in the way the state optput probability is calculated, the same decoding strategy can be applied. Depending on the BN complexity and the type of its variables, the output probability inference can be done in different ways. In general, we need to obtain $P(x_t, y_t^i, \ldots, y_t^j|q_t)$, where $y_t^i, \ldots, y_t^j$ are the instances of all additional observable variables. This requires a BN inference engine that should be coupled with the Viterbi decoder and feature extraction modules that will provide the $y_t^i, \ldots, y_t^j$ observations. In order to reduce the implementation costs, we can assume all additional variables hidden during recognition. Note that during the training, they are still observable. This is especially useful when during recognition the feature extraction is difficult or even impossible like in the case of articulatory features.

A further simplification can be achieved if all additional BN variables are chosen to be discrete. As we show in the HMM/BN application examples in this paper, the data likelihood inference can be reduced to a Gaussian mixture calculation. This is practically useful because in this case, the HMM/BN model is computationally equivalent to the HMM, and there is no need of inference engine or any modifications of the HMM decoder.

[注1] : This is true under the condition that continuous variables have no children.

3. HMM/BN application examples

3.1 Noisy speech recognition

When speech is contaminated by noise, speech feature vectors change their distributions and this change depends on the noise type as well as on the SNR value. Therefore, we can express these dependencies with a BN of the type shown in Fig.3.

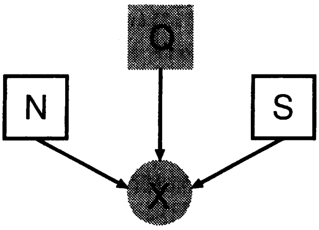

図 3 State BN with noise and SNR variables

Here, N and S are discrete variables representing noise type and SNR value. In most cases, prior probabilities $P(N)$ and $P(S)$ can reasonably be assumed equal for each type of noise and each SNR value and then:

$$P(x_t|q_i) = \frac{1}{N(n,s)} \sum_{n,s} P(x_t|N = n, S = s, q_i) \qquad (4)$$

Word models as well as sub-word models are made in the same way as in the conventional HMM case.

The evaluation experiments were performed using the AURORA2 database [15]. Of primary interest for us was to compare the HMM/BN system with Multi-condition trained HMM system. When training the BN, we labeled the training data by noise type and by SNR value and used the ML algorithm obtain parameters of each Gaussian $P(x_t|N = n, S = s, q_i)$. All other system parameters as feature vectors, word model state number and experimental conditions are kept the same. Note that, no adaptation or noise robust methods are used in our HMM/BN system. The main functional difference between the two systems is that HMM/BN system explores the hidden dependencies of speech features and noise.

Recognition results for test set A (same noise types as in training data) and test set B (different noises) are summarized in Table 1. As can be seen, the HMM/BN system performance is much higher for the closed noise condition test (A set) approaching the state-of-the-art results for this task obtained by much more complex systems. As for the B set condition, there is a degradation of the performance. This can be explained by the fact that no knowledge of dependencies for the new noises is available to the HMM/BN system in addition to the mismatch in the speech spectrum feature distributions. On the other hand, in the multi-condition HMM system, state Gaussian mixtures clearly do not model very well the complex distribution resulting from multiple noise and SNR conditions. However, this mismatch between data and model distributions has some smoothing effect which increases the model abilities to generalize over unseen data.

表 1 HMM and HMM/BN systems performance (%)

SNR	Test set A		Test set B	
	HMM	HMM/BN	HMM	HMM/BN
Clean	98.54	98.83	98.54	98.83
20 dB	97.52	98.12	96.96	97.26
15 dB	96.94	97.65	95.38	95.05
10 dB	94.59	96.04	92.58	90.27
5 dB	87.51	91.70	83.50	78.00
0 dB	59.84	76.11	58.91	48.70
-5 dB	23.46	35.79	23.86	3.18
Average*	87.29	91.92	85.46	81.85

* Calculated over values from 20dB to 0dB.

Another difference between the baseline HMM and the hybrid HMM/BN model is that latter has 17 times (4 noise types times 4 SNR values plus clean condition) more parameters. In order to prove that the better performance of the HMM/BN on test set A in not only due to increased number of parameters, we trained HMM model with the same number of parameters by increasing the mixture number. The overall average word accuracy rates of the three types of models is shown in Fig.4 where the newly trained model is denoted by HMM+.

