Modeling Auxiliary Information in Bayesian Network Based ASR

Todd A. Stephenson[*], *M. Mathew, Hervé Bourlard*[*]

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P. O. Box 592, CH-1920 Martigny, Switzerland

Abstract

Automatic speech recognition bases its models on the acoustic features derived from the speech signal. Some have investigated replacing or supplementing these features with information that can not be precisely measured (articulator positions, pitch, gender, etc.) automatically. Consequently, automatic estimations of the desired information would be generated. This data can degrade performance due to its imprecisions. In this paper, we describe a system that treats pitch as an auxiliary information within the framework of Bayesian networks, resulting in improved performance.

1. Introduction

Automatic speech recognition (ASR) uses statistical models that give the likelihood of acoustic observations $\mathbf{X}$ given the hidden states $\mathbf{Q}$:

$$P(\mathbf{X}|\mathbf{Q}). \qquad (1)$$

If relevant auxiliary, non-acoustic information, $\mathbf{A}$, is available, it can be used to reduce the variability of the models and improve their robustness, using the enhanced likelihood:

$$P(\mathbf{X},\mathbf{A}|\mathbf{Q}). \qquad (2)$$

Used as complementary observations, $\mathbf{A}$ could encode different auxiliary information such as the contour of the lips [1], the position of the articulators [2], or gender information [3]. Depending on the relevance of $\mathbf{A}$ and the reliability of its measurement or estimation (during training and/or recognition), its direct use in the estimation of (2) (during training and/or recognition) may improve or hurt the performance of the resulting system. For example, in the case of gender modeling [3], $\mathbf{A}$ is simply a two-value variable (male or female) which is directly used during training (where we thus assume that we know for sure the sex of the speaker). However, during testing, it is better not to try to estimate the speaker's gender, using it directly in (2), but instead to infer it automatically as a by-product of the recognition process; this is typically done by picking up the conditional model (conditional on the auxiliary information) yeilding the highest likelihood. While this approach, using a simple variant of HMM (i.e., simply duplicating hidden Markov models (HMM's) for each possible value of the auxiliary variable), has been shown to be quite effective in a number of cases, it quickly becomes untractable if the auxiliary variable can take many different values (since this would require a large number of conditional HMM models) or is a continuous value.

In this paper, we thus show how to deal with a multi-valued auxiliary variable by using Bayesian network (BN) as a generalization of the HMM formalism. In our case, this variable will be a discretized value of the pitch estimate (though BN's can also treat this as a continuous variable). Furthermore, we show that in this particular case, it seems more beneficial to use the estimated pitch value directly during training, while during recognition it is better to infer it (from the acoustic data and the trained BN parameters).

In Section 2, we recall some basic background for how to include auxiliary information in ASR. We follow that by explaining in Section 3 how our models were formed and trained. We then discuss the recognition results in Section 4

2. Auxiliary Data

For acoustic observations $\mathbf{X}$ auxiliary information $\mathbf{A}$, the task is to find the model M that maximizes the posterior probability:

$$\begin{aligned} P(M|\mathbf{X},\mathbf{A}) &= \frac{P(\mathbf{X},\mathbf{A}|M)P(M)}{P(\mathbf{X},\mathbf{A})} && (3) \\ &\propto P(\mathbf{X},\mathbf{A}|M), && (4) \end{aligned}$$

with (4) assuming there are equal prior probabilities $P(M)$ and recognizing that $P(\mathbf{X},\mathbf{A})$ remains constant for all $\mathbf{M}$. The likelihood $P(\mathbf{X},\mathbf{A}|M)$ in (4) can be further reduced, with a first-order Markov assumption for $\mathbf{A}$ and a time-independence assumption for $\mathbf{X}$:

$$P(\mathbf{X},\mathbf{A}|M) \cong \prod_{t=0}^{N} P(X_t|A_t,M)P(A_t|A_{t-1},M) \quad (5)$$

[*] Also affiliated with The Swiss Federal Institute of Technology of Lausanne, CH-1015 Lausanne, Switzerland

If the auxiliary data $\mathbf{A}$ happened to be independent of the model M, then (5) reduces to:

$$P(\mathbf{X}, \mathbf{A}|M) \cong \prod_{t=0}^{N} P(X_t|A_t, M)P(A_t|A_{t-1}) \quad (6)$$

