A NEW APPROACH FOR MULTI-BAND SPEECH RECOGNITION BASED ON PROBABILISTIC GRAPHICAL MODELS

Khalid Daoudi, Dominique Fohr and Christophe Antoine

INRIA-LORIA (Speech Group)
B.P. 101 - 54602 Villers les Nancy. France.
e-mails: daoudi,fohr,antoinec@loria.fr

ABSTRACT

In this paper, we introduce a new approach for multi-band speech recognition which allows interaction between sub-bands and does not require a recombination step. Moreover, this approach is a natural generalization of the HMMs paradigm and leads to fast learning and recognition algorithms.

1. INTRODUCTION

In the standard multi-band approach, the frequency axis is divided into several sub-bands, then each sub-band is independently modeled by a hidden markov model (HMM). The recognition scores in the sub-bands are then fusioned with some recombination module. While the ideas leading to the multi-band approach are attractive (analogy with human perception in particular), the latter has many drawbacks. For instance, the sub-bands are assumed mutually independent which is an unrealistic hypothesis. Moreover, the information contained in one sub-band is not discriminative in general. This can make the recombination a difficult task.

To overcome these limitations, we propose in this paper to interpret each HMM as a *probabilistic graphical model* (PGM) and build a more complex but uniform PGM on the time-frequency domain by "coupling" all the HMMs. We now give a brief introduction to PGMs and refer the reader to [1] for a detailed lecture. The PGM formalism consists in associating a graph to the joint probability distribution (JPD) $P(X)$ of a set of random variables $X = \{X_1, ..., X_N\}$. The nodes of this graph represent the random variables, while the edges encode the conditional independencies which (are supposed to) exist in the JPD. Using the relationships between the graphical structure and the conditional independencies, it is then possible to specify fast and exact algorithms to perform probability calculation. The conditional independence semantics, or the Markov property, of PGMs depend on the nature of the graph (directed, undirected or chain graph). In this paper, we consider directed PGMs (DPGMs) defined on directed acyclic graphs, also known as Bayesian networks or belief networks. In this case, the Markov property states that, conditioned on the configurations of its parents, a variable is independent of all the other variables except its descendants. A basic property of DPGMs is that the JPD can be factorized as[1]

$$P(x) = \prod_{n=1}^{N} P(x_n | pa(x_n)) \quad (1)$$

where $pa(x_n)$ denotes a configuration of the parents of X_n. From this point of view, it is easy to represent a HMM as a DPGM shown in Figure 1. In this figure, each node represents a random variable Q_t or O_t whose configuration specifies the state or the observation at time t. The arrows indicate the "causal influences" between variables. We shall use this representation of HMMs to build our multi-band DPGM.

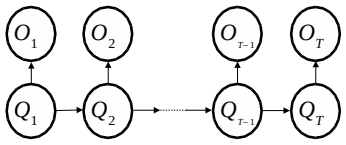

Figure 1: a HMM represented as DPGM

The basic idea behind our multi-band approach is the following: instead of considering an independent HMM for each sub-band, we couple all the HMMs by adding directed links between the variables and, then, use the computational algorithms associated with DPGMs to perform efficient learning and recognition. Our primary goal in this paper is to introduce these algorithms and provide a preliminary study of the potential of DPGMs in multi-band speech recognition. Thus, in order to make clear the different steps and given the lack of space, we will concentrate through this paper on the case of 2 sub-bands. Our 2-band DPGM is defined as follows. We link the hidden variables of sub-band 1 to those of sub-band 2 in such way that the configuration of a hidden variable in sub-band 2 at time t is conditioned by the configuration of two hidden variables: at time $t-1$ in the same sub-band and at time t in sub-band 1. Figure 2 shows the resulting DPGM. Each $Q_t^{(l)}$ is a discrete variable taking its values in $\{1, ..., m\}$. Each $O_t^{(l)}$ is a continuous variable with a Gaussian distribution representing the observation vector at time t in sub-band $l (l = 1, 2.)$.

[1] In the whole paper, upper-case (resp. lower-case) letters are used for random variables (resp. configurations).

