AUXILIARY VARIABLES IN CONDITIONAL GAUSSIAN MIXTURES FOR AUTOMATIC SPEECH RECOGNITION

Todd A. Stephenson *Mathew Magimai-Doss* *Hervé Bourlard*

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), CH-1920 Martigny, Switzerland
The Swiss Federal Institute of Technology Lausanne (EPFL), CH-1015 Lausanne, Switzerland

ABSTRACT

In previous work, we presented a case study using an estimated pitch value as the conditioning variable in conditional Gaussians that showed the utility of hiding the pitch values in certain situations or in modeling it independently of the hidden state in others. Since only single conditional Gaussians were used in that work, we extend that work here to using conditional Gaussian mixtures in the emission distributions to make this work more comparable to state-of-the-art automatic speech recognition. We also introduce a rate-of-speech (ROS) variable within the conditional Gaussian mixtures. We find that, under the current methods, using observed pitch or ROS in the recognition phase does not provide improvement. However, systems trained on pitch or ROS may provide improvement in the recognition phase over the baseline when the pitch or ROS is marginalized out.

1. INTRODUCTION

Hidden Markov models (HMMs) calculate at each time n the likelihood of the acoustic observation x_n being produced, given that the hidden state variable q_n has the discrete value of k (with K possible discrete values):

$$p(x_n|q_n=k). \tag{1}$$

This is typically computed using an ANN or a Gaussian mixture distribution, with mean $\mu_{k,m}$, covariance $\Sigma_{k,m}$, and mixtures $m=1,\dots,M$:

$$p(x_n|q_n=k) \sim \sum_{m=1}^{M} P(m|q_n=k)\cdot\mathcal{N}(\mu_{k,m},\Sigma_{k,m}). \tag{2}$$

There may be information not directly available in the acoustic observation x_n that may be of use in enhancing the models. Such auxiliary information a_n, which can be continuous or discrete, may be derived from the acoustic signal or may be obtained from a secondary source. q_n and a_n can then jointly condition the emission likelihoods, replacing (1) with:

$$p(x_n|q_n=k, a_n=z). \tag{3}$$

For the case of discrete a_n, $a_n = 1,\dots,L$, Gaussian mixture models are also used to estimate the emission likelihoods:

$$p(x_n|q_n=k,a_n=l) \sim \sum_{m=1}^{M} P(m|q_n=k)\cdot\mathcal{N}(\mu_{k,l,m},\Sigma_{k,l,m}), \tag{4}$$

resulting in L times as many Gaussians over that of (2). For the case of continuous a_n, it is more difficult to model the emission distributions of (3). We have chosen the framework of *conditional* Gaussians, as also done in [1], though this is not the definitive way. In conditional Gaussians the means of the emission probabilities for the Gaussian distributions (2) can then be shifted using the regression weights B_k upon the value of a_n:

$$p(x_n|q_n=k,a_n=z) \sim \sum_{m=1}^{M} P(m|q_n=k)\cdot\mathcal{N}(u_{k,m},\Sigma_{k,m}), \quad u_{k,m}=\mu_{k,m}+B_{k,m}^T z \tag{5}$$

So, instead of having L Gaussians for a given mixture of a state, one conditional Gaussian is defined whose mean changes dynamically according to a_n. The variance within the conditional Gaussian, however, does not itself depend upon a_n; doing this is itself a topic of future research.

We proceed as follows: we begin in Section 2 by specifying, in the framework of (conditional) Gaussian mixtures, how auxiliary information can be incorporated into the acoustic modeling. This is then transfered to the dynamic Bayesian network (DBN) framework in Section 3. These DBNs are then used in experimental testing in Section 4, followed by discussion in Section 5.

2. INTRODUCING AUXILIARY INFORMATION WITH MIXTURES

ASR with auxiliary information involves modeling $p(X,A,Q)$, the evolution of the observed space $X_1^N=\{x_1,x_2,\dots,x_N\}$ and the observed or hidden auxiliary space $A_1^N=\{a_1,a_2,\dots,a_N\}$ and the hidden state space $Q_1^N=\{q_1,q_2,\dots,q_N\}$ for time $n=1,\dots,N$ as[1]

$$p(X_1^N,A_1^N,Q_1^N) \approx \prod_{n=1}^{N} p(x_n,a_n|q_n)\cdot P(q_n|q_{n-1}) \tag{6}$$

