MODELING HMM STATE DISTRIBUTIONS WITH BAYESIAN NETWORKS

Konstantin Markov, Satoshi Nakamura

ATR Spoken Language Translation Research Labs,
2-2-2 Hikaridai, Seika-cho, Kyoto 619-0288, 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, Bayesian Networks (BN) allow for easy combination of different continuous as well as discrete features by exploring conditional dependencies between them. However, the lack of efficient algorithms has limited their application in continuous speech recognition. In this paper we propose new acoustic model, where HMM are used for modeling of temporal speech characteristics and state probability model is represented by BN. In our experimental system based on HMM/BN model, in addition to speech observation variable, state BN has two more (hidden) variables representing noise type and SNR value. Evaluation results on AURORA2 database showed 36.4% word error rate reduction for closed noise test without using any model adaptation or noise robust methods.

1. INTRODUCTION

For many years, since the introduction of the HMM for speech recognition [1, 2], observations conditional distributions $P(y|Q)$ for each state Q have been modeled by mixture of probability density functions (discrete HMM are not considered here). 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] milti-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.

Recently, the Bayesian Networks (BN) have attracted researchers attention as an alternative to the HMM. BN are well known and studied in Artificial Intelligence research field, but in speech recognition, they are relatively new research topic. Bayesian Networks 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 are the Dynamic BN (DBN)[5]. In some of the first reports on DBN in speech recognition, they were used as word models in isolated word recognition tasks [6, 7]. 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 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, 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 [9], increasing task vocabulary even to a few hundred words would be computationally prohibitive.

The method we are proposing in this paper aims at utilizing advantages of both HMM and BN while being free from their drawbacks described above. In our approach, HMM and BN are combined together in one hybrid HMM/BN model. In this model, 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.

This paper is organized as follows. Section 2 describes in detail our hybrid HMM/BN model and several possible BN structures. In Section 3, we show how to include additional information about noise type and noise SNR using HMM/BN framework and in Section 4 we describe the evaluation of our system on AURORA2 task. Section 5 offers discussion about our approach and some conclusions are drawn in Section 6.

2. THE HYBRID HMM/BN MODEL

In most cases, the use of BN in speech recognition has been based on the idea to represent HMM as a Dynamic BN. Such representation is shown in Fig. 1 where Q_t is the state variable and Y_t is the continuous observation variable at time $t = 1, 2, 3, 4, \ldots$. Arcs represent probabilistic dependencies between variables and it is easy to see that arcs between state instances represent HMM transition probabilities and arcs between state and observation instances represent HMM state conditional distributions. In figures below, variables shown in squares are discrete and variables in circles are continuous. Shaded circles/squares denote observable variables.

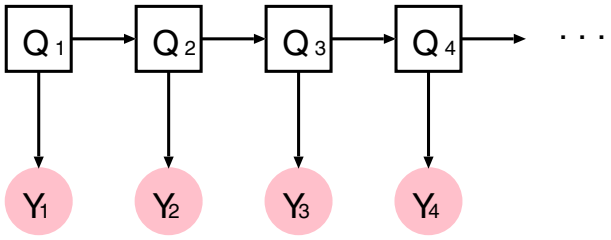


Fig. 1. Representation of HMM as DBN

Representing HMM as DBN requires BN inference algorithms to be used when we need to obtain the likelihood of input observation sequence $P(Y|M)$ where $Y = y_1, \ldots, y_T$ and M is our model. In doing so, the size of the network is adjusted to the size T of the input sequence and then $P(Y|M)$ is inferred from the whole network.

Let's now imaginary break arcs between state nodes. Then we get multiple, independent BN as shown in Fig. 2 corresponding to each time t. If we let the time transitions (broken arcs) be governed by conventional HMM, and assign those BNs to appropriate HMM states we can drop the time index and since all BNs have the same structure we can represent them as single BN shown in Fig. 3 where the variable Q takes values of state indexes (S_{ij}) of all HMMs in the acoustic model and the state probability distributions $P(Y|Q = S_{ij})$ are represented by the arc.

In other words, we modified the conventional HMM to have a BN as state distribution model instead of mixture of Gaussians. Combining HMM and BN in this manner

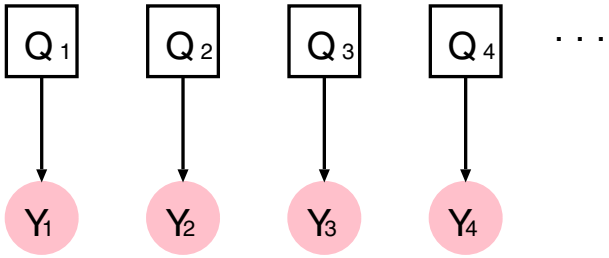


Fig. 2. Multiple BN for each time t

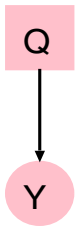


Fig. 3. State BN

makes the HMM/BN model hierarchical, where BN is at the bottom level and HMM is at the top level. Note that, the state variable Q (Fig. 3) has become observable for the BN, but at the upper HMM level it is still hidden.

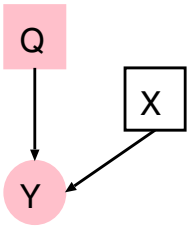
The state BN, can easily be extended to include other random variables representing additional knowledge. The graphical structure of the extended BN can be imposed according to our knowledge of the relationship between variables, rather than be learned from data, which is not a trivial task. Some possible structures of extended sate BN are shown in Fig. 4. For example, the variable X in this figure can represent the environment noise type and the other W and Z variables can represent speaker id and his/her native language.

