Automatic Speech Recognition using Pitch Information in Dynamic Bayesian Networks

Todd A. Stephenson [a,b]
Mathew Magimai Doss [a]　Hervé Bourlard [a,b]

IDIAP–RR 00-41

November 2000

submitted for publication

[a] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
[b] Swiss Federal Institute of Technology at Lausanne (EPFL)

IDIAP Research Report 00-41

Automatic Speech Recognition using Pitch Information in Dynamic Bayesian Networks

Todd A. Stephenson Mathew Magimai Doss Hervé Bourlard

November 2000

submitted for publication

Abstract. The challenge of automatic speech recognition (ASR) increases when speaker variability is encountered. Being able to automatically use different acoustic models according to speaker type might help to increase the robustness of ASR. We present a system that attempts to do so by augmenting the standard acoustic observations with pitch information. This allows the system to use acoustic models more appropriate to speech with the given pitch. Furthermore, pitch information is more easily detected in noisy conditions; thus, it may be of use in robust speech recognition. Using dynamic Bayesian networks (DBNs) allows further refinement of the system by eliminating unnecessary statistical dependencies and thus reducing the number of parameters. We show that when a system is trained on observed pitch data and performs recognition with missing pitch data, it can perform significantly better than a system that uses acoustics information only.

Acknowledgements: Todd A. Stephenson and Mathew Magimai Doss are supported by the Swiss National Science Foundation under grants # 2100-053960.98/1 and 2100-057245.99/1, respectively. We also thank Samy Bengio and Andrew C. Morris for their statistics help.

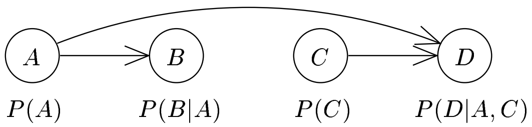

Figure 1: A sample Bayesian network with variables A, B, C, and D, along with the corresponding DAG and probability distributions.

1 Introduction

Speech is the most common mode of communication among human beings. The objective of speech recognition is to recognize the message being spoken. The speech recognition system performance is mostly affected by speaker variability, channel noise, etc. In the literature, different approaches have been proposed to adapt the model to a speaker to improve the performance of the system [1]. In this paper, we describe an approach to use pitch within the framework of dynamic Bayesian networks (DBNs) [2, 3] to reduce the effect of speaker variability on the performance of the system.

In [4], we showed how a DBN can be used to incorporate the auxiliary information of articulator positions into speech recognition. While keeping the standard acoustics variable, we included an additional variable to hold a causal variable for the acoustics; this variable held the causal articulator information. In that work, only certain articulatory information was used: the lips, the tongue, and the jaw. In this paper, we present a system that models another cause of the acoustics: the pitch. The goal behind using pitch is do speaker clustering within the acoustic models. That is, the acoustic model used depends on the pitch for the current frame. If the pitch is missing for the current frame, the DBN uses the expected distribution of the pitch values to apply a weight to the different acoustic model cluster possibilities.

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 [5]. In this paper, we define pitch to be F_0. Pitch is a speaker-specific feature and is used for speaker recognition [6]. In this work, by clustering the pitch we intend to cluster the speakers so as to derive appropriate models which could help in reducing the effect of speaker variability on the system performance. 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).

2 Bayesian Networks

We are using DBNs as our models because they allow more flexibility than hidden Markov models (HMMs) in modifying the model topology and in handling missing data [7]. A DBN is actually a generalization of an HMM and has the potential to take on a wider range of topologies. A Bayesian network (BN) is defined by the following three items (see Figure 1):

1. a set of variables $\mathbf{X}$ that represents all items that you are attempting to model (e.g., *Acoustics*, *Transitions*,*Phonemes*, *Pronunciation*, etc.)

2. a directed acyclic graph (DAG) whose edges incorporate the conditional (in-)dependencies of $\mathbf{X}$ (e.g., a directed edge from *Phoneme* to *Acoustics* indicates that *Acoustics* is dependent on *Phoneme*).

3. a local probability distribution for each $x_i \in \mathbf{X}$:

$$P(x_i|\text{parents}(x_i)) \tag{1}$$

The joint probability of all of the variables in the BN is represented by the product of all the component local probability distributions:

$$P(x_1, x_2, \ldots, x_N) = \prod_{1,\ldots,N} P(x_i|\text{parents}(x_i)) \tag{2}$$

It is important to see in Figure 1 that, unlike in HMMs, the edges in a BN do not themselves carry any probabilities. Rather, they dictate the conditional variables within the probability distributions of each variable.

