HYBRID HMM/BN LVCSR SYSTEM INTEGRATING MULTIPLE ACOUSTIC FEATURES

Konstantin Markov and Satoshi Nakamura

ATR Spoken Language Translation Research Labs.
Keihanna Science City, Kyoto, Japan
{konstantin.markov,satoshi.nakamura}@atr.co.jp

ABSTRACT

In current HMM based speech recognition systems, it is difficult to supplement acoustic spectrum features with additional information such as pitch, gender, articulator positions, etc. On the other hand, Dynamic Bayesian Networks (DBN) allow for easy combination of different features and make use of conditional dependencies between them. However, lack of efficient algorithms has prevented their application in large vocabulary continuous speech recognition. The hybrid HMM/BN acoustic model, where HMM are used for modeling of temporal speech characteristics and state probability model is represented by BN, provides a trade off solution to the problem. In this paper we describe the HMM/BN acoustic model and LVCSR system built upon this model. In the HMM/BN model, in addition to speech observation variable, state BN has two more discrete variables representing speaker gender and pitch frequency. Evaluation results on WSJ database showed lower word error rate with respect to the same complexity conventional HMM acoustic model when there is enough training data to estimate reliable HMM/BN parameters.

1. INTRODUCTION

In current LVCSR systems, HMM state probability distributions are commonly modeled by mixture of Gaussians. Also, there are hybrid HMM/NN systems [1] where Neural Networks are used to estimate HMM state likelihoods given input observation. Many researchers have tried to include additional features like pitch, speaking style, articulatory positions into their HMM systems. For example, in [2] 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. There is no common, flexible enough framework to deal with this problem.

On the other hand, Dynamic Bayesian Networks (DBN) [3] provide efficient framework for modeling joint probability distributions of many variables enabling easy integration of multiple speech features. In some of the first reports on DBN in speech recognition, they were used as word models in isolated word recognition tasks [4, 5]. In these works, DBN are regarded as generalization of the HMM, which in addition to speech spectral information can easily incorporate additional knowledge, such as sub-band correlation, speaking style, etc. In [6], acoustic features integrated with pitch information within the framework of DBN. Despite these attractive properties of BN, their application in speech recognition is still limited to small, isolated word recognition tasks. The reason is that existing algorithms for BN parameter learning and inference are not practically suitable for continuous speech recognition (CSR) and especially large vocabulary CSR tasks. Although, an extension of the DBN word model allowing recognition of continuously spoken digits was reported in [7, 8], increasing task vocabulary even to a few hundred words would be computationally intractable.

The hybrid HMM/BN acoustic model we proposed recently [9] aims at utilizing advantages of both HMM and BN while being free from their drawbacks described above. In this model, HMM and BN are combined together in one model where temporal characteristics of speech signal are modeled by HMM state transitions and the BN is used to model HMM state distributions. There is a two level hierarchy in which the BN is at the lower level and the HMM stays at the top level. The advantage of this is that existing recognition algorithms can be used without any modification since this model behaves as a conventional HMM and can be used to model both word and sub-word units which is essential for large vocabulary systems. Nevertheless, additional features (variables) can be easily integrated in the state BN.

2. THE HYBRID HMM/BN MODEL

2.1. Definitions

The hybrid HMM/BN model is a combination of HMM and BN, where temporal characteristics of speech signal are modeled by HMM state transitions and HMM state probability density is modeled by Bayesian Network. The HMM/BN model structure is shown in Fig. (1).

This model is described by two sets of probabilities: HMM transition probabilities $P(q_j|q_i)$ and joint probabil-

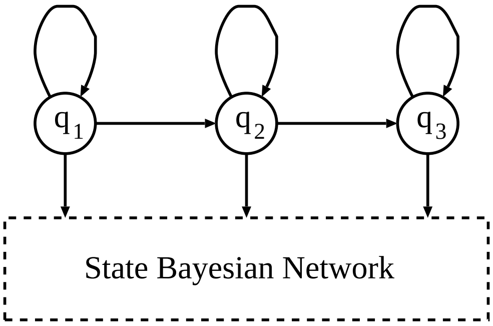


Fig. 1. HMM/BN model structure.

ity distribution of the Bayesian Network $P(X_1, \dots, X_K)$, where $X_i, i = 1, \dots, K$ are the BN variables. The BN joint pdf can be factorized as [10]:

$$P(X_1, \dots, X_K) = \prod_{i=1}^{K} P(X_i|Pa(X_i)) \qquad (1)$$

Here, $Pa(X_i)$ denotes the *parents* of the variable X_i.