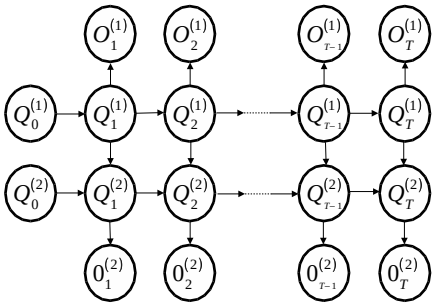

Figure 2: 2-band DPGM

We impose a left-to-right topology on each sub-band and assume that the hidden process is stationary. The numerical parameterization of our model is then the following: $a_{ij} \triangleq P(Q_t^{(1)} = j | Q_t^{(1)} = i)$; $u_{ijk} \triangleq P(Q_t^{(2)} = k | Q_t^{(1)} = i, Q_{t-1}^{(2)} = j)$; $b_i^{(l)}(o_t^{(l)}) \triangleq P(O_t^{(l)} = o_t^{(l)} | Q_t^{(l)} = i)$ where $b_i^{(l)}$ is a Gaussian with mean $\mu_i^{(l)}$ and covariance $\Sigma_i^{(l)}$.

Contrarily to the standard multi-band approach, our DPGM allows interaction between sub-bands and the possible asynchrony between them is taken into account. Moreover, our model uses the information contained in all sub-bands and no recombination step is needed. A related work has been proposed in [2] where a multi-band Markov random field is analyzed by mean of Gibbs distributions. This approach (contrarily to ours) does not lead however to exact and fast inference algorithms and assumes a linear model for asynchrony between sub-bands. In our approach, the asynchrony is learned from data. Moreover, the PGM framework is a natural generalization of the HMMs paradigm in the sense that all the computational aspects of HMMs are exactly recovered in the full-band case.

The paper is organized as follows. In the next section we present an inference algorithm for discrete DPGMs which will enables us to compute likelihoods and posterior probabilities efficiently and exactly. In section 3, we present two different but equivalent algorithms to perform learning. In section 4, we evaluate the performance of our approach by comparing it to HMMs and other systems on an isolated digit recognition task.

2. INFERENCE ALGORITHM FOR DPGM

In the last decade, major progress has been accomplished in the theory of PGMs. In particular, fast and exact inference algorithms has been developed when all the variables are discrete, all Gaussian or mixed discrete-Gaussian. In this section, we present the JLO algorithm [3] which applies to discrete DPGMs even though we are, in principle, in the mixed discrete-Gaussian case. The reason is that, as can be expected, all the quantities we will need to compute involve a complete set of observations of all the continuous variables (the $O_t^{(l)}$). Therefore the discrete case will be sufficient in our setting.

The JLO algorithm proceeds in two steps. The first one consists in using graph-theoretic tools to transform the initial graphical structure of the PGM into a specific graphical entity called the *junction tree*. In the second step, the junction tree is used as a channel to transmit and propagate the effect of observations (or evidence).

2.1. Construction of the junction tree

The first operation in the construction of the junction tree for DPGMs is the *moralization*. It consists in adding an extra undirected edge between any two nodes with a common child and subsequently removing directions. The undirected graph obtained this way is called the *moral* graph. The second operation consists in adding sufficient edges to the moral graph to make it *triangulated*. An undirected graph is triangulated (or chordal) if all cycles containing four or more nodes have a chord, i.e., an undirected edge between two non-consecutive nodes in the cycle. In the final operation, one identifies the set $\mathcal{C}$ of cliques[2] in the triangulated graph and forms a tree with these cliques in such way that resulting tree, the junction tree, satisfies the *running intersection property*. This property states that each variable which appears in two different cliques has to appear in all the cliques on the path between these two cliques. Attached with each edge linking two cliques C_1 and C_2 in the junction tree is a *separator* $S \triangleq C_1 \cap C_2$. We denote the set of separators by $\mathcal{S}$. Figure 3 shows the junction tree of our 2-band DPGM.

2.2. Propagation of evidence in the junction tree

Given the junction tree, the JPD $P(X)$ can be factorized as

$$P(x) = \frac{\prod_{C \in \mathcal{C}} \phi_C(x_C)}{\prod_{S \in \mathcal{S}} \phi_S(x_S)} \qquad (2)$$

where $\phi_C(x_C)$ (resp. $\phi_S(x_S)$) is a non-negative potential function on the clique $\mathcal{C}$ (resp. the separator $\mathcal{S}$). The collection of potentials $\Phi = \{\{\phi_C, C \in \mathcal{C}\}, \{\phi_S, \mathcal{S} \in \mathcal{S}\}\}$ is termed a *representation* of $P(X)$. A factorizable distribution $P(X)$ may have many different representations, i.e., many collections of potentials which satisfy (2). For DPGMs, an *initial* representation is obtained from (1) in the following way. First, assign each X_i to just one clique. Second, for each clique C, define the potential ϕ_C to be either the product of $P(X_i|pa(X_i))$ over all X_i assigned to C, or 1 if no variable is assigned to C. Then, if ϕ_S is set to be 1 for each separator S, one obtains a representation of $P(X)$.

