IDIAP RESEARCH REPORT

# Automatic Speech Recognition using Dynamic Bayesian Networks with the Energy as an Auxiliary Variable

Jaume Escofet Carmona [a]
Todd A. Stephenson [b]

IDIAP–RR 03-18

March 2003

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] with the Technical University of Catalonia (UPC), Barcelona, Spain. This work was performed while visiting IDIAP under the European Masters in Language and Speech

[b] also with the Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland, and supported during this work by the Swiss National Science Foundation under the grant BN_ASR (20-64172.00).

IDIAP Research Report 03-18

# Automatic Speech Recognition using Dynamic Bayesian Networks with the Energy as an Auxiliary Variable

Jaume Escofet Carmona Todd A. Stephenson

March 2003

**Abstract.** In current automatic speech recognition (ASR) systems, the energy is not used as part of the feature vector in spite of being a fundamental feature in the speech signal. The noise inherent in its estimation degrades the system performance. In this report we present an alternative approach for introducing the energy into the system so that it can help to enhance recognition. We present the experimental results of an ASR system based on dynamic Bayesian networks (DBNs) using the energy as an auxiliary variable. DBNs belong to the same family of statistical models as hidden Markov models (HMMs). However, DBNs are a more general framework and they allow more flexibility in defining new probabilistic relations between variables. We tried different network topologies and we noticed the benefit of conditioning the feature vector on the energy. Furthermore, hiding the value of the energy in recognition also improved the recognition performance.

# 1 Introduction

In standard ASR some acoustic features are extracted from the speech signal at each time frame. As a result, a sequence of feature vectors $X = \{x_1, \dots, x_n, \dots, x_N\}$ (from time 1 to time N) is produced and used as the input of a statistical model. The model computes at each time $n$ the likelihood of the feature vector $x_n$ given the current hidden state $q_n = k$:

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

Currently, the most widely used spectral representation are the MFCCs (mel frequency cepstral coefficients). The energy is a basic feature in the speech signal. However, it has been shown that it does not help to improve the recognition performance if it is introduced into the feature vector. The energy, though, is such a fundamental feature that it must have some effect on the parameters extracted from the speech signal. In our experiments we tried to model the correlation between the energy and the MFCCs. Instead of (1) we compute the following likelihood:

$$p(x_n|a_n = z, q_n = k), \tag{2}$$

where $a_n$ is an auxiliary variable (the energy in this work) with value $z$. Now the feature vector is not only dependent upon the hidden state, but also upon the energy.

We chose the framework of DBNs because of their flexibility in modifying both the topology and the distributions of the variables. Apart from the system which models (2) we have tested other topologies with different probabilistic relations between $x_n$, $q_n$ and $a_n$. The DBNs also allow us to marginalize out any variable in the network. In some of the experiments we used the energy in the training process to estimate the parameters of the model, but we marginalized out (*hid*) its value in recognition. We achieved good results following this procedure.

# 2 Dynamic Bayesian networks (DBNs)

A DBN is a Bayesian network (BN) whose variables evolve along time. A BN is a graphical model composed of the following items:

- A set of random variables $\boldsymbol{V} = \{V_1, \dots, V_w, \dots, V_W\}$. They can be discrete or continuous.
- A directed acyclic graph (DAG) where each of its vertices is associated with one variable $V_w \in \boldsymbol{V}$
- A set of probability distributions. Each variable has a distribution conditioned on the variables which have edges pointing on it (i.e., its parents):

$$P(V_w|parents(V_w))$$

The joint distribution of all the variables in the BN is defined as the product of all the local distributions:

$$P(\boldsymbol{V}) = \prod_{w=1}^{W} P(V_w|parents(V_w))$$

Figure 1 shows a DBN for isolated word speech recognition, as defined in [8]. It is composed of the following variables:

- **position**: sub-model index within the word model.
- **phoneme** ($q_n$): the hidden phonetic state associated with the current position.
- **transition**: binary variable that indicates whether the sub-model has to change or not.
- **feature vector** ($x_n$): acoustic observations extracted from the speech signal.

The variables *position, transition* and *phoneme* compose the control layer as they 'control' the legal sequences of phone models.

