Structured Models Design for Improved Speech Recognition

Konstantin Markov
Human Interface Laboratory, The University of Aizu
Aizu-Wakamatsu City, Fukushima, 965-8580 Japan
e·mail: markov@u-aizu.ac.jp

Abstract

Current statistical approach to model building is often called ”ignorance-based modeling” in the sense that any unwanted variability is assumed residual and is supposed to be accommodated with in the variances of the probability density functions (pdfs). This allows limited degree of prior model structure to be imposed. However, structure that explains systematic variations reduces the uncertainty which in turn increases the predictability and therefore, the model's performance. Bayesian Networks (BN) are an excellent tool which can efficiently and flexibly encode any structure through their topology, but it soon turned out that it's difficult to build large systems because of DBN's poor scalability. Our approach is to keep the hierarchical structure of the traditional ASR systems and use different, small BNs to model pdfs at different hierarchical levels independently. For example, at the lowest level, we use the BN to represent the HMM state pdf. At the next (phonetic) model level, we use the BN to factor the underlying pdf. We describe several examples of ASR models built using this approach and show that consistent performance improvement can be achieved in various tasks and settings.

Keywords: Bayesian networks, Hybrid HMM/BN model, Bayesian Multi-phone, Speech recognition.

1. Introduction

One of the fundamental technologies for achieving a speech-oriented interface is automatic speech recognition (ASR). The goal is to develop an intelligent machine that can automatically recognize naturally spoken words uttered by humans. However, extracting the underlying linguistic message from a complex acoustic signal is not an easy task due to many sources of variability contained in the signal [1]. With the introduction of statistical approaches and especially the hidden Markov models (HMM) [2, 3], a big change in the modeling method ocurred as well as in the whole system design paradigm. Statistical learning algorithms deal nicely with random variations and allow to take into account only those we are interested in. For example, in speech recognition we are interested in variations coming from the linguistical content, but not in those coming from individual speakers. In contrast, in the text-independent speaker recognition task, speaker variations are the important ones and phonetic variability is considered as unavoidable “noise”. Statistical models accomodate this “noise” in the variance of the probability density functions and thus we can just ignore it. This way, in order to build a model, no knowledge about the unvanted variability is necessary which leads to simple models with limited degree of structure and is known as “ignorance based modeling” approach [4]. However, there are some systematic variations, such as those coming from the environment, speaking style or speaker gender, and whose source is easily identified. We can use this additional knowledge by adding some model structure in a way that it explains those systematic variations. This will reduce the uncertainty, so that our models will have higher predictive power, which in turn means that thay will have better performance.

Bayesian networks (BN) [5, 6] are very well suited for modeling structured probability density functions by encoding the structure in the network topology. A special flavor of BN, called Dynamic BN (DBN) [7] is particularly suited for time-varying signals like speech. The DBN is very flexible and can express many known models. Fig. 1, for example, shows the conventinal HMM as a DBN.

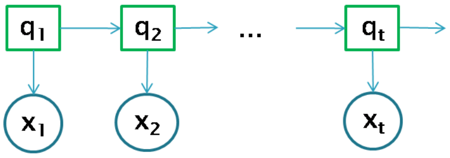

Figure 1: Traditional HMM represented as a DBN.

Many researchers have used DBN in their studies and encouraging results have been reported [8, 9, 10], but it turned out difficult to build large scale ASR systems entirely on DBN. The problem is that with the linear increase of the variable number, the computational complexity increases exponentially and often it becomes impractical or even intractable.