$$\approx \prod_{n=1}^{N}\sum_{m=1}^{M} p(x_n,a_n,m|q_n)\cdot P(q_n|q_{n-1}) \tag{7}$$

$$\approx \prod_{n=1}^{N}\sum_{m=1}^{M} p(x_n|a_n,m,q_n)\cdot p(a_n|q_n)\cdot P(m|q_n)\cdot P(q_n|q_{n-1}), \tag{8}$$

where we assume time-independence of x_n and a_n and a first-order Markov assumption (that is, $q_n \perp\!\!\!\perp Q_1^{n-2} \mid q_{n-1}$).[2] Furthermore, (8) assumes that a_n is not modeled by mixtures (that is, it has a single Gaussian).

[1] Assume throughout this paper that $P(q_1|q_0)=P(q_1)$.
[2] read, "q_n is conditionally independent of Q_n^{n-2} given q_{n-1}."

We are then interested in whether different assumptions related to a_n can be incorporated into (8). One is whether a_n even needs to be treated as a conditioning variable to x_n–that is, having the assumption $x_n \perp\!\!\!\perp a_n \mid q_n$, as in (9). A separate assumption involves whether the modeling of the auxiliary variable a_n can be done independently of the states q_n (that is, $a_n \perp\!\!\!\perp q_n$)[3] as in (10).

$$\prod_{n=1}^{N}\sum_{m=1}^{M} p(x_n|m,q_n)\cdot p(a_n|q_n)\cdot P(m|q_n)\cdot P(q_n|q_{n-1}) \quad (9)$$

$$\prod_{n=1}^{N}\sum_{m=1}^{M} p(x_n|a_n,m,q_n)\cdot p(a_n)\cdot P(m|q_n)\cdot P(q_n|q_{n-1}) \quad (10)$$

Standard HMM ASR estimates $p(X,Q)$ using (8) with references to A_1^N marginalized out:

$$p(X_1^N,Q_1^N) \approx \prod_{n=1}^{N}\sum_{m=1}^{M} p(x_n|m,q_n)\cdot P(m|q_n)\cdot P(q_n|q_{n-1}), \quad (11)$$

In summary, (11),(8),(9),(10) are used in our experimental section to test, respectively, a baseline system, an auxiliary baseline system, an auxiliary system with $x_n \perp\!\!\!\perp a_n \mid q_n$, and an auxiliary system with $a_n \perp\!\!\!\perp q_n$. The systems using (11) are equivalent to standard multi-Gaussian HMM-based ASR. The systems using (9) are equivalent to standard multi-Gaussian HMM-based ASR with a_n appended to the standard feature vector (though a_n itself is modeled by a single Gaussian).

3. AUXILIARY INFORMATION WITH DYNAMIC BAYESIAN NETWORKS

Dynamic Bayesian networks (DBNs), which are an extension of Bayesian networks (BNs)[4] [2], have been proposed as an alternative to HMMs that allows more flexibility in modeling the topology of the probability distributions within ASR [3]. For example, consider the four distributions that we proposed in Section 2: (11),(8),(9),(10). While they can be modeled with an HMM framework, a different version of the HMM programs used may need to be developed to handle each assumption. The DBN framework, however, is flexible enough to handle a wide range of assumptions, while using the same programs.

A BN, from which a DBN is built, is defined by three sets:

1. variables V (discrete or continuous)
2. directed acyclic graph (DAG), consisting of a node for each variable as well as directed arcs between nodes. These arcs indicate probabilistic dependencies between the underlying variables.
3. local probability distributions for each variable $v \in V$, whose topology is $p(v|\text{parents}(v))$. $\text{parents}(v)$ are the variables whose nodes have an arc going to v's node.

For continuous parent variables instantiated as $C = C'$ and for discrete parent variables instantiated as $D = D'$, the local probability distributions are defined as:

- v discrete:
  - $P(v|D=D')$: a table of probabilities.
  - $P(v|C=C')$ or $P(v|C=C',D=D')$: undefined in this framework.
- v continuous:
  - $p(v)$: Gaussian–$\mathcal{N}(\mu_v,\sigma_v^2)$
  - $p(v|D=D')$: Gaussians– $\{\mathcal{N}(\mu_{v,D'},\sigma_{v,D'}^2)\}_{D'}$
  - $p(v|C=C')$: conditional Gaussian– $\mathcal{N}(u_v,\sigma_v^2)$, where $u_v = \mu_v + B_v^T C'$ and B_v are regression weights on C'.
  - $p(v|C=C',D=D')$: conditional Gaussians– $\{\mathcal{N}(u_{v,D'},\sigma_{v,D'}^2)\}_{D'}$.