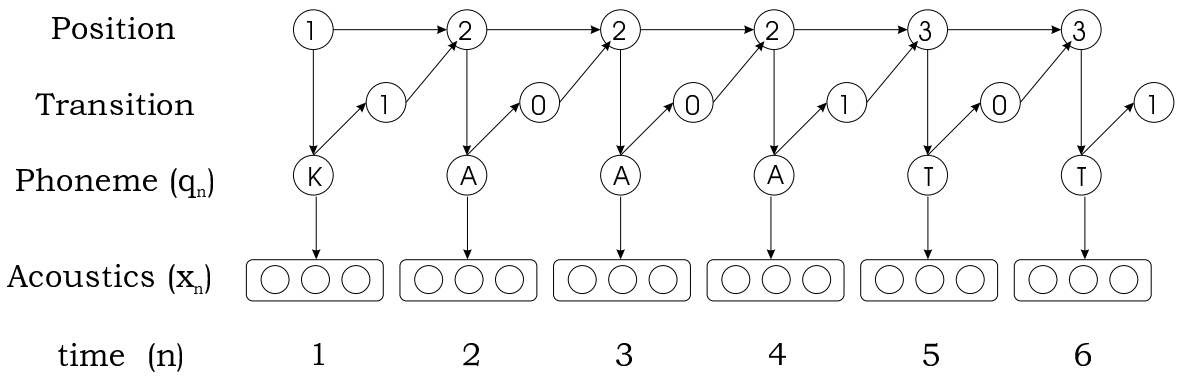


Figure 1: DBN modelling the word 'kat'.

# 3 Auxiliary information

We call auxiliary information a feature that is not part of the feature vector but which conditions it in some way. Along the time that the word is pronounced a sequence of auxiliary information $A = \{a_1, \ldots, a_n, \ldots, a_N\}$ (for time $n = 1, \ldots, N$) is obtained. Different statistical independence assumptions can be made between the feature vector $x_n$, the auxiliary information $a_n$ and the hidden state $q_n$.

In standard ASR the likelihood of the observations conditioned on the hidden state is computed using a Gaussian distribution with mean $\mu_k^x$ and covariance matrix $\Sigma_k^x$ (diagonal in standard approaches):

$$p(x_n|q_n = k) \sim \mathcal{N}_x(\mu_k^x, \Sigma_k^x) \tag{3}$$

In our work the short-term energy is used as an auxiliary variable. The traditional way to introduce the energy, or any other feature, into the ASR system is as part of the feature vector. Appending $a_n$ to the feature vector results in the following Gaussian distribution:

$$p(x_n, a_n|q_n = k) \sim \mathcal{N}_{x,a}(\mu_k^{x,a}, \Sigma_k^{x,a}) \tag{4}$$

In standard approaches the expanded covariance matrix $\Sigma_k^{x,a}$ would be diagonal, thus assuming independence between $a_n$ and $x_n$. This assumption is not reasonable in the case of the energy. Such a fundamental feature must be correlated with the MFCCs. Therefore we should model the correlation between the feature vector and the auxiliary variable. The distribution of $x_n$ must be changed according to the value of $a_n$. This can be done using conditional Gaussians:

$$p(x_n|q_n = k, a_n = z) \sim \mathcal{N}_x(\mu_k^x + B_k^T z, \Sigma_k^x) \tag{5}$$

The mean of this Gaussian is a regression on the mean of $x_n$ and the value of $a_n$. $B_k$ is the matrix containing the weights upon $z$, the value of $a_n$. The covariance matrix does not change with $z$. In this situation we assume that $a_n$ is independent of the hidden state $q_n$: $p(a_n|q_n) = p(a_n)$. However, we could model the correlation between $q_n$ and $a_n$ using a Gaussian distribution:

$$p(a_n|q_n = k) \sim \mathcal{N}_a(\mu_k^a, \Sigma_k^a) \tag{6}$$

Then, the joint emission likelihood of $x_n$ and $a_n$ can be computed by multiplying (5) and (6):

$$p(x_n|q_n = k, a_n) \cdot p(a_n|q_n = k) \sim \mathcal{N}_x(\mu_k^x + B_k^T z, \Sigma_k^x) \otimes \mathcal{N}_a(\mu_k^a, \Sigma_k^a), \tag{7}$$

where $\otimes$ is the combination operator for Gaussians.

### Marginalizing out auxiliary information

During training we use the auxiliary information to better estimate the parameters of the models. In recognition, however, we have the choice of using or not using the value of the auxiliary variable. There are many reasons for hiding the auxiliary information in recognition. For example, if we had a database with information about the articulators position, we could train some models using this information as auxiliary variable [4]. However, normally in recognition this information is not available and it is not possible to extract it from the speech signal; so we have to marginalize out the auxiliary variable. There is another type of information that we can estimate from the signal; but the estimation is not accurate enough, or there is noise on it. This is the case of pitch or energy [5]. In the training process we use a large amount of data from which we can extract some relevant statistical information. However, in recognition there is only one sample of A. In its estimation there might be noise that could hurt the recognition performance. We can hide the auxiliary variable $a_n$ by integrating over all its values. For the different topologies explained above, we can obtain the distribution of $x_n$ as follows:

$$\begin{aligned} p(x_n|q_n) &= \int p(x_n, a_n|q_n)\, \mathrm{d}a_n & (8)\\ &\approx \int p(x_n|q_n, a_n) \cdot p(a_n|q_n)\, \mathrm{d}a_n & (9)\\ &\approx \int p(x_n|q_n, a_n) \cdot p(a_n)\, \mathrm{d}a_n & (10) \end{aligned}$$

where (8), (9) and (10) apply to (4), (5) and (7), respectively.