In this paper, we describe a different approach, where we tried to achieve the best trade-off between the superior expressive power of the BN and the practical efficiency of HMM. The main idea is to keep the hierarchical structure of the conventional ASR system intact and to use small independent BN to model probability distributions at each hierarchical layer as shown in Fig. 2. At the lowest HMM state layer, all state pdfs are represented by a single BN. This approach is also known as hybrid HMM/BN model [11, 12, 13]. At the next layer, we have phoneme models and we use BN to decompose wide context dependent models into several less context dependent or context independent ones. This approach is very helpful when the amount of thaining data is limited and doesn't allow big number of different models to be sufficuently trained. This method is sometimes called Bayesain multiphone. One of the most static models in the ASR system is the pronunciation model or lexicon. In most cases, it is just a table with pronunciations of each word. Using BN, however, we can turn the lexicon into a probabilistic pronunciation model and use it at the next layer of the system [14]. Curently, we are working on language model prepresentation by a BN in a feasible and computationally reasonable way. Next sections provide some details about the above mentioned models and in Section 5. we describe several evaluation experiments and show that in all cases there is a benefit of applying BN based models.

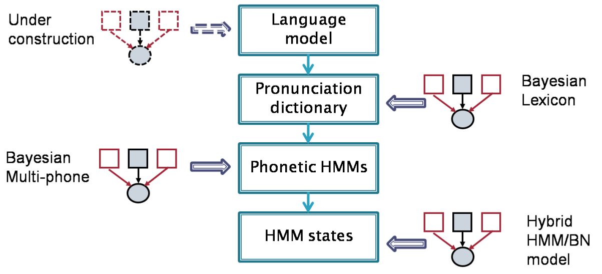

Figure 2: ASR system hieharchy and layer dependent structured model implemetation.

2. Hybrid HMM/BN model

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 [15]. A block diagram of the HMM/BN is shown in Fig.3.

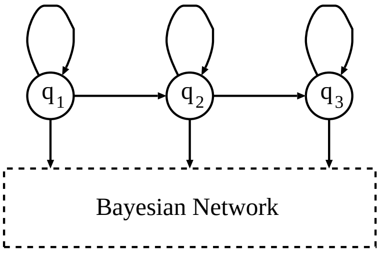

Figure 3: Hybrid HMM/BN model structure.

This model is described by two sets of probabilities: HMM transition probabilities $P(q_j|q_i)$ and joint probability distribution of the Bayesian Network $P(X_1,,X_k)$, where $X_i, i = 1,\ldots,K$ are the BN variables. The BN joint probability density function (PDF) can be factorized as [6]:

$$P(X_1,X_2\ldots X_K) = \prod_{i=1}^{K} P(X_i|Pa(X_i)) \qquad (1)$$

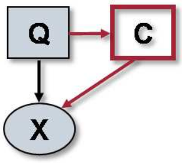

Figure 4: Hybrid HMM/BN model with one additional variable.

Figure 4 shows an example of a simple state BN structure with three variables. By circle we denote continuous variables, and the squares are used for discrete ones. Therefore, Q and C are discrete and X is continuous. The arcs represent dependencies between parent and child nodes which can be modeled by Conditional Probability Tables (CPT) if the child is discrete or by Gaussian pdf if the child is continuous. State output probability for the BN of Fig.4 can be calculated from the joint PDF in a closed form. According to Eq.(1:

$$P(X,Q,C) = P(X|C,Q)P(C|Q)P(Q) \qquad (2)$$

If all the BN variables are observable, then state output probability is just $P(X|C,Q)$ which is one of the BN parameters. However, when the additional variable C is hidden, are looking for $P(X|Q)$:

$$\begin{aligned} P(X|Q) &= \frac{P(X,Q)}{P(Q)} \\ &= \frac{\sum_c P(X,C=c,Q)}{P(Q)} \\ &= \sum_c P(C=c|Q)P(X|C=c,Q) \end{aligned} \tag{3}$$

Analyzing Eq.(3) we can see that it is same as the conventional mixture of Gaussians equation, where $P(C=c|Q)$ are the mixture weights and $P(X|C=c,Q)$ 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.

3. Bayesian Multi-phone

Let us consider a simple case where there are two additional variables, L and R representing the left and right phone context. The causal relationship between X, C, R, and L is described by the BN in Fig.5.

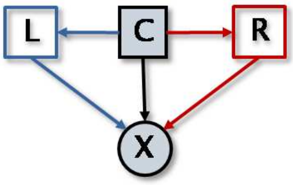

Figure 5: BN representing a triphone model.

