CONTINUOUS MULTI-BAND SPEECH RECOGNITION USING BAYESIAN NETWORKS

Khalid Daoudi, Dominique Fohr and Christophe Antoine

INRIA-LORIA (Speech Group, www.loria.fr/equipes/parole)
B.P. 101 - 54602 Villers les Nancy. France.

ABSTRACT

Using the Bayesian networks framework, we present a new multi-band approach for continuous speech recognition. This new approach has the advantage to overcome all the limitations of the standard multi-band techniques. Moreover, it leads to a higher fidelity speech modeling than HMMs. We provide a preliminary evaluation of the performance of our new approach on a connected digits recognition task.

1. INTRODUCTION

In standard multi-band (SMB) speech recognition, the frequency axis is divided into several sub-bands, then each sub-band is independently modeled by a HMM. The recognition scores in the sub-bands are then merged with some recombination module. The introduction of multi-band speech recognition [1, 2] has been essentially motivated by two desires. The first one is to mime the behavior of the auditive nerve which decomposes the speech signal into different sub-bands before recognition [3]. The second one is to improve the robustness to band-limited noise. While the ideas leading to multi-band speech recognition are attractive, the SMB approach 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. In addition, it is not easy to deal with asynchrony, particularly in continuous speech recognition. As a consequence, the recombination step can be a very difficult task.

In this paper, we present an alternative approach to perform continuous multi-band speech recognition which has the advantage to overcome *all* the limitations (mentioned above) of the SMB approach. Moreover, our experiments show clearly that asynchrony between sub-bands is extremely important in multi-band ASR. Furthermore, in clean conditions, we outperform HMMs without using the full-band parameterization as an additional "sub-band". In this sense, our approach can be also seen as a new way to model speech with higher fidelity than HMMs. In our opinion, this is mostly due to the fact that, contrarily to HMMs, we introduce a certain modeling of the frequency dynamics, namely, the asynchrony and dependency between sub-bands.

Our multi-band approach is based on the *Bayesian networks* (BNs) formalism. Using the interpretation of a HMM as a BN [4], we build a more complex but uniform BN on the time-frequency domain by "coupling" all the HMMs associated with the different sub-bands. In [4] we have presented a preliminary study of the same methodology, however only the case of 2 sub-bands and isolated speech recognition has been considered. In the present work, we provide the extension to *continuous* speech recognition with an *arbitrary* number of sub-bands.

The use of BNs in speech recognition has also been recently investigated in [5], but multi-band modeling is not addressed in that work. Briefly, the BNs formalism consists in associating a directed acyclic 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 arrows encode the conditional independencies (CI) which (are supposed to) exist in the JPD. The set of all CI relations, which are implied by the separation properties of the graph, are termed the *Markov properties*. The latter can be read as follows: conditioned on its parents, a variable is independent of all the other variables except its descendants. Given the graph structure, a BN is completely defined by the conditional probabilities of the variables given their parents. Indeed, the JPD can be expressed in a factored way as[1] $P(x) = \prod_{i=1}^{N} P(x_i|pa(x_i))$, where $pa(x_i)$ denotes an outcome of the parents of X_i.

In the next section, we define our multi-band model. In section 3 and 4, we present the decoding and learning algorithms. In section 5, we evaluate the performance of our model on a connected digits recognition task.

2. DEFINITION OF OUR MULTI-BAND MODEL

Let us assume that we are given a vocabulary V of $|V|$ words. For each word $v \in V$, instead of considering an independent HMM for each sub-band (as in SMB), we couple all the HMMs by adding directed links between the variables in order to capture the dependency between sub-bands. A natural question is: what are the "appropriate" links to

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

add? Probably the best answer is to learn the graphical structure (i.e., the dependencies between variables) from data. However, this strategy (which is extremely interesting and which we are currently investigating) is beyond the scope of this paper. Instead, we impose a graphical structure (for all words) which is motivated by the following criteria. We want a model where no continuous variable has discrete children in order to apply an exact inference algorithm. We also want a model with a small number of parameters and for which the inference algorithm is tractable. Finally, we want to have links between the hidden variables along the frequency axis in order to capture the asynchrony between sub-bands. A simple model which satisfies these criteria is the one shown in Figure 1. In this BN, the hidden variables of sub-band n are linked to those of sub-band $n+1$ in such way that the state of a hidden variable in sub-band $n+1$ at time t is conditioned by the state of two hidden variables: at time $t-1$ in the same sub-band and at time t in sub-band n. Each $Q_t^{(n)}\left(= Q_t^{(n)}(v)\right)$ is a discrete variable taking its values in the set of ordered labels $I_v = \{1_v, ..., m_v\}$. Each $O_t^{(n)}\left(= O_t^{(n)}(v)\right)$ is a continuous variable with a Gaussian distribution representing the observation vector at time t in sub-band n $(n = 1, ..., B)$, B is the number of sub-bands. We impose a left-to-right topology on each sub-band and assume that the hidden process is stationary. Therefore, given a word $v \in V$ and for each $(i, j, k) \in I_v^3$, the numerical parameterization of our model is: $a_{ij}(v) \triangleq P(Q_t^{(1)}(v) = j | Q_{t-1}^{(1)}(v) = i)$; $u_{ijk}^{(n)}(v) \triangleq P(Q_t^{(n)}(v) = k | Q_t^{(n-1)}(v) = i, Q_{t-1}^{(n)}(v) = j)$; $b_{i,v}^{(n)}(\cdot) \triangleq P(O_t^{(n)}(v) = \cdot | Q_t^{(n)}(v) = i)$;
where $b_{i,v}^{(n)}$ is a Gaussian with mean $\mu_i^{(n)}(v)$ and covariance $\Sigma_i^{(n)}(v)$. The asynchrony between sub-bands is taken into account by allowing all the $u_{ijk}^{(n)}(v)$ to be non-zero, except when $k < j$ or $k > j+1$ because of the left-to-right topology.