## 4 Experiments

Three sets of experiments were performed. In the first one, the short-term energy was defined as the auxiliary variable. In the second set of experiments we used the logarithm of the energy as the auxiliary information. In the last set of experiments we tested a new topology based on equivalence classes. In this new structure $a_n$ depends upon the class of phoneme, instead of the phoneme itself. In the following sections we describe in detail the systems tested in each set of experiments.

### 4.1 Setup

We did experiments of speaker-independent, task-independent, isolated word recognition. The telephone speech corpus *Phonebook* [3] was used. For the training we used the *small training set* defined in [1] (composed of 19421 utterances). The test set consisted of 6598 utterances, as defined in [1]. The lexicon was composed of 75 words. The transcription was done with a dictionary of 41 context-independent phonemes. We defined 3 states per phoneme, as well as initial silence and end silence models. The EM algorithm was used in the training, with a convergence criterion of stopping one iteration after the log-likelihood of the training data increased by less than 0.1% over that of the previous iteration.

The speech signal had a sampling frequency of 8 kHz. A set of features was extracted from the signal, with a shift of 8.3 ms for each successive frame, using a Hamming window of 25 ms. The feature vector $x_n$ was composed of 10 mel-frequency cepstral coefficients (MFCCs) (from 1 to 10), their deltas

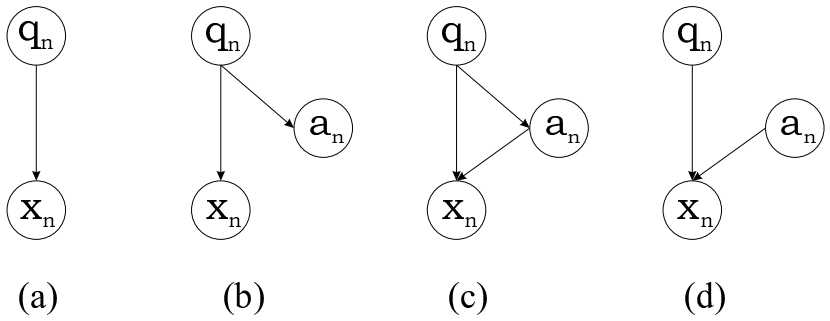


Figure 2: Different topologies that define different relations of probabilistic dependency between the state, the auxiliary variable and the feature vector.

(first derivative) plus the delta of the 0th coefficient. In all of our systems, $x_n$ was modelled with 4 mixtures of Gaussians.

For the first and second sets of experiments we tested three systems with different topologies, apart from the baseline. In Figure 2 we have the portion of DBN that describes the relation between $q_n$ (state), $x_n$ (observations) and $a_n$ (energy) for each system. The systems defined in these experiments are the following:

- **Baseline**: Figure (2a) represents a portion of DBN for ASR where there is no auxiliary information. It is equivalent to a standard HMM. The underlying equation is (3).
- **System 1**: $a_n$ depends upon $q_n$; $x_n$ independent of $a_n$. Figure (2b) uses Equation (4). $q_n$ is parent of $x_n$ and $a_n$. This system is equivalent to an HMM with $a_n$ appended to the feature vector $x_n$ (in our experiments this is true except that $x_n$ is modelled with 4 mixtures of Gaussians while $a_n$ is modelled with one single Gaussian).
- **System 2**: $a_n$ conditioned on $q_n$; $x_n$ depends upon $a_n$. Figure (2c) uses Equation (7). Now $x_n$ has 2 parents: $q_n$ and $a_n$. Furthermore, the value of $a_n$ depends on the state $q_n$.
- **System 3**: $a_n$ independent of $q_n$; $x_n$ conditioned on $a_n$. Figure (2d) uses Equation (5). The difference with System 2 is that in this case $a_n$ does not depends on the phoneme $q_n$.

Having used observed auxiliary information in training, each one of the systems with auxiliary information was tested twice: with $a_n$ (energy) observed and with it hidden.

