An HMM Acoustic Model Incorporating Various Additional Knowledge Sources

Sakriani Sakti[1,2], *Konstantin Markov*[1,2], *Satoshi Nakamura*[1,2]

[1]National Institute of Information and Communications Technology, Japan
[2]ATR Spoken Language Communication Research Laboratories, Japan
{sakriani.sakti,konstantin.markov,satoshi.nakamura}@atr.jp

Abstract

We introduce a method of incorporating additional knowledge sources into an HMM-based statistical acoustic model. The probabilistic relationship between information sources is first learned through a Bayesian network to easily integrate any additional knowledge sources that might come from any domain and then the global joint probability density function (PDF) of the model is formulated. Where the model becomes too complex and direct BN inference is intractable, we utilize a junction tree algorithm to decompose the global joint PDF into a linked set of local conditional PDFs. This way, a simplified form of the model can be constructed and reliably estimated using a limited amount of training data. Here, we apply this framework to incorporate accents, gender, and wide-phonetic knowledge information at the HMM phonetic model level. The performance of the proposed method was evaluated on an LVCSR task using two different types of accented English speech data. Experimental results revealed that our method improves word accuracy with respect to standard HMM.

Index Terms: acoustic model, knowledge incorporation, Bayesian network, junction tree algorithm, wide-phonetic knowledge.

1. Introduction

There are several approaches that have been developed to build an automatic speech recognition (ASR) system; an intelligent machine that can automatically recognize naturally spoken words uttered by humans. They can generally be classified into two main types: "knowledge-based" and "corpus-based". The idea underlying the former was to use explicit speech knowledge in a rule-based system and produce an acceptable rate of speech recognition. This was based mainly on human ability to interpret spectrograms or other visual representations of speech signals [1, 2]. However, problems lay in the fact that it greatly depended on human experts' ability to interpret spectrograms, and as the number of rules increased, inconsistency among rules also occurred. In contrast, the latter approach was usually based on modeling speech signals using well-defined statistical algorithms that could automatically extract knowledge from the data. This modeling approach has achieved encouraging results, and has outperformed the previous knowledge-based scheme. That is why most current ASR systems usually use statistical data-driven methods based on hidden Markov models (HMMs).

Although such statistical approaches have proved to be efficient choices, ASR systems still often perform much worse than human listeners, especially in the presence of unexpected acoustic variability. Only a limited level of success can be achieved by relying only on statistical models and mostly ignoring the available additional knowledge. Various attempts to integrate more explicit knowledge-based and statistical approaches have been undertaken. The work in [3] proposed that acoustic phonetic knowledge sources be incorporated using neural networks. Others such [4, 5] proposed that articulatory features, sub-band correlation, or speaking styles be incorporated by utilizing dynamic Bayesian Networks (DBNs). However, there are often cases when developing such complex models and achieving optimal performance is not feasible. This is especially the case when resources we have, i.e., available training data and memory space, are insufficient to properly train the model parameters. The best we can do is to choose a simplified form of the model that can be reliably estimated using the training data available.

We introduce a method of incorporating additional knowledge sources into an HMM-based statistical acoustic model in this paper. The approach we adopted here was to utilize the benefits of the Bayesian network framework. Since it allows the probabilistic relationship between information sources to be learned, any additional knowledge sources from any domain can be integrated in a unified way and the global probability function of the model can be formulated. Where the model becomes too complex and direct BN inference is intractable, we utilize a junction tree algorithm to decompose the global joint PDF into a linked set of local conditional PDFs. This way, a simplified form of the model can be constructed and reliably estimated using a limited amount of training data. We applied this framework to incorporate accent, gender, and wide-phonetic knowledge information, and experimentally verified it in an LVCSR task using accented English speech data.

We first describe the general framework to incorporate additional knowledge sources in the next section and give details on junction tree decomposition in Section 3. We then show how to apply this framework to incorporate additional knowledge sources of accent, gender and wide-phonetic information at the HMM phonetic model level in Section 4. Details of the experiments are then presented in Section 5, including the results and discussion. Finally, conclusions are drawn in Section 6.

2. General Framework

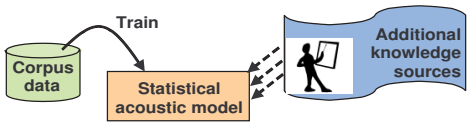

Figure 1: *Incorporating knowledge sources into statistical acoustic model.*

There is a schematic of the incorporation of additional knowledge sources into a statistical acoustic model in Fig. 1, and the procedure basically consists of several steps, as outlined in Fig. 2.