Figures 1,2,3,4 present how the DAG of a DBN looks for isolated word recognition [3, 4] according to (11),(8),(9),(10), respectively. The variables are defined as follows:

- Deterministic variables
  - Index (discrete): the index of the phoneme state (submodel) within the word model.
  - q_n (discrete): the phoneme state mapped to each index.
- Random variables
  - Trans (discrete): the exit transition from a sub-model.
  - x_n (continuous): the acoustics.
  - m (discrete): the (conditional) Gaussian mixture of x_n.
  - a_n (continuous): the auxiliary information, in this case, pitch or ROS.

We use the BN inference algorithm in [5] to compute $P(v|O)$, the posterior marginal distribution of v given all of the observations O, as well as $P(O|V)$, the likelihood of the observations. Any variable can be observed, hidden, or partially observed, regardless of whether it is continuous or discrete valued. The computed posterior marginal distributions can be used for the expected counts in expectation-maximization (EM) training [6] for learning the discrete probabilities $P(\cdot)$, the means μ, the regression weights B, and the covariances Σ.

4. EXPERIMENTAL TESTING

4.1. General Setup

Using the PhoneBook speech corpus [7] with the small training set defined in [8], we train four mixed BN systems to do speaker-independent, task-independent, isolated-word recognition. There are 41 context-independent, three-state phones in these systems, as well as initial silence and end silence models.

Training was done using the EM algorithm, using a convergence criterion of stopping one iteration after the log-likelihood of the training data increased by less than 0.1%. Each system with auxiliary information was tested two times on the test utterances defined in [8], using lexicons of 75 words:

1. with both X and A observed.
2. with X observed and A hidden; this marginalizes out A and, hence, converts an auxiliary DBN to a baseline DBN (Figure 1), though with different parameter values than the regular baseline DBN.

[3] read, "a_n is conditionally independent of q_n."
[4] also known as directed graphical models

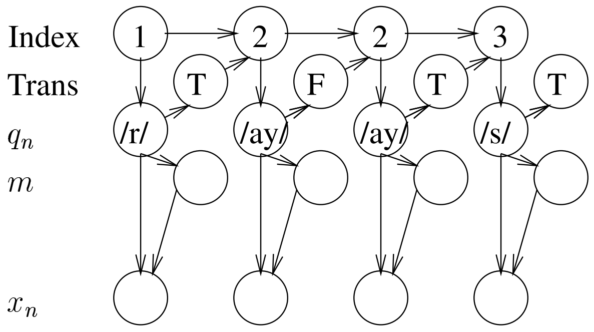

Fig. 1. Baseline Dynamic Bayesian network for isolated word recognition, corresponding to (11)

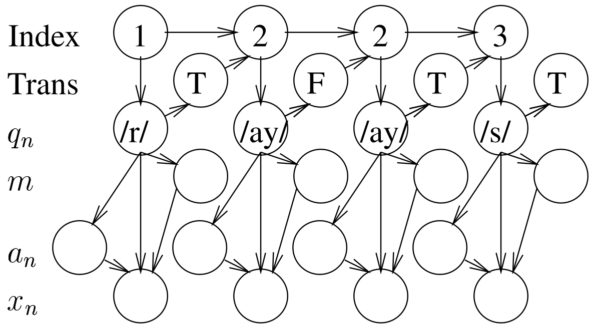

Fig. 2. Auxiliary Baseline Dynamic Bayesian network for isolated word recognition, corresponding to (8)

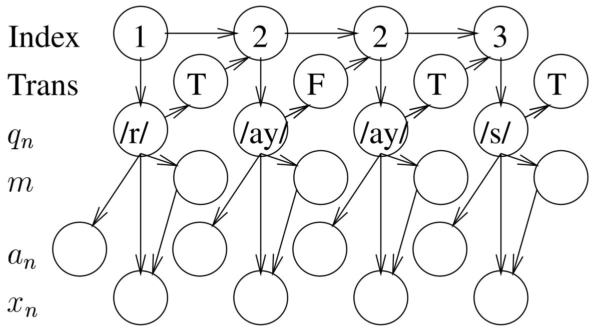

Fig. 3. Auxiliary Dynamic Bayesian network for isolated word recognition with $x_n \perp\!\!\!\perp a_n \mid q_n$, corresponding to (9)