For the cases where $\mathbf{X}$ are our only observations (the $\mathbf{A}$ are hidden), the likelihood calculation becomes a sum over all K possible assignments to $\mathbf{A}$:

$$P(\mathbf{X}|M) = \sum_{k=1}^{K} P(\mathbf{X}, \mathbf{A_k}|M) \quad (7)$$

3. Model Setup

3.1. Dynamic Bayesian Networks

3.1.1. Definition

We have set up our models within the framework of dynamic Bayesian networks (DBN's) [4, 5] as opposed to the conventional HMM's. DBN's are in the same family of models as HMM's [6], but using them specifically allows more ease in experimenting with different topologies and with hidden vs. observable data. [7, 8] provide the foundation for how we do ASR with DBN's. A BN, of which a DBN is a specific type, has the following components:

1. A set of variables $\mathbf{Z}$ to model
2. A directed acyclic graph (DAG), $G = <\mathbf{V}, \mathbf{E}>$, with vertices $\mathbf{V}$ and edges $\mathbf{E}$ and with a one-to-one mapping between $\mathbf{Z}$ and $\mathbf{V}$.
3. A local, conditional probability distribution for each $Z \in \mathbf{Z}$ based only upon the parent vertices:

$$P(Z|\text{parents}(Z)) \quad (8)$$

The joint probability distribution of a model is then defined as the product of the local, conditional probability distributions:

$$P(\mathbf{Z}) = \prod_{\mathbf{Z}} P(Z|\text{parents}(Z)) \quad (9)$$

As with HMM's, probabilistic inference, given new observations, is done in two-passes over the vertices in the DBN. Given the observations $\mathbf{e}^-$ below a variable Z in the DAG, and the observations $\mathbf{e}^+$ "above" the variable Z in the DAG, the joint distribution $P(\mathbf{Z}, \mathbf{e})$ can be factored [9]:

$$\begin{aligned} & P(Z, \mathbf{e}) && (10)\\ &= P(Z, \mathbf{e}^-, \mathbf{e}^+) && (11)\\ &= P(Z, \mathbf{e}^+)P(\mathbf{e}^-|Z, \mathbf{e}^+) && (12)\\ &= P(Z, \mathbf{e}^+)P(\mathbf{e}^-|Z) && (13) \end{aligned}$$

The reduction from (12) to (13) is due to (8). The factors in (13) are also known as $\pi(Z)$ and $\lambda(Z)$, respectively and are analogous to the α and β parameters used in HMM inference:

$$\begin{aligned} \pi(Z) &= P(Z, \mathbf{e}^+) && (14)\\ \lambda(Z) &= P(\mathbf{e}^-|Z) && (15) \end{aligned}$$

λ parameters are computed in the first-pass and the π parameters in the second. Probabilistic inference in the DBN is the same regardless of which variables are or are not observed. In the case of the unobserved variables, their posterior distributions are determined based on the observed variables' values. This will enable us to leave the auxiliary variable as hidden, with the DBN inferring its distribution over all of its possible values instead of the estimator finding a hard, uncertain value.

3.1.2. Use in ASR

Figure 1 gives the baseline system which does not account for any auxiliary variable A. It models the three streams of the acoustic emissions $\mathbf{X}$ as well as as the following component variables of M:

- Index - The hidden state number.
- Exit - The boolean value of whether to exit from the current phonetic state.
- Phone - The phonetic state corresponding to Index;

'Exit' is a stochastic variable while 'Index' and 'Phone' are deterministic (i.e., have probabilities of 1's and 0's), as defined by the given model M.

Figures 2 & 3 present two stages for incorporating the auxiliary information $\mathbf{A}$ into the DBN, based on (5) and (6), respectively. They only show a single time-step of the DBN, based on Figure 1.

3.2. Data

3.2.1. MFCC's

We have selected PhoneBook [10] as the corpus to use for training and testing our models. The bulk of this corpus contains isolated word utterances collected from a large number of speakers with a large variety of words. We have chosen a similar partition to that in [11]: we used their "small" training set for training our models and their cross-validation set for testing our models.