3 Using Acoustic and Pitch Variables in a Dynamic Bayesian Network

One approach to using *Acoustics* and *Pitch* variables in an *HMM* would be to concatenate them into a single feature vector. You would then be utilizing it in the HMM emission probability,

$$P(x_t, p_t|q_k), \tag{3}$$

where x_t is the current acoustic vector at time t, p_t the associated pitch value (estimated as explained in Section 5.2.1), and q_k the hypothesized phonetic state. DBNs facilitate making changes to (3) to allow better modeling. For example, you can easily take the following steps (though there are other possibilities as well):

1. the joint probability of *Acoustics* data and of *Pitch* data given the hidden phonetic state can be factored as follows:
$$P(x_t, p_t|q_k) = P(x_t|p_t, q_k)\, P(p_t|q_k). \tag{4}$$
This step itself does not theoretically change the system but allows the succeeding changes to be done. Note that the BN framework allows you to easily deal with either variable (*Acoustics* or *Pitch*) being missing with the other being observed.

2. you can easily introduce temporal dependencies to only one of the variables while leaving the other independent of the past, given its parent(s). For example, you can add a temporal dependency for the *Pitch* variable: $p_{t-1} \Longrightarrow p_t$. Thus, adding this dependency to (4) gives:
$$P(x_t, p_t|q_k) = P(x_t|p_t, q_k)\, P(p_t|p_{t-1}, q_k). \tag{5}$$

3. you can make (conditional) independence assumptions [8] that affect only certain of the factors and not others. For example, if p_t and q_k are independent then
$$P(p_t|p_{t-1}, q_k) = P(p_t|p_{t-1}). \tag{6}$$
This then simplifies (5):
$$P(x_t, p_t|q_k) = P(x_t|p_t, q_k)\, P(p_t|p_{t-1}). \tag{7}$$

4 The Dynamic Bayesian Network for Automatic Speech Recognition

Figure 2 presents the DBN, based on [7], for doing ASR of isolated-words (our chosen task) with both acoustic and pitch information. It has the following variables:

- Deterministic variables
    1. *Position*: the current sub-model number of the word model.
    2. *Phone*: maps the sub-model number from *Position* to an actual phonetic model.

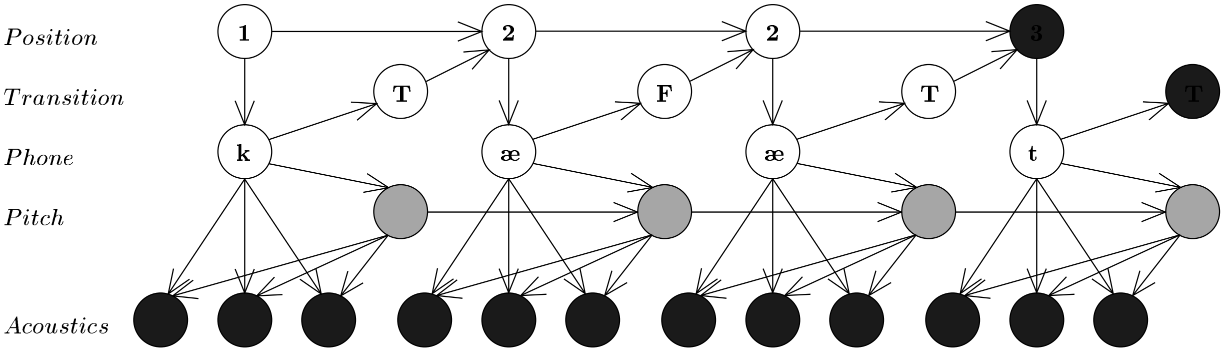

Figure 2: Pitch-Acoustics DBN, based on equation (5). The black variables (*Acoustics* and the final *Position* and *Transition*) are always observed in training and recognition. The grey variables (*Pitch*) are always observed in training but are hidden in some recognition tests.

- Stochastic variables

  1. *Transition*: the probability of exiting from this phonetic model (that is, the next frame will be in a new phonetic model).
  2. *Pitch*: the probability of the pitch for this frame.
  3. *Acoustics*: the probability of the acoustics for this frame (also known as the emission probability).

If the temporal dependencies between the *Pitch* variables($Pitch_{t-1} \Longrightarrow Pitch_t$) in Figure 2 are removed, then the DBN is functionally equivalent to a standard ASR HMM. That is, assuming that both the *Acoustics* and the *Pitch* variables are always observed, it is theoretically the same as an HMM where the emission is a vector composed of both the acoustics and the pitch. However, as explained in Section 3, using the DBN framework allows you to easily explore different relations among the variables such as the temporal dependency over time of *Pitch* and the independence of *Pitch* from *Phone*.