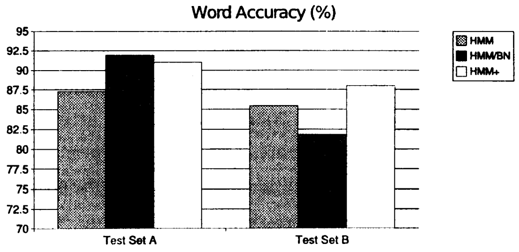

図 4 Comparison between baseline HMM, HMM/BN and HMM+ with the same number of parameters as the hybrid model.

This comparison clearly shows that the hybrid HMM/BN model is still better than the HMM+ for the known environments case due to better modeling of the noise-observation dependency which is learned explicitly. In contrast, the conventional HMM learns it implicitly. This advantage comes, however, at the expense of lesser generalization ability.

3.2 Articulatory and Acoustic Feature Integration

The articulatory data used in this experiment were collected by using the Electromagnetic Midsagittal Articulographic (EMA) system at NTT, Japan [16]. In the EMA system, a number of miniature coils are attached to points in the vocal tract. The subject's head is then placed in an electromagnetic field, allowing the movement of the coils to be inferred from the corresponding induced voltages. The output of the system is a set of x and y traces for articulatory movement. Acoustic signal and articulatory traces were recorded simultaneously. The sampling rate was 250 Hz for the articulatory channels and 12 kHz for the acoustic channel. All articulatory data were subsequently corrected for head movements and rotated to bring the occlusal plane into coincidence with the horizontal axis. The speech material consisted of 350 Japanese sentences that were read at normal speed by three male subjects. 300 sentences were selected for training and the rest 50 were used for evaluation.

Since both the acoustic and articulatory features are real valued vectors, direct integration using the HMM/BN model is difficult. In order to make this task feasible, we transform the articulatory parameters into discrete data by using Vector Quantization (VQ). Of course, some information will be lost, but this is a trade-off between the model's accuracy and its complexity.

The BN structure we used to combine the acoustic MFCC data and the articulatory parameters is shown in Fig.5.

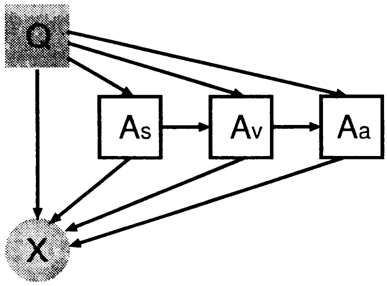

図 5 BN structure modeling dependencies between acoustic X, articulatory position A_s, velocity A_v and acceleration A_a variables.

Here, X depends on three articulatory variables: position, velocity and acceleration. In addition, the possible correlation between these variables is taken into account by making them dependent on each other. The output likelihood obtained from this BN structure is as follows:

$$\begin{aligned} p(x_t|q_i) = & \sum_{j=1}^{K_s}\sum_{n=1}^{K_v}\sum_{m=1}^{K_a} P(a_j^s|q_i) \cdot P(a_n^v|a_j^s, q_i) \cdot \\ & \cdot P(a_m^a|a_n^v, q_i) \cdot \\ & \cdot P(x_t|a_j^s, a_n^v, a_m^a, q_i) \end{aligned} \tag{5}$$

A closer look at this equation reveals that it is a mixture of Gaussians equation. Indeed, the first three terms of the right side are discrete probabilities, and their product $P(a_j^s|q_i)P(a_n^v|a_j^s, q_i)P(a_m^a|a_n^v, q_i)$ is simply the weight of the corresponding Gaussian mixture component $P(x_t|a_j^s, a_n^v, a_m^a \cdot q_i)$, which can be calculated in advance.

Since the BN articulatory variables are discrete, before HMM/BN training, all of the articulatory data had to be quantized. First, we reduced vector dimension to four by the principal component analysis (PCA) technique. The estimated information loss from this procedure in all cases was less than 15%. Then, for each articulatory parameter type (position, velocity, acceleration as well as concatenation of all three) we trained VQ codebooks of different sizes ranging from 4 to 1024. These codebooks were used to quantize the corresponding type of data, and their VQ labels served as articulatory observations for the BN training. Observations of the state variable Q were obtained using Viterbi alignment as described in section 2.2. Thus, all BNs were fully observable, and ML training was sufficient for the BN parameter estimation. Instead of initializing its parameters randomly, we used a HMM trained on acoustic data only as a bootstrap model which also serves as a baseline. Transition probabilities of this model were taken as initial values of the corresponding HMM/BN state transitions. The bootstrap model was also used in the Viterbi alignment step of the first training iteration to obtain good initial state segmentation. After such initialization, one or two training iterations were performed for all of the HMM/BN models. Since the number of states of both the baseline and HMM/BN models is the same, the only difference between them becomes the number of mixtures and the way they are trained.