By performing graphical transformations we can obtain a junction tree [16, 17] as shown in Fig.6.

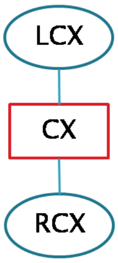

Figure 6: Junction tree corresponding to BN from Fig. 5.

The joint probability distribution is then defined as the product of all cluster potentials divided by the product of the separator potentials [6] and becomes:

$$P(X,C,L,R) = \frac{P(L,C,X)P(R,C,X)}{P(C,X)} \tag{4}$$

This indicates a new way of representing the joint probability function, $P(X,C,L,R)$, as a composition of several smaller joint probability functions $P(L,C,X)$ and $P(R,C,X)$, which leads to:

$$P(X|C,L,R) = \frac{P(X|L,C)P(X|R,C)}{P(X|C)} \tag{5}$$

and shows that the calculation of triphone probability can be decomposed into calculation of two bi-phone probabilities (for the left and right bi-phone) and one monophone probability. Similarly, we can obtain the relation for a pentaphone model [18, 19]:

$$\begin{aligned} & P(X|C,L1,L2,R1,R2) \\ &= \frac{P(X|L1,L2,C)P(X|R1,R2,C)}{P(X|C)} \end{aligned} \tag{6}$$

4. Bayesian Lexicon

Applying the graphical framework to the pronunciation model, we focus on predicting the realized phone label of conversational speech (the surface form), given the expected phone from the canonical dictionary (the baseform). As shown in Fig. 7, the baseform in denoted as B, S is the realized surface form, and L, R, P, and D are the additional knowledge sources, which are defined as follows:

- Preceding baseform phoneme contexts (L)
- Succeeding baseform phoneme contexts (R)
- Position of baseform phoneme in words (P)
- Previous surface phoneme condition (D) - deleted/not deleted.

The BN joint probability then becomes:

$$\begin{aligned} & P(D,S,P,L,R,B) \\ &= P(S|D,P,L,R,B)P(L|B)P(R|B) \\ & \quad P(P|B)P(D|B) \end{aligned} \tag{7}$$

from where we can easily find the expression for the surface form probability:

$$P(S|B) = \sum_{D,P,L,R} P(S|D,P,L,R,B) P(L|B)P(R|B)(P|B)P(D|B) \quad (8)$$

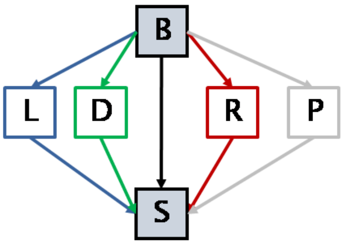

Figure 7: BN model of a single phone.

5. Experiments

In this section, we briefly decsribe several experiments where the performance of each of the methods above is evaluated and compared with other commonly used techniques.

5.1. Gender Dependent ASR

In this experiment, we use very simple BN to model HMM state probability function, in which there is only one additional variable G representing the speaker gender [20] as shown in Fig. 8.

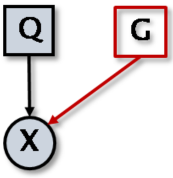

Figure 8: BN topology when speaker gender is represented by the additional variable G.

The traditional approach is to separate the training data into male and female portions and train two models with the same structure for each speaker gender. During recognition, these models are used in parallel. In our case, however, this is not necessary as the speaker gender information is already embedded into the state pdf. For comparison, we also built speaker independent model by pooling male and female data alltogether. Fig. 9 shows the word accuracy (the higher, the better) for an ASR system using all three kinds of models: GI-HMM - gender independent, GD-HMM - gender dependent, and HMM/BN - our model. As can be seen, the HMM/BN model achieves the best result.