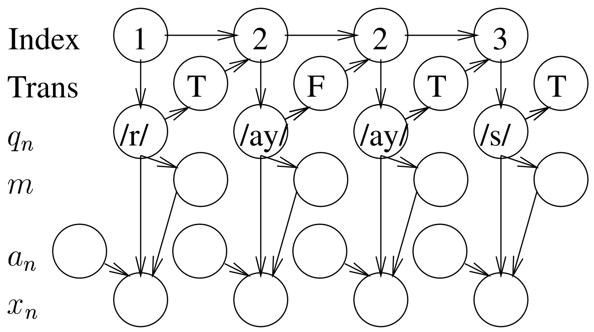

Fig. 4. Auxiliary Dynamic Bayesian network for isolated word recognition with $a_n \perp\!\!\!\perp q_n$, corresponding to (10)

As the DBNs with auxiliary information have different numbers of free parameters depending both upon the assumptions used in Section 2 and upon whether A is observed or hidden, two baseline acoustics-only DBNs, each with a different number of free parameters, are presented with the DBNs with auxiliary information.

Similarly to [3], mel-frequency cepstral coefficients (MFCCs) are extracted from the speech signal, sampled at 8 kHz, using a window of 25 ms with a shift of 8.3 ms for each successive frame. Ten MFCCs with mean subtraction as well as the deltas (first-derivatives) of those ten coefficients and of C_0 are computed for each frame.

4.2. Auxiliary Features

Two different sets of experiments were performed: one with pitch and another with ROS.

4.2.1. Pitch

Pitch is estimated using the simple inverse filter tracking (SIFT) algorithm [9], 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. Results are shown in Table 1.

4.2.2. Rate of Speech (ROS)

Different units for ROS include word rate, syllable rate, phone rate, and normalized phone rate. While a word ROS has been utilized in ASR, work such as [10] has chosen a phone ROS as the phone's length is more stable than that of a word, which can range between containing a single phone or as many as a dozen or more phones. As different phones have different average lengths, the deviation from the normalized length of a phone has been used in [11] as part of the measure of ROS. A syllable ROS measure arose during the development of an estimator of ROS directly from the speech signal [12].

Our work continues in the tradition of [12] of investigating the use of an ROS estimate computed by their *mrate* program directly from the signal and, hence using a syllable ROS measure. *mrate* works best if it has one to two seconds of speech, which typically cover an entire word. Since we are dealing with isolated words, we have computed one ROS value, ros, per isolated word utterance. Therefore, $a_n = ros, \forall n$. Future work would entail using other ROS units. The literature on ROS in ASR looks at incorporating it into the state transition probabilities, the language and pronunciation models, and the acoustic models. It is the incorporation of ROS into the acoustic models that we investigate here. Results are shown in Table 2.

We have used the silence markers provided with PhoneBook so as to run *mrate* only upon the speech segment of the utterance but with the ROS value being assigned to both the speech and non-speech portions of the utterance. We also used these silence markers in the testing, which is unrealistic for real applications.

5. DISCUSSION

With both pitch and ROS DBNs, the performance with the auxiliary variables observed does not improve over that of the baseline

	Mix.	Obs. Pitch	Hid. Pitch
Baseline	4	5.9% (21k)	
Baseline	6	4.3% (32k)	
Pitch Baseline	4	48.9% (32k)	6.2% (21k)
Pitch ($x_n \perp\!\!\!\perp a_n \mid q_n$)	4	60.5% (22k)	19.2% (21k)
Pitch ($a_n \perp\!\!\!\perp q_n$)	4	5.3% (32k)	6.0% (21k)

Table 1. Word error rate for the two Baseline (non-Pitch) DBNs and the three Pitch DBNs. Results for the Pitch DBNs are given with observed and hidden Pitch. For each result, the effective number of parameters is given (i.e., parameters for A subtracted if A is marginalized out). The number of mixtures is given as well.