To propagate the effect of an observed evidence e, the JLO algorithm operates by transforming one representation to another, starting from the initial one modified by the incorporation of the evidence. The algorithm finishes with the *marginal* representation in which, for each clique C (resp. separator S), the potential ϕ_C (resp. ϕ_S) is equal to the marginal (joint) probability distribution for the variables in C (resp. S) and the evidence. The incorporation of evidence in the initial representation is

[2] A clique is a subset of nodes which are fully connected and maximal, i.e, if a node is added to the subset, the latter does not remain fully connected.

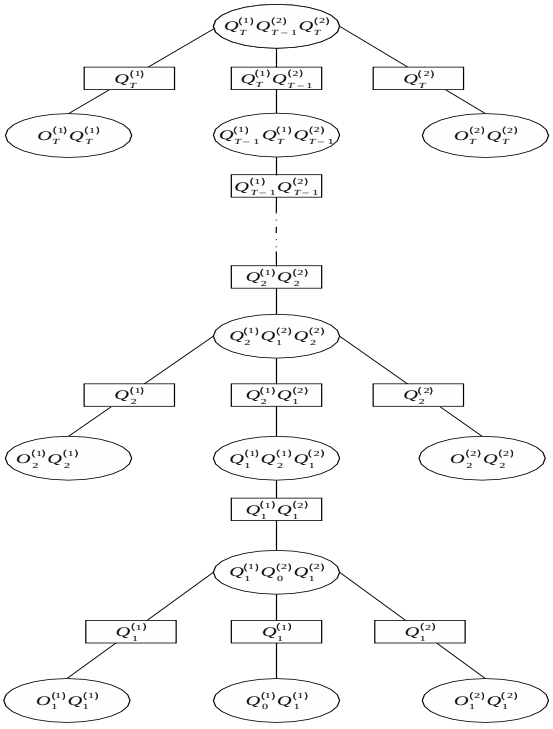

Figure 3: Junction tree of the 2-band DPGM. Cliques and separators are respectively represented by ellipsoids and rectangles.

done simply by setting $\phi_C(x_C)$ to 0 for any clique C containing an observed variable and for any configuration x_c involving a different state of the one observed. After this incorporation, the algorithm proceeds by passing a sequence of *flows* along the edges of the junction tree. Each flow from clique C_1 to C_2 updates the potentials of C_2 and the separator $S = C_1 \cap C_2$ in the following manner. Suppose that, prior to this flow, we have a representation Φ. Then, the activation of the flow yields a new representation Φ^* where the new potentials of C and S are[3]

$$\phi^*_S = \sum_{C_1 \setminus S} \phi_{C_1} \quad ; \quad \phi^*_{C_2} = \frac{\phi^*_S}{\phi_S} \phi_{C_2}$$

and all the other potentials being unchanged. A *schedule* of such flows consists in updating all the cliques using the the available information. This is done by choosing a clique C_r to be the *root* clique and, then, operating a recursive two-phase propagation scheme: *collecting* evidence and *distributing* evidence. In the collection phase, flows are activated along all the edges of the junction tree toward C_r. In the distribution phase, flows are activated out from C_r in the reverse direction. Once a schedule is complete, one obtains a new (final) representation Φ^f in which the potential ϕ^f_C of each clique C equals $P(x^h_C, e)$ (where x^h_C is a configuration of the hidden variables in C). Therefore, by marginalizing over the unobserved variables in any clique, on gets the likelihood of observations $P(e)$. Also, by normalizing the potential at a clique C to sum 1, one get the posterior conditional probability $P(x^h_C|e)$ of the hidden variables in C given the evidence e. The complexity of the JLO algorithm scales as the sum of the clique state-spaces. Thus, the complexity for inferring our 2-band DPGM is $\mathcal{O}(m^3 T)$.

In our experiments, we will consider an isolated digit recognition application. Therefore, we only need to compute likelihoods to perform recognition, thus the JLO algorithm will be sufficient for this purpose. We point out however that a modified version of this algorithm allows to identify, with the same time complexity, the most likely sequence of hidden states given observations [4]. This modified algorithm can be used in the case of continuous speech for instance. It is important to note that the Forward-Backward and Viterbi algorithms are recovered when applying these generic inference and identification algorithms to the particular case of HMMs [5]. This shows that the HMMs paradigm fits completely in the PGM framework.