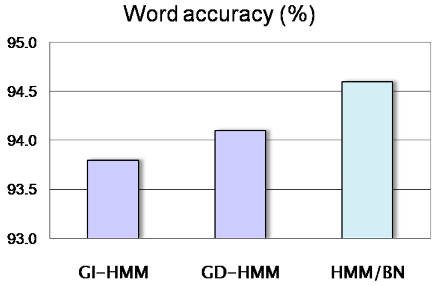

Figure 9: Comparison between gender-independent (GI), gender-dependent (GD) and Hybrid HMM/BN models performance.

5.2. Noisy ASR

In another experiment set to compare the HMM/BN aproach and a DBN based system was conducted for digit recognition task in variaous noisy environments. In this case, the BN representing HMM states has two additional variables corresponding to the noise type N and the signal-to-noise ratio (SNR) S of the environment [11]. The topology of this BN is shown in Fig. 10.

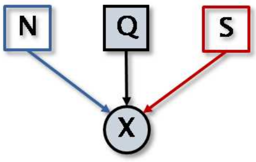

Figure 10: BN topology when noise type and SNR are represented by the additional variables N and S.

The topology of each DBN section also has the N and S variable, but in addition, they are dependent on the same variables from the previous section. As a baseline system, we used atandard HMM based ASR. All three systems were trained using the same data from the Aurora2 databse [11] and the evaluation results in terms of ward accuracy averaged for different noises and with respect to various SNR values are shown in Fig. 11. It shows that for relatively low noise conditions, all systems perform similarly, but as the noise level increases (high SNR), the hybrid HMM/BN system achieves better results.

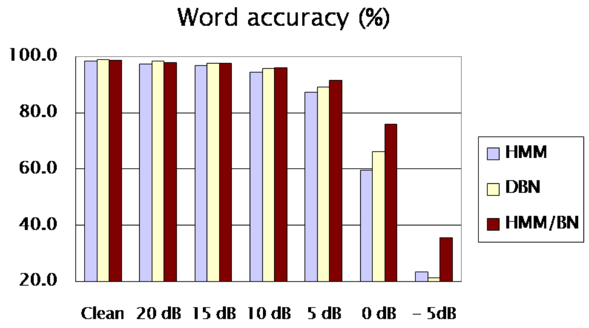

Figure 11: Comparison between standard HMM, DBN and hybrid HMM/BN models performance.

5.3. Bayesian Pentaphone ASR

In this experiment, we evaluated the performance of a pentaphone model, i.e. phoneme model with context dependency of two phonemes to the right and two phonemes to the left using both the standard and Bayesian multi-phone approaches [21, 22]. Normal ptiphone and pentaphone models were trained using all available data, whereas for the Bayesian multiphone, we used the decomposition of Eq.(6, i.e. two triphone and one monophone models. In addition, we implemented aceent (US, British and Australian speakers of English) and gender dependency in all cases by either splitting the data and training separate models or by inserting appropriate BN variables. Evaluation results we obtained are shown in Fig. 12. As in the previous experiments, our method achieved the best results.

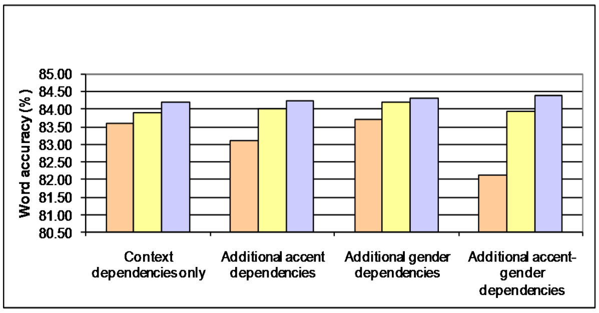

Figure 12: Comparison between triphone, pentaphone and the Bayesian pentaphone modeling approaches.