Sampled in 8 kHz, the speech signal was parmaterized using mel-frequency cepstral coefficients (MFCC's) with a Hamming window of 25 msec in width, shifting 8.3 msec per frame. Ten MFCC's as well as C_0, the energy coefficient, were retained for the models. We then created three codebooks for the acoustic data, each of 256 values:

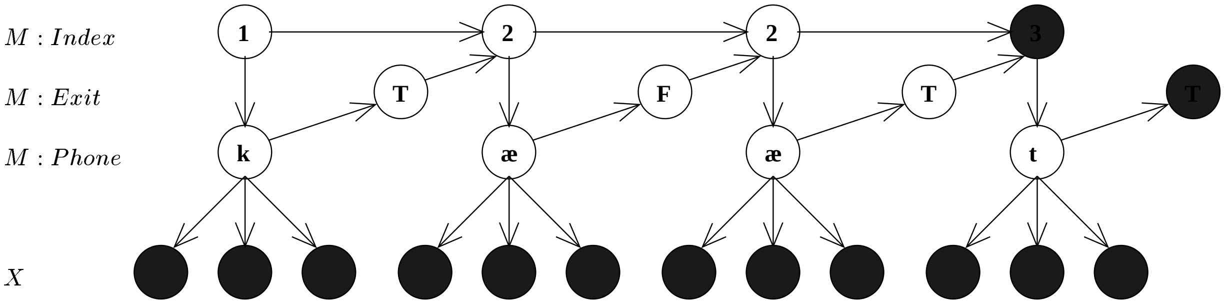

Figure 1: Baseline DBN for modeling $P(\mathbf{X}|M)$. Equivalent to a standard HMM.

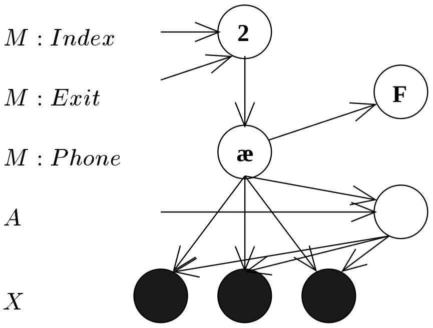

Figure 2: DBN for modeling $P(\mathbf{X},\mathbf{A}|M)$. A_t is conditioned upon $Phone_t$ and A_{t-1}. X_t is conditioned on $Phone_t$ as well as A_t.

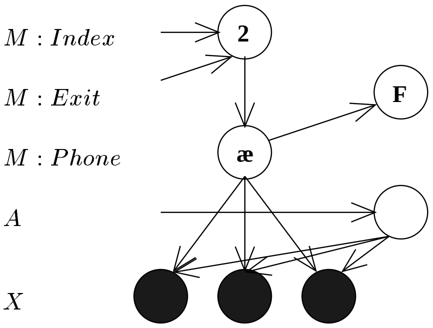

Figure 3: DBN for modeling $P(\mathbf{X},\mathbf{A}|M)$, where $\mathbf{A}$ is independent of M. A_t is conditioned only on A_{t-1}. X_t is conditioned on $Phone_t$ as well as A_t.

- Ten MFCC's.
- Approximated derivative of the ten MFCC's.
- C_0 combined with its approximate derivative.

3.2.2. Auxiliary Parameters

In examining the effectiveness of our claim of the effectiveness of hiding estimations in recognition, it is worthwhile to work with features that are of interest to ASR yet are hard to estimate. Pitch meets these qualities well; pitch is a distinguishing feature in spoken language, and Hess described pitch determination as "among the most difficult problems in speech analysis" [12]. Speech is produced by the excitation of a time-varying vocal tract by a time-varying source (vibration of the vocal cords). The acoustic correlate of the vibration of the vocal cords is the fundamental frequency (F_0) or pitch frequency [12]. In this paper, we define pitch to be F_0. Pitch is a speaker-specific feature and is used for speaker recognition [13]. The voice source parameters include the type of phonation (voiced or unvoiced) and the measure of periodicity (F_0) of the speech signal, if it is voiced. Thus estimation of pitch implicitly provides information about voicing. The presence vs. absence of voicing plays a vital role in phonetics. That is, a language can have two phonemes whose characteristics differ only regarding whether there is voicing or not. An example of this is the phonemes /z/ (voiced) and /s/ (unvoiced).