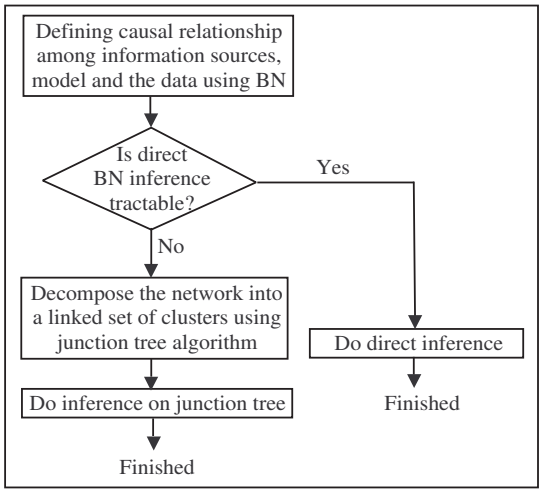

Figure 2: *General procedure of incorporating additional knowledge sources.*

Let us start from a simple case, where given some observation data D, we train model M. Then, suppose that we incorporate two additional knowledge sources K_1 and K_2 into the model. The conditional relationship between D, M, K_1 and K_2 using BN is like that shown in Fig. 3, assuming that both K_1 and K_2 are conditionally independent given M.

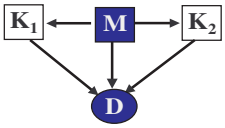

Figure 3: *Conditional relationship between M, D, and additional knowledge sources K_1, K_2.*

The BN joint probability function can be factorized [6] as

$$P(Z_1, Z_2, ..., Z_K) = \prod_{k=1}^{K} P(Z_k|Pa(Z_k)), \quad (1)$$

where $Pa(Z_k)$ denotes the parents of BN variable Z_k, so that we obtain

$$P(D, K_1, K_2, M) = P(D|K_1, K_2, M)P(K_1|M)P(K_2|M)P(M), \quad (2)$$

from Fig. 3. Our primary interest is to calculate the probability, $P(D|K_1, K_2, M)$, which predicts data that can be expected, given current knowledge about the model. The computation of inference can be easy or difficult depending on the complexity of the $P(D|K_1, K_2, M)$. If this PDF allows direct calculation and all variables can be observed, we can simply calculate the output probability as $P(D = d|K_1 = k_{1_j}, K_2 = k_{N_j}, M = m)$. If some variables, such as additional knowledge sources can not be observed or are hidden, we still can carry out direct inference by marginalization over all possible values of these hidden knowledge sources.

However, direct BN can be intractable in some cases, due to too many variables and/or computational complexity. The BN directed graphs need to be decomposed into clusters of variables, on which the relevant computations can be performed, to solve this problem. This can be done with the junction tree algorithm [6], which is briefly described in the next section.

3. Junction Tree Decomposition

Several graphical transformations are applied to obtain a junction tree [6, 7]. We first construct a **moral graph** (an undirected graph with a link between any pair of variables that have a common child). We then **triangulate** the moral graph (add links until all cycles consisting of more than three links have a chord). Figure 4 shows a moral and triangulated version of the BN from Fig. 3.

Then, for each variable Z_k in the triangulated graph with $Pa(Z_k) \neq 0$, we form a subset containing $Pa(Z_k) \bigcup Z_k$ which is called a **cluster/clique**, and build a **junction tree**, starting with clusters/cliques as the nodes, in which each link between two cliques is labeled by using a **separator** of a non-empty intersection between these cliques. However, we can only obtain one cluster/clique from this triangulated graph with the full set of variables $\{D, M, K_1, \text{and } K_2\}$ and can not decompose it any further.

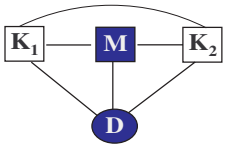

Figure 4: *Moral and triangulated graph of Fig. 3*

Fortunately, since K_1, and K_2 are assumed to be independent, we could solve this problem by reversing some arrows to obtain an equivalent graph, as in Fig. 5(a). Figure 5(b) shows the moral and triangulated version of this graph. We can then identify the clusters/cliques and obtain the junction tree in Fig. 5(c), where the cluster sets are represented by oval nodes and the separator sets are represented by square nodes.