## 4.2 Short-term energy

In a first set of experiments the auxiliary variable was defined as the short-term energy and was computed as follows:

$$a_n = \frac{1}{C} \sum_{t=1}^{T} s_n^2[t] \cdot w^2[t] \tag{11}$$

where $\{s_n[1], \ldots, s_n[t], \ldots, s_n[T]\}$ is the speech signal of $T$ samples associated with time frame $n$, and $\{w[1], \ldots, w[t], \ldots, w[T]\}$ is a Hamming window, and $C$ is a normalizing constant used to give manageable values for the short-term energy.

Table 1 shows the results (expressed in Word Error Rate) obtained with each system. System 1 presents a very poor performance when the energy is observed. This result prove that incorporating the energy into the feature vector damage the system performance (from 5.9% in the baseline we rise

| DBN | Fig | Eq | Observed energy | Hidden energy |
|---|---|---|---|---|
| Baseline | (2a) | (3) | 5.9% | |
| System 1 | (2b) | (4) | 28.9% | 6.3% |
| System 2 | (2c) | (7) | 27.4% | 5.9% |
| System 3 | (2d) | (5) | 5.9% | 19.4% |

Table 1: Word Error Rate obtained with each system using the short-term energy as an auxiliary variable. The correspondent figure and equation are also given.

| DBN | Fig | Eq | Observed energy | Hidden energy |
|---|---|---|---|---|
| Baseline | (2a) | (3) | 5.9% | |
| System 1 | (2b) | (4) | 6.9% | 5.3% |
| System 2 | (2c) | (7) | 6.1% | 5.6% |
| System 3 | (2d) | (5) | 5.8% | 6% |

Table 2: WER using the logarithm of the energy as an auxiliary variable.

up to 28.9%). However, hiding the energy provides a very positive effect, thus achieving a WER close to the baseline. System 2 has a similar behaviour. We achieve a great improvement by marginalizing out the energy, so as to equalise the baseline.

System 3 models the correlation between $a_n$ and $x_n$, and we assume independence between $a_n$ and $q_n$. This scheme has a beneficial effect (compared to the other two systems) when the energy is observed. However, in this case marginalization seems to be hurtful. This may be produced because of a bad estimation of $p(a_n)$ used in (10).

### 4.3 Logarithm of the energy

In a new set of experiments we try to compute the energy in a different way in order to achieve better results. Specifically, the auxiliary variable was defined as the logarithm of the short-term energy. It was computed by applying the logarithm function to (11). This computation is meant to have a better behaviour because the logarithmic function emulates the human auditory system response.

The systems illustrated in Figure 2 were tested, under the same conditions as before. The results are given in Table 2. In general there has been an improvement in all the systems, compared to the results with the short-term energy. The difference between systems have been relaxed, but they are still significant.

System 1 with $a_n$ observed provides again the poorest performance. This confirms again the difficulty in incorporating the energy in the traditional way. Hiding the energy has a positive effect. We even decrease the WER in the baseline by a relative 10% (5.9% to 5.3%). Similar results are obtained with System 2, which also improves the system performance when $a_n$ is hidden. System 3 provides a slightly better result when the energy is observed.

### 4.4 Equivalence classes

Finally, in the last set of experiments a new topology, hybrid between (5) and (7) (Systems 2 and 3) was defined. It is difficult to assure that the energy has a clear dependency on $q_n$. However, there are groups of phonemes characterised by a certain value of energy. For example, in general vowels present higher energy than consonants. The idea of these experiments is to assume a correlation between the energy and the class of phoneme. We did two groups of experiments with equivalence classes. Firstly we defined 2 classes: *Silence* and *Phone*. In a second set of experiments 3 classes were defined: *Silence, Vowel* and *Consonant*. Figure 3 illustrates the new topology based on equivalence classes.

The auxiliary variable was defined as short-term energy as well as logarithm of the energy. Each system was tested with the auxiliary variable observed and hidden, having always been observed in

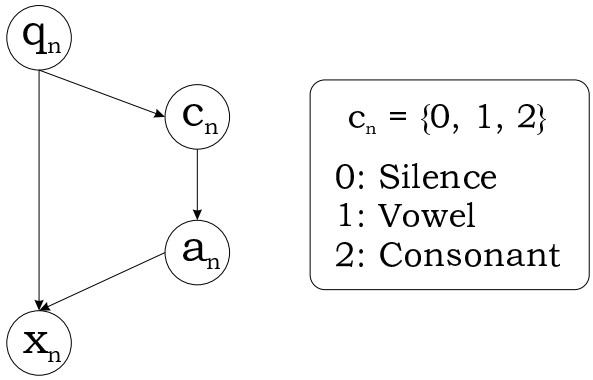


Figure 3: Portion of BN for ASR using auxiliary information an equivalence classes. In this example there are 3 classes defined. $c_n$ is a deterministic variable that changes its value according to the current phone. The auxiliary variable depends upon the class of phone, given in $c_n$.