3. MODEL PARAMETERS ESTIMATION

In the previous section, we showed how to perform inference for a DPGM when its parameters are known. In this section, we present an algorithm for learning the parameters our 2-band DPGM from available data (in the case of a single Gaussian per state). Suppose that we have an observation vector $O = (o^{(1)}_1, ..., o^{(1)}_T, o^{(2)}_1, ..., o^{(2)}_T)$. Following the same procedure as in the standard Baum-Welch algorithm for HMMs, we obtain the re-estimation formulae as follows. Suppose that we have estimated the parameters at iteration n. Define $\xi_t(i,j) \triangleq P(Q^{(1)}_{t-1} = i, Q^{(1)}_t = j|O)$; $\psi_t(i,j,k) \triangleq P(Q^{(1)}_t = i, Q^{(2)}_{t-1} = j, Q^{(2)}_t = k|O)$; $\xi(i,j) \triangleq \sum_{t=1}^{T} \xi_t(i,j)$; $\psi(i,j,k) \triangleq \sum_{t=1}^{T} \psi_t(i,j,k)$; $\gamma^{(1)}_t(j) \triangleq \sum_{i=1}^{m} \xi_t(i,j)$; $\gamma^{(2)}_t(k) \triangleq \sum_{i,j=1}^{m} \psi_t(i,j,k)$. Then, the new parameters at iteration $n+1$ are given by[4]

$$a_{ij} = \frac{\xi(i,j)}{\sum_j \xi(i,j)} \; ; \; u_{ijk} = \frac{\psi(i,j,k)}{\sum_k \psi(i,j,k)}$$

$$\mu^{(l)}_i = \frac{\sum_{t=1}^{T} \gamma^{(l)}_t(i) b^{(l)}_i(o^{(l)}_t)}{\sum_{t=1}^{T} \gamma^{(l)}_t(i)} \; ;$$

$$\Sigma^{(l)}_i = \frac{\sum_{t=1}^{T} \gamma^{(l)}_t(i)(o^{(l)}_t - \mu^{(l)}_i)(o^{(l)}_t - \mu^{(l)}_i)'}{\sum_{t=1}^{T} \gamma^{(l)}_t(i)}.$$

The remaining question is how to compute efficiently $\xi_t(i,j)$ and $\psi_t(i,j,k)$. Notice that, at each time t, the two subsets of hidden variables involved in these posterior conditional probabilities are both included in some

[3] The summation $\sum_{C_1 \setminus S}$ is over the state-space of variables that are in C_1 but not in S.

[4] For sake of notational simplicity, we drop the iteration index.

cliques of the junction tree. Therefore, these quantities can be efficiently computed using the JLO algorithm.

We now give an "alternative" way to compute ξ_t and ψ_t (and also likelihoods) which is analogous to the one used for HMMs. More precisely, we define (for our 2-band DPGM) new Forward and Backward coefficients and give the formulae to compute ξ_t and ψ_t using these coefficients. Therefore, if an HMM-based system is available, one can easily modify it to implement our approach. We emphasize however that this "alternative" way is just a re-writing of the JLO algorithm (which applies to any DPGM) when applied to our 2-band DPGM.

Let $o_t^{t'} \triangleq (o_t^{(1)}, ..., o_{t'}^{(1)}, o_t^{(2)}, ..., o_{t'}^{(2)})$. Define the Forward and Backward coefficients to be respectively

$$\alpha_t(i,k) \triangleq P(o_1^t, Q_t^{(1)} = i, Q_t^{(2)} = k)$$

$$\beta_t(i,k) \triangleq P(o_{t+1}^T | Q_t^{(1)} = i, Q_t^{(2)} = k),$$

where $\alpha_0(i,k) \triangleq \delta_1(i)\delta_1(k)$ and $\beta_T(i,k) \triangleq 1$. These coefficients can be recursively computed as follows:

$$\alpha_t(i,k) = b_i^{(1)}(o_t^{(1)}) b_k^{(2)}(o_t^{(2)}) \sum_j u_{ijk} \sum_n a_{ni} \alpha_{t-1}(n,j)$$

$$\beta_t(i,k) = \sum_n a_{in} b_n^{(1)}(o_{t+1}^{(1)}) \sum_m u_{nkm} b_m^{(2)}(o_{t+1}^{(2)}) \beta_{t+1}(n,m).$$

Then, ξ_t and ψ_t are given by

$$\xi_t(i,j) = \frac{a_{ij} b_j^{(1)}(o_t^{(1)}) \sum_k \beta_t(j,k) b_k^{(2)}(o_t^{(2)}) \sum_n u_{jnk} \alpha_{t-1}(i,n)}{P(O)}$$

$$\psi_t(i,j,k) = \frac{b_i^{(1)}(o_t^{(1)}) b_k^{(2)}(o_t^{(2)}) u_{ijk} \beta_t(i,k) \sum_n a_{ni} \alpha_{t-1}(n,j)}{P(O)}$$

where the likelihood $P(O) = \sum_{i,k} \alpha_T(i,k)$.