Therefore, in addition to the acoustic features above, we generated some auxiliary parameters to supplement them. Specifically, we generated some pitch estimates using the *S*imple *I*nverse *F*ilter *T*racking (SIFT) algorithm [14], which is based on an inverse filter formulation. This method retains the advantages of the autocorrelation and cepstral analysis techniques. The speech signal is prefiltered by a low pass filter with a cut-off frequency of 800 Hz, and the output of the filter is sampled at 2 kHz before computing the inverse filter coefficients using the Durbin algorithm. Codebooks of size one, three, and seven were generated for the data where the pitch was non-zero. With an additional entry in each reserved for the case where the pitch is zero, this resulted in codebooks of two, four, and eight. The codebooks were each reserved for a different set of experiments.

4. Recognition Results

All systems used three hidden states for each sub-word model; there were 41 monophone sub-word models plus models for beginning silence and ending silence. A baseline system, with no auxiliary information was trained using the baseline model and data described above. It is theoretically equivalent to a discrete HMM. The two sets of

Baseline DBN			
	7.8%		33k
	Estimated Auxiliary	Ignored Auxiliary	
Auxiliary DBN (Figure 2)			
2 Prototypes	8.5%	7.6%	66k
4 Prototypes	7.9%	7.1%	133k
8 Prototypes	8.6%	7.2%	270k
Auxiliary DBN (Model-independent, Figure 3)			
2 Prototypes	8.5%	7.7%	66k
4 Prototypes	8.0%	7.2%	133k
8 Prototypes	8.9%	N/A	270k

Table 1: Word Error Rates. Both results within the same given line were from the same system, which was trained on estimated auxiliary data. The final column is the number of parameters.

DBN's (see Section 3.1) with auxiliary information were each trained using the different sized codebooks for the auxiliary information, as explained in Section 3.2.2.

Using these trained DBN's, we tested their performance with the estimated auxiliary information provided and also with it left hidden. Results are given in Table 1. In all cases, the Auxiliary DBN performed significantly better when the auxiliary information was left hidden than when the auxiliary information was observed. Furthermore, the best performing Auxiliary DBN is significantly better than the Baseline DBN: 7.1% verses 7.8%.

5. Conclusion

In this paper, we have compared different approaches towards including pitch as auxiliary information in state-of-the-art DBN based speech recognition. During training, we explicitely estimated (using the SIFT algorithm) the instantaneous pitch values. They were used to clamp the DBN variable coding this auxiliary information and to estimate the DBN parameters maximizing the enhanced likelihood. Although our pitch estimator is noisy, its use was shown to provide enough information during training to enhance the resulting acoustic model. However, given the fact that the pitch estimator is not perfect, it was also shown that during recognition it was better not to explicitely estimate the pitch value. Rather it was better to use the DBN parameters to infer the value of the auxiliary information that maximizes the joint likelihood (i.e., integrating over all possible pitch values), resulting in significant performance gain.

6. Acknowledgements

Todd A. Stephenson and M. Mathew are supported by the Swiss National Science Foundation under grants # 2100-053960.98/1 and 2100-057245.99/1, respectively.

7. References

[1] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 141–151, September 2000.

[2] J. Frankel, K. Richmond, S. King, and P. Taylor, "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces," in *ICSLP*, October 2000, pp. IV:254–257.

[3] C. Neti and S. Roukos, "Phone-context specific gender-dependent acoustic-models for continuous speech recognition," in *ASRU*, December 1997.

[4] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc., 1988.

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

[6] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural Computation*, vol. 11, no. 2, 1999.

[7] G. Zweig and S. Russell, "Speech recognition with dynamic Bayesian networks," in *AAAI*, July 1998, pp. 173–180.

[8] G. G. Zweig, *Speech Recognition with Dynamic Bayesian Networks*, Ph.D. thesis, University of California, Berkeley, 1998.

[9] M. A. Peot and R. D. Shachter, "Fusion and propagation with multiple observations in belief networks," *Artificial Intelligence*, vol. 48, pp. 299–318, 1991.

[10] J. F. Pitrelli, C. Fong, S. H. Wong, J. R. Spitz, and H. C. Leung, "PhoneBook: A phonetically-rich isolated-word telephone-speech database," in *ICASSP*, May 1995.

[11] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite, "Hybrid HMM/ANN systems for training independent tasks: Experiments on phonebook and related improvements," in *ICASSP*, 1997, pp. 1767–1770.

[12] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer-Verlag, 1983.

[13] B. S. Atal, "Automatic speaker recognition based on pitch contours," *JASA*, vol. 52, no. 6, pp. 1687–1697, 1972.

[14] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio and Electroacoustics*, vol. 20, pp. 367–377, 1972.