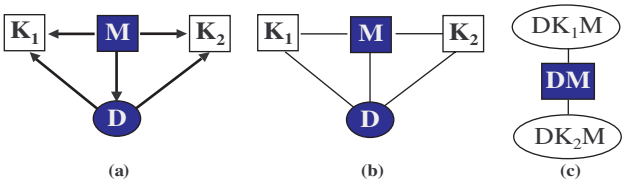

Figure 5: *(a) Equivalent BN topology of Fig. 3 (b) Moral and triangulated graph of Fig. 5a (c) Junction tree of Fig. 5b.*

The joint probability distribution is then defined as the product of all cluster potentials, divided by the product of the separator potentials [7] as

$$P(Z_1, Z_2, ..., Z_K) = \frac{\prod_i \phi_{C_i}}{\prod_j \phi_{S_i}}, \quad (3)$$

where ϕ_{C_i} is the cluster potential (the probability over cluster C_i), and ϕ_{S_i} is the separator potential (the probability over separator S_i), respectively. Thus, according to Fig. 5(c), the joint probability function, $P(D, K_1, K_2, M)$, becomes

$$P(D, K_1, K_2, M) = \frac{P(D, K_1, M)P(D, K_2, M)}{P(D, M)}, \quad (4)$$

Then, using Eqs. (2) and (4), we finally obtain

$$P(D|K_1, K_2, M) = \frac{P(D|K_1, M)P(D|K_2, M)}{P(D|M)}. \quad (5)$$

This indicates a new way of representing probability function $P(D|K_1, K_2, M)$, as a composition of several local probability functions $P(D|K_1, M), P(D|K_2, M)$, corresponding to the probability of observation data, D, given the specific additional knowledge K_1 and K_2. The term, $P(D|M)$, serves as a normalization constant here.

Now, it should be much easier to define, estimate and calculate several simple $P(D|K_i, M)$ than a single but complex $P(D|K_1, ..., K_N, M)$.

4. Incorporating Accents, Gender, and Wide-Phonetic Context Information at HMM Phonetic Model Level

We apply the theoretical framework described in Section 2 to the problem of incorporating additional knowledge sources into HMM. The model, M, is currently our HMM phonetic model λ, and D is $X_s = X_t, ..., X_{t+s}$, an observation data segment of length s.

We first incorporate additional wide-phonetic context knowledge, where K_1 represents preceding contexts C_L and K_2 represents succeeding contexts C_R. The topological structure is similar to the one in Fig. 3, and the probability function of HMM phonetic units is now represented by the BN joint probability function, similar to Eq. (2)

$$\begin{aligned} &P(X_s, C_L, C_R, \lambda) \\ &= P(X_s|C_L, C_R, \lambda)P(C_L|\lambda)P(C_R|\lambda)P(\lambda). \end{aligned} \quad (6)$$

Our primary interest is now to calculate $P(X_s|C_L, C_R, \lambda)$, given input segment X_s. However, it is difficult to obtain a simple functional form for this conditional PDF, because it involves HMM model λ, segment X_s of variable duration, and wide-phonetic context knowledge. We thus need to decompose $P(X_s|C_L, C_R, \lambda)$ with the junction tree algorithm described in Section 3 in this case. It can be decomposed as

$$P(X_s|C_L, C_R, \lambda) = \frac{P(X_s|C_L, \lambda)P(X_s|C_R, \lambda)}{P(X_s|\lambda)}, \quad (7)$$

according to Eq. (5). If we assume that λ is monophone unit model $/a/$, and C_L and C_R are preceding and following context unit models $/a^-/$ and $/a^+/$, we can thus define

$$P(X_s|C_L, C_R, \lambda) = P(X_s|[a^-, a, a^+]), \quad (8)$$

and Eq. (8) becomes

$$P(X_s|[a^-, a, a^+]) = \frac{P(X_s|[a^-, a])P(X_s|[a, a^+])}{P(X_s|[a])}. \quad (9)$$

This equation has the same factorization as the one proposed in [8], where a triphone model is constructed from monophone and biphone models and is known as a Bayesian triphone.

One simple way of representing the composition of a wider phonetic context such as pentaphone $/a^{--}, a^-, a, a^+, a^{++}/$ is by setting λ to represent a monophone, $/a/$, and the second preceding and succeeding contexts, C_L and C_R, to represent $/a^{--}, a^-/$ and $/a^+, a^{++}/$, respectively. We then obtain

$$\begin{aligned} &P(X_s|[a^{--}, a^-, a, a^+, a^{++}]) \\ &= \frac{P(X_s|[a^{--}, a^-, a])P(X_s|[a, a^+, a^{++}])}{P(X_s|[a])}, \end{aligned} \quad (10)$$

which indicates that pentaphone $P(X_s|[a^{--}, a^-, a, a^+, a^{++}])$ can be composed from a left/preceding-triphone-context unit (L3), a right/following-triphone-context unit (R3), and a monophone unit (C1). We call this composition C1L3R3 in this paper.