When doing recognition with this HMM/BN model, as in the case of conventional HMM, we need to calculate the $P(y|Q)$ for each state $Q = q_{ij}$ where i is the HMM index and j is the state index of the i^{th} HMM. We can infer this value from the BN probability model and there are many exact as well as approximate inference algorithms to do this. For simple BN, as that of Fig. 4.a, even "brute force" method is applicable. The joint probability model for this BN can be expressed by chain rule as follows:

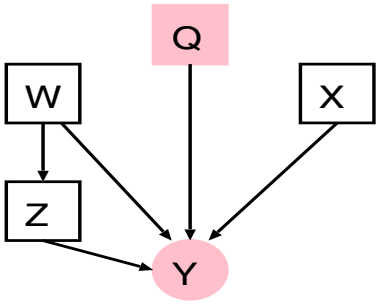
$$P(Y, X, Q) = P(Y|X, Q) * P(X|Q) * P(Q) \quad (1)$$

and since X and Q are independent variables (there are no arcs linking them), above equation can be rewritten as:

$$P(Y, X, Q) = P(Y|X, Q) * P(X) * P(Q) \quad (2)$$



a) State BN with one additional discrete variable.



b) State BN with more complex structure.

Fig. 4. Possible state BN structures.

Then, probability of interest $P(Y|Q)$ is calculated by marginalization over X:

$$\begin{aligned} P(Y|Q) &= \frac{P(Y,Q)}{P(Q)} = \frac{\sum_x P(Y,X=x,Q)}{P(Q)} \\ &= \frac{\sum_x P(Y|X=x,Q) * P(X=x) * P(Q)}{P(Q)} \\ &= \sum_x P(Y|X=x,Q) * P(X=x) \end{aligned} \quad (3)$$

In many practical cases, we can assume that $P(X)$ is the same for all $X = x$ and then Eq.(3) reduces to:

$$P(Y|Q) = \frac{1}{N(x)} \sum_x P(Y|X=x,Q) \quad (4)$$

where $N(x)$ is the number of values X can take.

Training of the BN parameters can be done independently for each state, in much the same way as conventional HMM state parameters are trained. For simple BN, as in our example (Fig. 4.a), we need to estimate only $P(Y|X = x, Q)$ for each x. If we use Gaussian mixture pdfs to represent each $P(Y|X = x, Q)$, their parameters (means, weights and covariances) can be estimated by the standard EM algorithm from data labeled for each condition x. For more complex BN, BN training algorithms should be integrated with the Baum-Welch training. However, this topic is out of the scope of this paper and will be researched in the future.

3. HMM/BN MODEL IN NOISY SPEECH RECOGNITION SYSTEM

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 state BN of the type shown in Fig. 5. Here, N and S are hidden discrete variables representing noise type and SNR value. In this case, the state likelihood can be expressed analytically in the same way as we derived Eq.(3). 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:

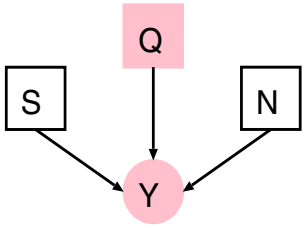


Fig. 5. State BN with noise and SNR variables

$$P(Y|Q) = \frac{1}{N(n,s)} \sum_{n,s} P(Y|N=n,S=s,Q) \quad (5)$$

Word models as well as sub-word models are made in the same way as in the conventional HMM case. Decoding also can be performed as in standard HMM based systems without any changes in the decoder.

4. EVALUATION ON AURORA2 TASK

In these experiments, we followed closely the evaluation scenario suggested by the official AURORA2 task. Of primary interest for us was to compare the HMM/BN system with Multi-condition trained HMM system. When training the HMM/BN state conditional distributions, we divided the training data by noise type and by SNR value and used HTK to train parameters for each condition separately. 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 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.

Table 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.

5. DISCUSSION

Obviously, the proposed hybrid HMM/BN model is applicable not only in noisy speech recognition systems, but in many other cases, where performance can benefit from additional observable or hidden features. This approach is more like a general framework for increasing modeling capabilities of the system by combining together features from different spaces and exploring dependencies between them. Especially interesting is the possibility to infer the probabilities of the hidden variables of the BN. This way, HMM/BN system can be used for recognition of those additional parameters. For example, if an additional hidden variable X represents language in a multi-lingual system, we can calculate $P(X|Q)$ for each frame and accumulate these probabilities over the input utterance. Then, $x = \arg\max_x P(x|Q_S)$, where Q_S is the best hypothesis state sequence, shows the most probable language the utterance has been spoken in. Thus, in addition to recognizing multi-lingual speech, such system can perform language recognition as well.

6. CONCLUSION

We have proposed a method for combining HMM and BN in a single model which benefits from strengths of both HMM and BN. The hybrid HMM/BN model allows for easy addition of other information in the speech recognition systems increasing their performance at minimal cost. Furthermore, HMM/BN model can represent sub-word phonetic units like the conventional HMM. This way, it becomes possible to use the BN framework in large vocabulary continuous speech recognition. Experimental evaluation of the method in noisy speech recognition task showed that adding noise type and SNR values as additional parameters and exploring dependency between them and the spectrum feature parameters resulted in 36.4% less errors in the AURORA2 task.

7. ACKNOWLEDGMENT

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus".

8. REFERENCES

[1] S. E. Levinson, 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, 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, pp. 257–285, Feb. 1989.

[3] H. Bourlard and N. 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*, pp. 229–232, 1999.

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

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

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

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

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