Some examples of possible state BN structures are shown in Fig. 2. BN variables can be both discrete and continuous variables and some of them can be hidden. For example, variable X_1 can represent HMM state, X_2 - speech spectrum observation vector, and others can represent additional speech characteristics as pitch, articulators positions, speaker gender, etc. Dependency between two variables is denoted by an arc and is described by a conditional probability function. Since it is difficult to learn such dependency automatically, most often BN structure is designed manually based on our knowledge about the data.

2.2. HMM/BN model training

For HMM/BN model training, the same approach as for HMM/NN training [11] can be adopted. It is based on the Viterbi training algorithm and proceeds in several steps:

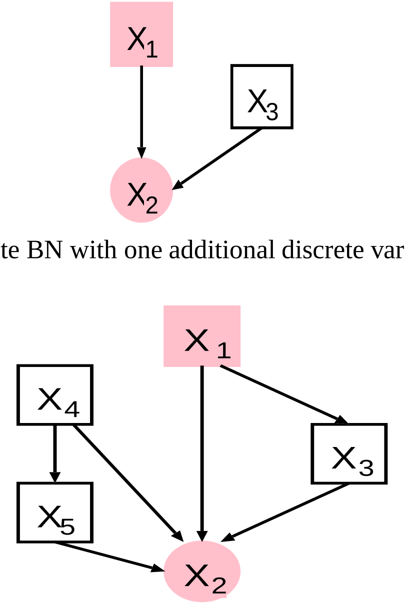
Step 1. Initialization: Choose topologies of HMMs and state Bayesian network and initialize model parameters (either randomly or performing Step2 using conventional bootstrap HMM recognizer).

Step 2. Viterbi alignment: Perform Viterbi alignment of the training data. This gives a time-aligned state segmentation. It is used to produce training data for the state Bayesian network.

Step 3. BN training: Train State BN using maximum likelihood algorithm if all variables are observable or EM algorithm if some of them are hidden.

Step 4. HMM transition probabilities training: Perform standard forward-backward training of HMM transition probabilities.

Step 5. Convergence check: Check whether convergence criterion is met (either specified number of iterations or data likelihood increase threshold) and go to Step 2 or finish training.



a) State BN with one additional discrete variable.

b) State BN with more complex structure.

Fig. 2. Possible state BN structures.

2.3. Recognition with the HMM/BN model

When doing recognition with this HMM/BN model, as in the case of conventional HMM, the standard Viterbi decoding algorithm is used. Here, we need to calculate input observation likelihood $P(x_t|Q)$ for each state $Q = q_{ij}$ where i is the HMM index and j is the state index of the i^{th} HMM. For simple state BN, $P(x_t|Q)$ can be derived analytically using "brute force" inference method. For more complex state BN, standard exact inference algorithms (as for example "junction tree" algorithm [10]) can be used.

3. LVCSR SYSTEM WITH HMM/BN MODEL

For large vocabulary speech recognition systems it is essential to use sub-word unit modeling approach where each HMM represents short speech unit: phone for example. Also, in order to achieve good recognition performance, it is preferable to use context dependent models. In our system, we adopted crossword triphone HMM/BN models with 3-state left-to-right topology. As for the state Bayesian Network, we used the structure shown in Fig. 3 where variables Q and X are observable and represent discrete HMM state and continuous observation vector. The other two discrete variables - F and G represent two additional speech features: pitch frequency and speaker gender.