We next extend C1L3R3 with other additional knowledge variables, such as gender or accent information. We can extend it with gender information only, accent information only, or with both accent and gender information. The BN topology and its corresponding junction tree for the case with additional accent and gender information becomes that shown in Fig. 6, and the conditional probability function is obtained as

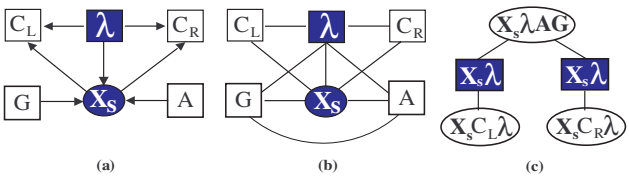

Figure 6: *(a) BN topology. (b) Moral and triangulated graph. (c) Corresponding junction tree.*

$$\begin{aligned} &P(X_s|C_L, C_R, \lambda, A, G) \\ &= P(X_s|\lambda, A, G)\frac{P(X_s|C_L, \lambda)}{P(X_s|\lambda)}\frac{P(X_s|C_R, \lambda)}{P(X_s|\lambda)} \\ &= \frac{P(X_s|C_L, \lambda, A, G)P(X_s|C_R\lambda, A, G)}{P(X_s|\lambda, A, G)}. \end{aligned} \quad (11)$$

Thus, following the same setting as before, the pentaphone likelihood becomes

$$\begin{aligned} &P(X_s|[a^{--}, a^-, a, a^+, a^{++}], A, G) \\ &= \frac{P(X_s|[a^{--}, a^-, a], A, G)P(X_s|[a, a^+, a^{++}], A, G)}{P(X_s|a, A, G)}, \end{aligned} \quad (12)$$

which indicates that $P(X_s|[a^{--}, a^-, a, a^+, a^{++}], A, G)$ can be simplified by factorization into $P(X_s|a, A, G)$, $P(X_s|[a^{--}, a^-, a], A, G)$, and $P(X_s|[a, a^+, a^{++}], A, G)$.

5. Experiments

We used the ATR accented English speech corpus of travel domain expressions, which consists of American (US) and Australian (AUS) accents. The training data consisted of 90% of the total data or about 40k utterances (80 speakers: 40 male and 40 female). The test data included 200 randomly selected utterances from the remaining 10% of all accent data. A sampling frequency of 16 kHz, a frame length of a 20-ms Hamming window, a frame shift of 10 ms, and 25 dimensional vectors (12-order MFCC, Δ MFCC and Δ log power) were used as feature extraction parameters. Three states were used as the initial HMM for each phoneme. A shared state HMnet topology was then obtained using a successive state splitting (SSS) training algorithm based on the minimum description length (MDL) optimization criterion [9]. A context-dependent triphone HMM model (having a total of 2,126 states) and a pentaphone HMM model (having a total of 2,202 states) were used as the baseline. Incorporation of additional knowledge such as gender and accents was also possible for the baseline models by training gender and/or accent dependent acoustic models. Only an embedded training procedure was carried out with the specific accent or gender training data so that all models had the same topological structure.

All component of the Bayesian pentaphone model, C1L3R3, were trained separately using the same amount of training data and the same SSS training algorithm. There was a total of 3,403 states (sum of C1: 132 st., L3: 1,645 st., R3: 1,626 st.). An embedded training procedure was then undertaken for the extended C1L3R3 on specific accent or gender training data.

The pentaphone HMM baseline and the proposed pentaphone C1L3R3 models were applied to rescoring the N-best list

generated from a standard and unmodified triphone ASR system to simplify decoding, as we did previously [10]. The rescoring was done using a 10-best list, and a 0.3 weight parameter α for deleted interpolation was used, as in our previous study [10].