5.4. Probabilistic Lexicon

For statistical model training, a lot of training data are necessary for robust parameter estimation. However, for the lexicon model obtaining a large amount of data is very expensive since it requires manual labeling of the data by experienced phoneticians. That is why, we used only a small training set that was publicly available at the time of the experiment. It consists of part of Switchboard database with sort duration which has surface form phonetic labels [14]. In order to compare, the effect of the probabilistic lexicon implementation, we trained several systems with or without this lexicon. The baseline system used standard table based lexicon. Next, we used the probabilistic lexicon only for the model training, and in the last system, we used it for both training and testing. As Fig. 13 shows, the best result is achieved with the last system.

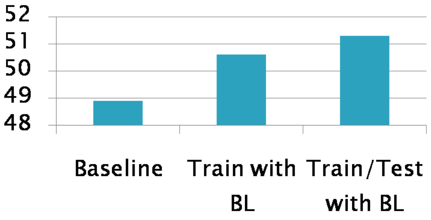

Figure 13: Performance comparison for three systems: Baseline, when Bayesian lexicon is not used, when is used only for training, and when is used for both training and test.

6. Conclusion

We presented an approach to builing structured models for speech recognition using small scale Bayesian networks. In contrast to the Dynamic BN based technique, our method can be applied in both the large and small vocabulary systems and has the advantage of being practicle, computationally tractable and easy to implement. Although not as powerfull as a full scale DBN in expressing multiple time-varying processes, our approach has shown consistent improvements in various tasks and scenarios.

References

[1] W. Holmes and M. Huckvale, "Why have hmms been so successful for automatic speech recognition and how might they be improved,"

Speech, Hearing, and Language, vol. 8, pp. 207–219, 1994.

[2] S. Levinson, L. Rabiner, and 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.

[3] 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.

[4] Roger Moore, "Spoken language processing: Piecing together the puzzle," *Speech Communication*, vol. 49, pp. 418–435, 2007.

[5] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, California, 1988.

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

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

[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 multi-band speech recognition using Bayesian Networks," in *Proc. ASRU*, 2001.

[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, "Modeling HMM State Distributions with Bayesian Networks," in *ICSLP*, 2002, pp. 1013–1016.

[12] K. Markov and S. Nakamura, "Forward-Backwards Training of Hybrid HMM/BN Acoustic Models," in *Interspeech*, 2006, pp. 621–624.

[13] K. Markov and S. Nakamura, "Using Hybrid HMM/BN Acoustic Models: Design and Implementation Issues," *IEICE Trans. Inf. & Syst.*, vol. E86-D, pp. 981–988, 2006.

[14] S. Sakti, K. Markov, and S. Nakamura, "Probabilistic Pronunciation Variation Model Based on Bayesian Network for Conversational Speech Recognition," in *ISUC*, 2008, pp. 405–410.

[15] K. Markov and S. Nakamura, "A Hybrid HMM/BN Acoustic Model for Automatic Speech Recognition," *IEICE Trans. Inf. & Syst.*, vol. E86-D, pp. 438–445, 2003.

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

[17] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. thesis, University of California, Berkeley, 2002.

[18] S. Sakti, K. Markov, and S. Nakamura, "A Hybrid HMM/BN Acoustic Model Utilizing Pentaphone-Context Dependency," *IEICE Trans. Inf. & Syst.*, vol. E89-D, pp. 953–961, 2006.

[19] S. Sakti, K. Markov, and S. Nakamura, "A method to integrate additional knowledge sources into HMM based on junction tree decomposition," in *EUSIPCO*, 2007, pp. 2404–2408.

[20] K. Markov and S. Nakamura, "Acoustic Modeling of Accented English Speech for Large Vocabulary Speech Recognition," in *SRIV*, 2006, pp. 113–118.

[21] S. Sakti, S. Nakamura, and K. Markov, "Improving Acoustic Model Precision by Incorporating a Wide Phonetic Context Based on a Bayesian Framework," *IEICE Trans. Inf. & Syst.*, vol. E89-D, pp. 946–953, 2006.

[22] S. Sakti, K. Markov, and S. Nakamura, "Incorporating Knowledge Sources into a Statistical Acoustic Model for Spoken Language Communication Systems," *IEEE Trans. Computers*, vol. 56, pp. 1199–1211, 2007.