5 Experiments

5.1 Database

Our experiments were set within the task of speaker-independent, task-independent, isolated word recognition. That is, none of the speakers used in training were used in the recognition evaluation; likewise, none of the isolated-words in training were used in the recognition evaluation. The database used in our experiments was PhoneBook [9], which was initially used in speech recognition experiments in [10].

5.2 Preprocessing

5.2.1 Feature Calculation

Similarly to [7], mel-frequency cepstral coefficients (MFCCs) were extracted from the 8 kHz signal using a window of 25 ms with a shift of 8.3 ms for each successive frame. Cepstral mean subtraction and energy normalization were performed. Ten MFCCs plus C_0 (the energy coefficient) as well as the deltas (first-derivatives) of those eleven coefficients were computed for each frame, using 20 filterbanks and a pre-emphasis coefficient of 0.97.

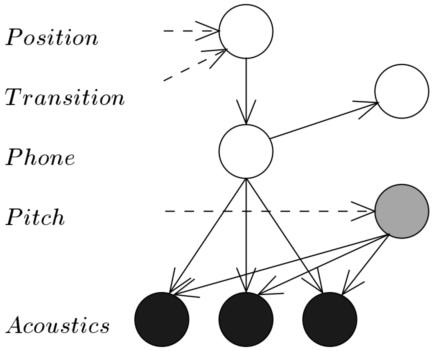

Figure 3: Phone-independent Pitch-Acoustics DBN, based on equation (7). This is one time-slice of Figure 2 with the edge between *Phone* and *Pitch* removed, making *Pitch* independent of *Phone*. The dashed lines are connections coming from the previous time-slice.

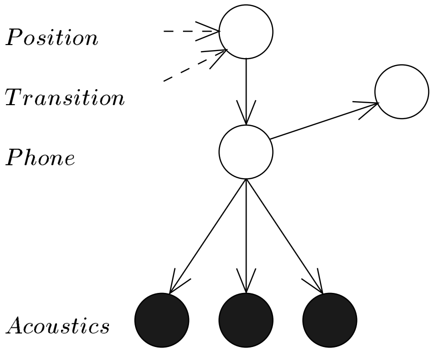

Figure 4: Acoustics DBN (the baseline DBN). This is one time-slice of Figure 2 with *Pitch* removed. Thus, this is equivalent to the standard, acoustics-only, discrete ASR HMM. The dashed lines are connections coming from the previous time-slice.

The pitch for our studies is estimated using the *S*imple *I*nverse *F*ilter *T*racking (SIFT) algorithm [11], 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.

5.2.2 Discretization

BNs are mostly easily used when all variables are in the discrete domain. Therefore, all real-valued data (acoustic features and pitch) were discretized using K-means clustering.

As in [7], the ten MFCCs, the ten delta-MFCCs, the one C_0, and the one delta-C_0 were clustered according to 256, 256, 16, and 16 prototypes, respectively. The C_0 and delta-C_0 values were then combined into a single, 256 prototype stream. Thus, there were three acoustic streams: MFCC, delta-MFCC, and C_0/delta-C_0, each with 256 prototypes.

Different prototype sizes were tried for the *Pitch* data: 2, 4, and 8. In all cases, one of the prototypes was reserved for the pitch value of 0 (i.e., unvoiced speech). All of the non-zero pitch values (i.e., voiced speech) were then clustered according to the number of remaining prototypes (1, 3, and 7, respectively). Note that in the 2 prototype case, this effectively means that the two prototypes only distinguish between voiced and unvoiced speech and would therefore not be clustering speakers; the DBN from Figure 2 with only 2 pitch prototypes would thus be similar to the 'articulator' DBN

	WER		# Parameters
Baseline (No Pitch)	7.8%		33k
	Observed Pitch	Missing Pitch	
2 Pitch Prototypes	8.5%	7.6%	66k
4 Pitch Prototypes	7.9%	7.1%	133k
8 Pitch Prototypes	8.6%	N/A	270k

Table 1: Pitch-Acoustics DBN vs. Acoustics DBN performance, in terms of word error rate (WER). Using the same trained systems, recognition was done in two different ways: one with the *Pitch* variable observed and the other with it being *missing.* (N/A: Not Available).

	WER		# Parameters
Baseline (No Pitch)	7.8%		33k
	Observed Pitch	Missing Pitch	
2 Pitch Prototypes	8.5%	7.7%	65k
4 Pitch Prototypes	8.0%	N/A	131k
8 Pitch Prototypes	8.9%	N/A	263k

Table 2: Phone-independent Pitch-Acoustics DBN vs. Acoustics DBN performance, in terms of word error rate (WER). Using the same trained systems, recognition was done in two different ways: one with the *Pitch* variable observed and the other with it being *missing.* (N/A: Not Available).

in [7] except that he kept the *Pitch* variable (which he called the *Context* variable) hidden while we kept it observed (except as noted below).