4. EXPERIMENTS

In this section, we evaluate the performance of our approach on an isolated digit recognition task and compare it to HMMs and 2 other models. Our experiments are carried out on the Tidigits database in which 112 (resp. 113) speakers are used for training (resp. test). Each speaker utters 11 digits twice. The parameterization for the classical full-band HMM is done as follows: 25ms frames with a frame shift of 10ms, each frame is passed through a set of 24 triangular filters resulting in a vector of 35 features, namely, 11 static MFCC (the energy is dropped), 12 Δ and 12 $\Delta\Delta$. The parameterization for our 2-band DPGM is done as follows: each frame is passed through the 15 first (resp. last 9) filters resulting in the acoustic vector of sub-band 1 (resp. sub-band 2). Each vector contains 17 features: 5 static MFCC, 6 Δ and 6 $\Delta\Delta$. The resulting bandwidths of sub-bands 1 and 2 are $[0..1777Hz]$ and $[1467Hz..10000Hz]$ respectively. As mentioned earlier, we also compare our approach to 2 other models. The first one is a "classical" 2-band model where the recombination is performed by a multi-layer perceptron (MLP). The topology of this MLP is: 22 input, 15 hidden and 11 output neurons. In the second one, for each frame, we concatenate the acoustic vectors of sub-band 1 and 2 and use the resulting vector (34 features) as input for the HMM-based system. The number of hidden states is six for every digit and all models ($m = 6$). The training of all models is done on clean speech only. The test however is performed on clean and noisy speech. The latter is obtained by adding, at different SNRs, a band-pass filtered white noise with a bandwidth of $[3000Hz..6000Hz]$. Figure 4 shows the recognition rates we obtain using the four models. For clean speech, the 2-band DPGM performs 3% better than HMMs and 2% worse than MLP. We point out however that the isolated speech task, plus the fact of 2 sub-bands only, is the most ideal case for MLP. For noisy speech, our model performs always better than the 3 others models. In conclusion, our 2-band DPGM models speech with higher fidelity than HMMs and is more robust to noise than all the other models. This shows that the PGM framework is very promising in the field of speech recognition and should be investigated further. This will be done in a future work.

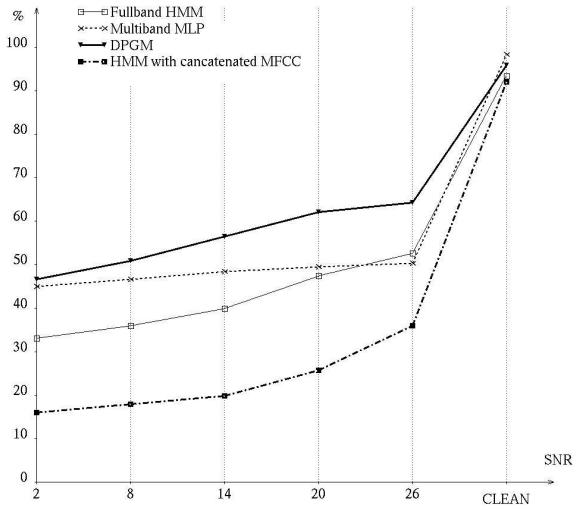

Figure 4: Recognition rates for clean and noisy speech

5. REFERENCES

[1] S.L. Lauritzen. *Graphical models.* Clarendon Press, 1996.

[2] G. Gravier, M. Sigelle, and G. Chollet. A markov random field based multi-band model. ICASSP'2000.

[3] F.V. Jensen and S.L. Lauritzen and K.G. Olsen . Bayesian updating in recursive graphical models by local computations. *Computational Statistics and Data Analysis*, (4):269–282, 1990.

[4] A.P. Dawid . Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, (2):25–36, 1992.

[5] P. Smyth and D. Heckerman and M.I. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9(2):227–269, 1997.