	Mix.	Obs. ROS	Hid. ROS
Baseline	4	5.9% (21k)	
Baseline	6	4.3% (32k)	
ROS Baseline	4	6.0% (32k)	5.8% (21k)
ROS ($x_n \perp\!\!\!\perp a_n \mid q_n$)	4	6.0% (22k)	5.9% (21k)
ROS ($a_n \perp\!\!\!\perp q_n$)	4	5.8% (32k)	5.7% (21k)

Table 2. Word error rate for the two Baseline (non-ROS) DBNs and the three ROS DBNs. Results are presented as in Table 1.

systems. The auxiliary DBNs perform approximately the same whether they have their auxiliary variables A observed in recognition or whether they are hidden and, thus, marginalized out in recognition; the notable exceptions are the two Pitch DBNs whose performance rises dramatically once the A are marginalized out. However, when the A are marginalized out of the auxiliary DBN, its number of parameters and complexity is reduced while maintaining or improving over the performance achieved with the A observed. In most of these cases with a reduced number of parameters, the performance of the auxiliary DBNs statistically equals the baseline DBN of four mixtures, which has a similar number of parameters. In past work [13], a Pitch DBN ($a_n \perp\!\!\!\perp q_n$) with a single conditional Gaussian and its A marginalized actually performed better than a baseline DBN with a single Gaussian.

Regarding Pitch, the DBN with $x_n \perp\!\!\!\perp a_n \mid q_n$ does very poorly. As mentioned in Section 2, this DBN is nearly the same as standard HMM-based ASR with a_n appended to the standard feature vector. This confirms past difficulty in ASR research in incorporating pitch into ASR. However, the Pitch DBN with $a_n \perp\!\!\!\perp q_n$, in which a_n conditions the distribution of x_n, shows a better way to incorporate pitch into ASR, as also proposed by [1].

Regarding ROS, it may be an error to condition every element in the acoustic vector upon the speaking rate as this may have introduced too much noise. Assuming that MFCC *derivatives* are different in fast speech, we would like to make only the MFCC derivatives be dependent upon a_n. Furthermore, our system assumes a linear relationship between x_n and a_n within the conditional Gaussian. Perhaps this relationship is better modeled non-linearly. If this is so and could be incorporated within future systems, this may help to improve the performance in fast speech. Finally, other units for ROS, specifically phone ROS, should be looked at in this framework. These can be estimated using a forced alignment of the data.

Acknowledgments

Todd A. Stephenson and Mathew Magimai-Doss are supported by the Swiss National Science Foundation under grants FN 2000-064172.00/1 and FN 2100-057245.99/1, respectively. We thank Jaume Escofet for his contributions to setting up the experiments and Samy Bengio for his contributions to the analysis.

6. REFERENCES

[1] Katsuhisa Fujinaga, Mitsuru Nakai, Hiroshi Shimodaira, and Shigeki Sagayama, "Multiple-regression hidden Markov model," in *ICASSP*, Salt Lake City, Utah, USA, May 2001, vol. 1, pp. 513–516.

[2] Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, Statistics for Engineering and Information Science. Springer-Verlag New York, Inc., 1999.

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

[4] Geoffrey Zweig and Mukund Padmanabhan, "Dependency modeling with Bayesian networks in a voicemail transcription system," In Eurospeech '97 [14], pp. 1135–1138.

[5] Steffen L. Lauritzen and Frank Jensen, "Stable local computations with conditional Gaussian distributions," *Statistics and Computing*, vol. 11, no. 2, pp. 191–203, April 2001.

[6] Steffen L. Lauritzen, "The EM algorithm for graphical association models with missing data," *Computational Statistics & Data Analysis*, vol. 19, pp. 191–201, 1995.

[7] 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 '95 [15], pp. 101–104.

[8] 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*, Munich, April 1997, vol. 3, pp. 1767–1770.

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

[10] Matthew A. Siegler and Richard M. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," In ICASSP '95 [15], pp. 612–615.

[11] F. Martínez, D. Tapias, J. Álvarez, and P. León, "Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition," In Eurospeech '97 [14], pp. 469–472.

[12] Nelson Morgan, Eric Fosler, and Nikki Mirghafori, "Speech recognition using on-line estimation of speaking rate," In Eurospeech '97 [14], pp. 2079–2082.

[13] Todd A. Stephenson, Mathew Magimai-Doss, and Hervé Bourlard, "Mixed Bayesian networks with auxiliary variables for automatic speech recognition," in *ICPR*, Quebec City, PQ, Canada, August 2002, to appear.

[14] *Eurospeech*, Rhodes, Greece, September 1997.

[15] *ICASSP*, Detroit, MI, May 1995.