The evaluation experiments were done using models trained on data from all three speakers. The test set consisted of each speaker's test data pooled together. In addition to the baseline model, we trained a HMM using concatenated acoustic and articulatory feature vectors. Note that this model reqires articulatory observations to be available during recognition. We will refer to this models as HMM(AC) and HMM (AC+ART) respectively.

The phoneme recognition accuracies obtained from all the three types of models are plotted in Fig.6. As the results show, the HMM/BN model performed much better than HMM(AC), achieving the same accuracy as the HMM(AC+ART). This suggests that the lack of additional speech features during recogntion can be effectively compensated by the correlation information learned during HMM/BN training.

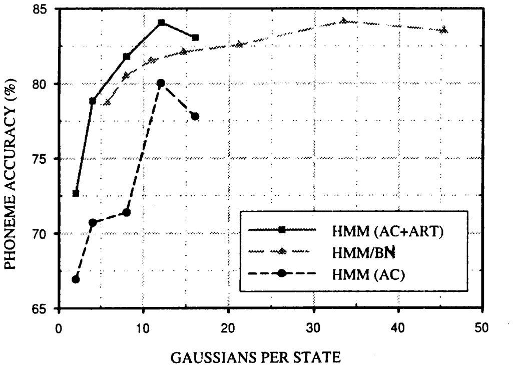

図 6 Performance of HMM/BN and two HMM model (multi-speaker case).

4. Conclusion

In this paper, we described the hybrid HMM/BN model and discussed some issues related to its design and implementation. Although this model can be regarded as a pure Bayesian Network, its structure allows simple algorithms to be used for training and recognition instead of general BN learning and inference methods which depending on the task may often become computationally intractable.

Since the HMM/BN has the same state topology as the HMM, the way we build acoustic models is not altered at all. The only difference is the need of BN training which in many cases can be reduced to an easy ML parameter estimation. The implementation of the HMM/BN can be simplified by forcing all the additional BN variables to be discrete. This way, the inference becomes equivalent to a Gaussian mixture computation.

As the provided examples of HMM/BN application show, even with a few additional variables and simple BN topologies, the hybrid model achieved better performance than the conventional Gaussian mixture HMM.

5. Acknowledgment

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

文 献

[1] S. E. Lavinson, L. R. Rabiner, and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, Apr. 1983.

[2] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, Feb. 1989.

[3] Herve Bourlard and Nelson Morgan, "A continuous speech recognition system embedding MLP into HMM," in *Advances in Neural Information Processing 2*, D. Touretzky, Ed., pp. 186–193. Morgan Kaufmann, 1990.

[4] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov Models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP*, 1999, pp. 229–232.

[5] T. Dean and K. Kanazawa, "Probabilistic temporal reasoning," in *AAAI*, 1988, pp. 524–528.

[6] Geoffrey Zweig and Stuard Russell, "Probabilistic modeling with Bayesian Networks for automatic speech recognition," in *Proc. ICSLP*, 1998, pp. 3010–3013.

[7] K. Daoudi, D. Fohr, and C. Antoine, "A new approach for multi-band speech recognition based on probabilistic graphical models," in *Proc. ICSLP*, 2000, vol. I, pp. 329–332.

[8] T. Stephenson, M. Mathew, and H. Bourlard, "Modeling auxiliary information in Bayesian Network based ASR," in *Proc. Eurospeech*, 2001, pp. 2765–2768.

[9] K. Daoudi, D. Fohr, and C. Antoine, "Continuous multiband speech recognition using Bayesian Networks," in *Proc. ASRU*, 2001.

[10] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, Boston, 1994.

[11] F. Jensen, *An introduction to Bayesian networks*, UCL Press, 1998.

[12] D. Heckerman, "A tutorial on learning with Bayesian Networks," in *Learning in Graphical Models*, M. Jordan, Ed., pp. 301–354. Kluwer Academic Publishers, 1998.

[13] 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.

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

[15] H. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance evaluations of Speech Recognition Systems under Noisy Conditions," in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*. Paris, France, Sept. 2000.

[16] T. Okadome and M. Honda, "Generation of articulatory movements by using a kinetic triphone model," *J. Acoust. Soc. Am.*, vol. 110, pp. 453–463, 2001.