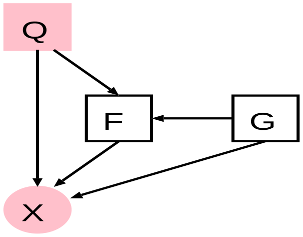


Fig. 3. State BN structure with pitch frequency F and speaker gender G as additional variables.

When building a LVCSR system, we always face the problem of limited training data. There are various methods to deal with this problem, most widely used of which is the state tying method. Such method for HMM/BN model, however, does not exist yet and its development would require considerable time. Our solution was to initialize HMM/BN with already tied state bootstrap HMM and use its tying scheme. Of course, such state tying would not be optimal with respect to HMM/BN model, but still greatly reduces the limited data problem. Once initialized, HMM/BN model is trained as described in section 2.2.

State output probability can be calculated in closed form from the joint pdf for the BN of Fig. 3 which according to Eq.(1) is:

$$P(X, F, G, Q) = P(X|F, G, Q)P(F|G, Q)P(G)P(Q) \tag{2}$$

Since X is continuous variable, $P(X|F, G, Q)$ is modeled by Gaussian density. F and G are discrete and therefore $P(F|G, Q)$ is represented by conditional probability table (CPT). Assuming that speaker gender and pitch frequency of the training data are known, during training all BN variables are fully observable and Gaussian parameters are estimated using ML algorithm, while CPTs are obtained from sample counts.

During recognition, speaker gender is usually unknown and variable G can be left hidden. Furthermore, it is equally probable that the test speaker will be either male or female, so prior probabilities $P(G = g), g = \{male, female\}$ are both set to 0.5 As for the pitch variable, it can be hidden as well as observable. The case when pitch variable is hidden is especially interesting because during recognition we don't have to estimate pitch frequency. Then, state output probability is:

$$\begin{aligned} P(X|Q) &= \frac{P(X, Q)}{P(Q)} \\ &= \frac{\sum_{f,g} P(X, F = f, G = g, Q)}{P(Q)} \\ &= \sum_{f,g} \frac{P(F = f|G = g, Q)}{2} P(X|F = f, G = g, Q) \end{aligned} \tag{3}$$

If the pitch frequency is available during recognition, then we can use this information and calculate conditional probability of X with respect to both state Q and pitch F:

$$\begin{aligned} P(X|F, Q) &= \frac{P(X, F, Q)}{P(F, Q)} \\ &= \frac{\sum_g P(X, F, G = g, Q)}{P(F|Q)P(Q)} \\ &= \frac{\sum_g P(X|F, G = g, Q)P(F|G = g, Q)}{2\sum_g P(F|G = g, Q)} \end{aligned} \tag{4}$$

Analyzing Eq.(3) we can see that it is same as the conventional mixture of Gaussians equation, where $P(F = f|G = g, Q)/2$ are the mixture weights and $P(X|F = f, G = g, Q)$ are the Gaussian components. In this case, the HMM/BN structure is equivalent to the standard HMM and therefore existing HMM decoders can work with HMM/BN without any modification. However, the difference between HMM/BN and HMM lays in the way they are trained. As for the Eq.(4), it also can be viewed as mixture of Gaussians, however, which Gaussians are to be used depends on the pitch variable value.

4. EXPERIMENTS

We evaluated our HMM/BN based LVCSR system using WSJ database. The experimental setup followed closely the HUB2 (Nov93) evaluation specifications. For training we used SI-284 training set. Language model was standard bigram provided for the HUB2 evaluation. The test set consists of 215 utterances with 0% OOV and 5000 word dictionary.

Speech data were transformed into 39 dimensional feature vectors (pow + 12 MFCC, Δ, $\Delta\Delta$) from 20ms long frames with 10ms shift. Pitch frequency was extracted from speech signal such that for each feature vector there was a corresponding pitch value. Zero pitch was set for silence and non-voiced parts. From all non-zero pitch data two VQ codebooks were trained with 3 and 7 centroids respectively. Later, a zero centroid was added manually to each of the codebooks, so the number centroids became 4 and 8. All pitch data (train and test) were then quantized and codebook labels were obtained. Thus, we had each speech feature vector labeled with pitch and speaker gender label.