5.3 Tests

Theoretically equivalent to a standard, discrete ASR HMM, an Acoustics DBN (Figure 4) was trained as a baseline for all tests. Two DBNs with an included *Pitch* variable were also used: the Pitch-Acoustics DBN (Figure 2) and the Phone-independent Pitch-Acoustics DBN (Figure 3). When referring to the Pitch-Acoustics DBN and the Phone-independent Pitch-Acoustics DBN collectively, we will use the short term 'Pitch DBN'. Forty-one monophones, were used; with beginning and ending silence, this gives a total of 43 'phones'. There were three states per phone, resulting in 129 hidden phonetic states for each DBN. We used the 'small' training set and validation set as defined in [10] for PhoneBook. All of the systems presented here were trained on this training set using expectation-maximization (EM) training with observed acoustics and observed pitch information; we considered the training to have converged when the data log likelihood increased by less than 1% from the previous EM iteration. All of the recognition results presented here were performed on the validation set, which has its own unique speakers and tasks.

Table 1 presents results using the Pitch-Acoustic DBN from Figure 2 while Table 2 presents results using the Phone-independent Pitch-Acoustics DBN from Figure 3. The baseline Acoustics DBN is only significantly[1] better than the Pitch DBNs that have 8 pitch prototypes (with *Pitch* observed). The baseline system is not significantly better than any of the other systems with *Pitch* observed. Furthermore, the Pitch-Acoustics DBN does not do significantly better than the Phone-independent Pitch-Acoustics DBN.

Having had good results in [4] in treating a causal variable (i.e., one in the same position as *Pitch*) as hidden, we also hid the *Pitch* data for some additional recognition tests. That is, we used the exact same systems as used in the "Observed" experiments of Tables 1 & 2 (trained with *observed*

[1] All significance tests used for this paper are at 95% confidence.

Pitch data) but treated the *Pitch* data as missing during recognition. In all of the results provided in Tables 1 & 2, any given Pitch DBN does significantly better with the pitch data missing than the same Pitch DBN with observed pitch data. Furthermore, the Pitch-Acoustics DBN with 4 pitch prototypes with missing pitch data does *significantly* better than the baseline Acoustics DBN.

6 Discussion

We have the following conclusions from these initial experiments:

1. A Pitch-Acoustics DBN, as in Figure 2, with four pitch prototypes and missing *Pitch* data, performs significantly better than the baseline Acoustics DBN.
2. A Pitch DBN, as in both Figures 2 & 3, performs significantly better with the *Pitch* variable having missing values than with its having observed values.
3. A Pitch DBN, as in both Figures 2 & 3, does not perform better than the baseline Acoustics DBN when the pitch data is observed.
4. A Pitch-Acoustics DBN, as in Figure 2, does not perform significantly better than a Phone-independent Pitch-Acoustics DBN, as in Figure 3.

Conclusion 1 does show that there is more research to be done to see if an even better Pitch DBN can be constructed; a significant improvement was achieved over the baseline Acoustics DBN.

From conclusions 2 & 3, we propose that our pitch estimator is not robust enough to be used directly in recognition. However, relevant information and statistical relationships can be extracted from its output during training. The DBN training indeed seems to have extracted relevant correlation information among the *Phone*, *Acoustics*, and *Pitch* variables. Furthermore, the *Pitch*'s temporal dependency may have helped to smooth and correct the output from the pitch estimator (the pitch estimator does not use temporal information). So, in recognition it is better to use the statistical properties learned from training instead of the observations from the pitch estimator; the DBN can then infer the distribution of the hidden pitch variables given the observed acoustics. Furthermore, more study is needed to see if perhaps the resulting DBN would be a more accurate pitch estimator than the one we used.

References

[1] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-39, no. 4, pp. 806–814, April 1991.

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

[3] Thomas Dean and Keiji Kanazawa, "Probabilistic temporal reasoning," in *Proceedings of the Seventh National Conference on Artificial Intelligence*, 1988, pp. 524–528.

[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 *ICSLP 2000*, October 2000, vol. II, pp. 951–954.

[5] W. Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, 1983.

[6] Bishnu S. Atal, "Automatic speaker recognition based on pitch contours," *Journal of the Acoustical Society of America*, vol. 52, no. 6, pp. 1687–1697, 1972.

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

[8] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc., San Francisco, CA, revised second printing edition, 1988.

[9] 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 *ICASSP 95*, May 1995.

[10] 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 *ICASSP 97*, 1997, pp. 1767–1770.

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