| DBN | Obs. Energy | Hid. Energy |
|---|---|---|
| Baseline | 5.9% | |
| STEn Sil Phon | 6% | 4.8% |
| STEn Sil Vow Cons | 6.3% | 4.6% |
| LogEn Sil Phon | 4.7% | 4.7% |
| LogEn Sil Vow Cons | 5.1% | 5.3% |

Table 3: WER obtained with systems based on equivalence classes. Short-term energy (*STEn*) and logarithm of the energy (*LogEn*) were used as auxiliary information. Two classes (Silence, Phone) and three classes (Silence, Vowel, Consonant) were tested.

training. The results are given in Table 3. Marginalization produces a great improvement if short-term energy is used as auxiliary information. The best result is achieved with 3 classes, a reduction of a relative 22% is achieved on baseline WER (from 5.9% to 4.6%). On the other hand, the logarithm provides good results no mater $a_n$ is observed or hidden.

The energy provides its best result with 3 classes. Instead, the logarithm performs better with 2 classes, probably because in this case there is a greater difference between *Phone* and *Silence*, as opposed to *Vowel* and *Consonant*.

The new approach presented in this section works better than the systems tested before. There was not need of computing new features, or transforming the energy computation. We get a significant improvement just re-defining the statistical relations between variables in the model in a more 'reasonable' way.

# 5 Conclusion

We have presented an appropriate approach for introducing the energy into an ASR system. The standard systems incorporate the energy into the feature vector. This approach have been proved to performs poorly. Our work demonstrates the great benefit of conditioning the feature vector on the energy or hiding the energy in recognition. The benefit can be greater if the logarithm of the energy is computed. In future experiments, other estimations of the energy could be done, as long-term energy or the energy of a frequency sub-band [2].

The best results were achieved with the equivalence classes approach. For future work we propose to use other classifications. For example, we could join the nasal and liquid consonants (e.g., m, n, l) with the vowels, as they present on average similar values of energy.

The database used in our experiments was composed of utterances where the speakers were asked to read some words, so the speech is not spontaneous. It would be interesting to test our systems in natural, spontaneous speech, as the intensity of the voice strongly changes during the speech in natural conditions [7].

Finally we note that other experiments have been done using other features as auxiliary information. This is the case of pitch [6], speaking rate or articulatory information [4]. In our work, we achieved significant performance improvement by using the simple and easily computable feature of energy.

# References

[1] Stéphane Dupont, Hervé Bourlard, Olivier Deroo, Vincent Fontaine, and Jean-Marc Boite. Hybrid HMM/ANN systems for training independent tasks: Experiments on phonebook and related improvements. In *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97)*, volume 3, pages 1767–1770, Munich, April 1997.

[2] Katsuhisa Fujinaga, Mitsuru Nakai, Hiroshi Shimodaira, and Shigeki Sagayama. Multiple-regression hidden Markov model. In *ICASSP*, volume 1, pages 513–516, 2001.

[3] John F. Pitrelli, Cynthia Fong, Suk H. Wong, Judith R. Spitz, and Hong C. Leung. PhoneBook: A phonetically-rich isolated-word telephone-speech database. In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95)*, volume 1, pages 101–104, Detroit, MI, May 1995.

[4] Todd A. Stephenson, Hervé Bourlard, Samy Bengio, and Andrew C. Morris. Automatic speech recognition using dynamic Bayesian networks with both acoustic and articulatory variables. In *6th International Conference on Spoken Language Processing: ICSLP 2000 (Interspeech 2000)*, volume 2, pages 951–954, Beijing, October 2000.

[5] Todd A. Stephenson, Jaume Escofet, Mathew Magimai-Doss, and Hervé Bourlard. Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables. In *2002 IEEE International Workshop on Neural Networks for for Signal Processing (NNSP 2002)*, Martigny, Switzerland, September 2002.

[6] Todd A. Stephenson, Mathew Magimai-Doss, and Hervé Bourlard. Mixed Bayesian networks with auxiliary variables for automatic speech recognition. In *International Conference on Pattern Recognition (ICPR 2002)*, Quebec City, PQ, Canada, August 2002.

[7] Todd A. Stephenson, Mathew Magimai-Doss, and Hervé Bourlard. Speech recognition of spontaneous, noisy speech using auxiliary information in Bayesian networks. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, April 2003. To appear.

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