Using the HTK speech toolkit we trained three tied state crossword triphone bootstrap models HMM with 10071, 7870 and 5666 states respectively. They were used for initialization of three HMM/BN models. During HMM/BN training, data aligned for each state were divided into 8 or 16 sets in accordance to their labels and from each set one Gaussian pdf parameters were calculated using ML method. If the number of vectors in certain set was less than a threshold, Gaussian was not made thus, effectively reducing the number of mixtures for the state. In some cases, data from both genders for a given pitch label were pooled in order to exceed this threshold. HMM/BN model training was stopped after 5 iterations.

Decoding with HMM/BN model was done using the same HTK software without any modification, so the gender and pitch variables were left hidden and state output probability takes the form of Eq.(3). Table 1 and Table 2 show the results using 4 and 8 level quantized pitch data respectively. In HMM/BN case, since the mixture number varies from state to state, the average number of Gaussians per state is given. For comparison, results of similar complexity standard HMM model are shown.

Table 1. Results using 4 level CB quantized pitch data

Model	states	mix/state	WER (%)
HMM	10071	4	12.4
HMM/BN		3.7	11.8
HMM	7850	4	14.7
HMM/BN		4.1	14.0
HMM	5666	5	13.6
HMM/BN		4.5	12.4

Table 2. Results using 8 level CB quantized pitch data

Model	states	mix/state	WER (%)
HMM	10071	6	11.2
HMM/BN		5.6	12.1
HMM	7850	6	13.3
HMM/BN		5.9	13.8
HMM	5666	7	12.5
HMM/BN		6.6	12.8

These are first results we got and they are definitely not conclusive. Nevertheless, they show that sparse training data is a serious problem for the HMM/BN model. As we mentioned in section 3, we used state tying derived from one mixture per state conventional HMM and is clearly not suitable for the HMM/BN model and obtained results confirm this. For the case of 4 level quantized pitch data and less number of model parameters, we got better results than the baseline HMM and the relative improvement is highest for the models with smallest state number. On the other hand, with 8 level quantized data, HMM/BN did not improve the baseline HMM performance, but for the case of smallest state number, WERs are almost the same. Also, the HMM/BN model using 8 level quantized pitch data is more sensitive to pitch extraction errors than the one which uses 4 level quantized pitch data.

5. CONCLUSION

We have presented a LVCSR system based on the hybrid HMM/BN model which allows easy integration of additional speech information. In our system, state BN had two additional variables representing pitch and speaker gender. Evaluation on WSJ database showed that HMM/BN model performs better than conventional HMM of similar complexity given there is enough training data. Also, the state tying algorithm for HMM seems to produce inefficient tying with respect to the HMM/BN model.

Future work will include research on new HMM/BN state tying algorithm and new HMM/BN model evaluations with different data and various state BN structures.

6. ACKNOWLEDGMENT

This research was supported in part by the Telecommunications Advancement Organization of Japan.

7. REFERENCES

[1] 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.

[2] 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.

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

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

[5] 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.

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

[7] K. Daoudi, D. Fohr, and C. Antoine, "Continuous multi-band speech recognition using Bayesian Networks," in *Proc. ASRU*, 2001.

[8] G. Zweig, J. Bilmes, T. Richardson, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz, J. Torres, and B. Byrne, "Structurally discriminative Graphical Models for automatic speech recognition - Results from the 2001 Johns Hopkins summer workshop," in *Proc. ICASSP*, 2002, vol. I, pp. 93–96.

[9] K. Markov and S. Nakamura, "Modeling HMM state distributions with Bayesian Networks," in *Proc. ICSLP*, 2002, vol. 2, pp. 1013–1016.

[10] F. V. Jensen, *An Introduction to Bayesian Networks*, UCL Press, London, 1996.

[11] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionist Probability Estimators in HMM Speech Recognition," *IEEE Trans. SAP*, vol. 2, no. 1, Part II, pp. 161–173, Jan. 1994.