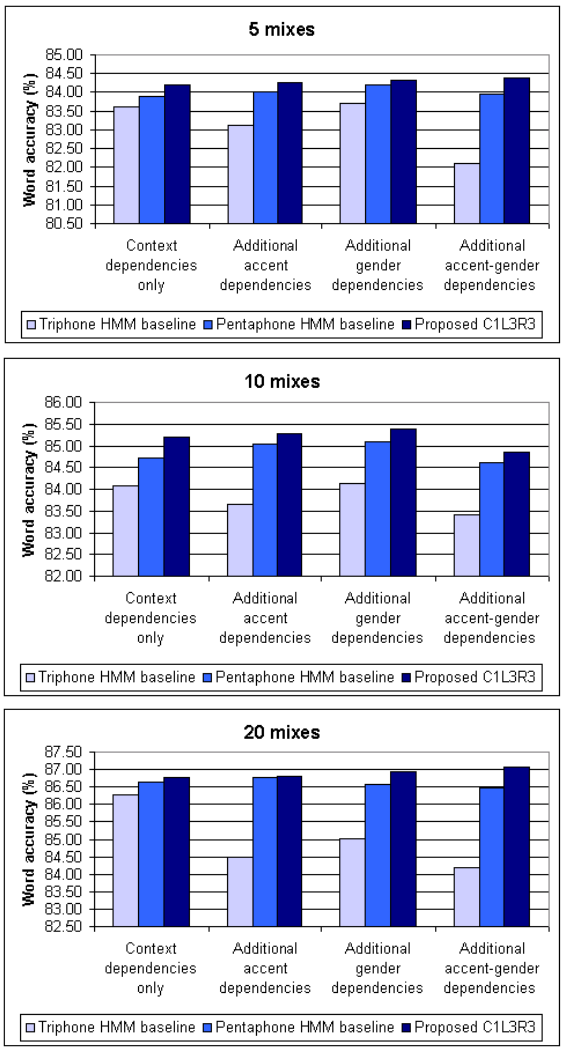

Figure 7: Comparing recognition accuracy rates of different systems triphone HMM baseline, pentaphone HMM baseline, and the proposed pentaphone models having the same 5,10,20 mixture components per state.

How well each of the models performed having the 5, 10, 20 mixture components per state is shown in Fig. 7. The triphone baseline without any additional knowledge achieved 83.60% word accuracy for the 5 mixture components per state. However, it decreased to 82.11% word accuracy for the accent-gender-dependent models. This might be due to the size of the training data, which is much smaller than that of the other baseline models. Performance could be improved by rescoring with a more precise pentaphone model. Of the pentaphone models, the performance of the model we propose was always better than that of conventional pentaphone HMM. This might be because given the amount of training data, the training of the conventional pentaphone model using the MDL-SSS algorithm resulted in a model having a total of 2,202 states, which is not that different from the total number of states in the triphone HMM. As many different pentaphone contexts seemed to share the same Gaussian components, the context resolution was reduced. Thus, approximating a pentaphone model using the composition of several less context-dependent models could help to reduce the loss of context resolution and improve performance. Performance did not decrease when gender and accent were incorporated, as in the case of the triphone baseline, which is probably due to the use of deleted interpolation. The best performance was obtained by the model that incorporated additional knowledge of accent A, gender G, second preceding context C_L, and succeeding context C_R. Overall, the results revealed that, through different mixture components per state, the proposed pentaphone models consistently outperformed the standard HMM baseline.

6. Conclusion

We introduced a general framework to incorporate additional knowledge sources into statistical HMM acoustic models. We also demonstrated the implementation of this new framework by integrating accents, gender, and wide-phonetic context information. The framework is based on a junction tree algorithm and allows us to construct models with wider contexts from several others with narrower contexts. As this leads to a reduction in the number of context units to be estimated, the loss of context resolution can be considerably reduced. We applied these composition models at the post-processing stage with N-best rescoring. Performance was evaluated on an LVCSR task using two different types of accented English speech data. The experimental results revealed that our method improves word accuracy with respect to standard HMM with or without additional knowledge sources. The best performance was obtained by the model that incorporated additional knowledge of accent A, gender G, second preceding context C_L and succeeding context C_R.

7. References

[1] V.W. Zue and R.A. Cole, "Experiments on spectrogram reading," in *Proc. ICASSP*, Washingtton D.C., USA, 1979, pp. 116–119.

[2] D.H. Klatt, "Review of the ARPA speech understanding project," *Acoustical Society of America*, vol. 62, pp. 1345–1366, 1977.

[3] J. Li, Y. Tsao, and C.-H. Lee, "A study on knowledge source integration for candidate rescoring in automatic speech recognition," in *Proc. ICASSP*, Philadelphia, USA, 2005, pp. 837–840.

[4] G. Zweig and S. Russell, "Probabilistic modeling with Bayesian networks for automatic speech recognition," in *Proc. ICSLP*, Sydney, Australia, 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*, Beijing, China, 2000, pp. 329–332.

[6] F. Jensen, *An Introduction to Bayesian Network*, UCL Press, 1998.

[7] C. Huang and A. Darwiche, "Inference in belief networks: A procedural guide.," *International Journal of Approximate Reasoning*, vol. 11, pp. 1–158, 1994.

[8] Ji Ming, P. O Boyle, M. Owens, and F. Jack Smith, "A Bayesian approach for building triphone models for continuous speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 6, pp. 678–684, November 1999.

[9] 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, pp. 2121–2129, 2004.

[10] 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, no. 3, pp. 946–953, 2006.