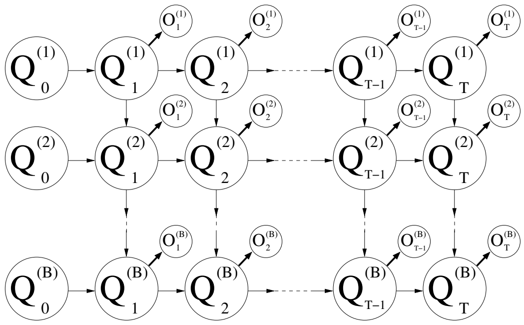

Fig. 1. B-band Bayesian network

Contrarily to HMMs, our BN provides "a" modeling of the frequency dynamics of speech. Contrarily to SMB, our BN 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 [6] 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 nor fast inference algorithms and assumes a linear model for asynchrony between sub-bands. In our approach, the asynchrony is learned from data.

3. CONTINUOUS SPEECH DECODING ALGORITHM

Given a B-band model of each word in the vocabulary and a speaker utterance, we need to identify the most likely sequence of words given the observation. A naive solution would be to use a B-dimensional Viterbi algorithm which is computationally very expensive. In this section, we present an efficient decoding algorithm which relies essentially on a state-augmented B-band model. The basic idea is to build a "super" B-band model which represents all the words in the vocabulary. Precisely, each variable $Q_t^{(n)}$ in Figure 1 takes now its values in the set $I = \bigcup_{v \in V} I_v$. To define this "super" model we need to specify the conditional probabilities of the hidden and the observed variables. The latter are simply given by those corresponding to each word, namely: $P(O_t^{(n)} = \cdot | Q_t^{(n)} = i_v) \triangleq b_{i_v,v}^{(n)}(\cdot)$, for each $i_v \in I$.
To specify the former, we need to include the language model, we also make some (a)synchrony assumptions. We still allow complete asynchrony *inside* a word, but we impose a full synchrony of all sub-bands when *transiting* between words. Precisely, since we have a left-to-right topology, the only non-zero conditional probabilities are the following:

• The synchronous transition between two (not necessarily different) words v and v':
$P(Q_t^{(1)} = 1_v | Q_{t-1}^{(1)} = m_{v'}) \triangleq P(v|v')$,
$P(Q_t^{(n)} = 1_v | Q_t^{(n-1)} = 1_v, Q_{t-1}^{(n)} = m_{v'}) \triangleq P(v|v')$,
where $P(v|v')$ is given by the language model.

• The inside-word conditional probabilities:
$P(Q_t^{(1)} = j_v | Q_{t-1}^{(1)} = i_v) \triangleq a_{ij}(v)$,
$P(Q_t^{(n)} = k_v | Q_t^{(n-1)} = i_v, Q_{t-1}^{(n)} = j_v) \triangleq u_{ijk}^{(n)}(v)$.

Now we have a completely defined B-band model on which we can perform decoding. To do so, we use the Dawid algorithm [7] which allows the identification (with the same time complexity as the JLO algorithm [8]) of the most likely sequence of hidden states given observations. The Dawid algorithm proceeds in two steps (as JLO). The first one (which is exactly the same as in JLO) consists in using graph-theoretic tools (moralization and triangulation) to transform the initial graphical structure of the BN into a specific graphical entity called the *junction tree*. We recall that the junction tree is a tree where the nodes are *cliques* and *separators*. The cliques are clusters of variables. A separator is simply the intersection between two adjacent cliques. The second step (the message propagation scheme) differs from the JLO one in potential updating and in the distribution phase. We refer the reader to [7] for details.

The construction of the junction tree is of particular interest to us because the decoding efficiency requires a junction trees with "small" clique state-spaces. In the 2-band case [4], finding a minimal junction tree is obvious because the moral graph is triangulated as it is. This is not true any more when $B > 2$. Since the problem of automatically finding minimal junction trees (for arbitrary BNs) is NP-hard, we need to find an appropriate (analytical) technique to derive a minimal junction tree for our particular B-band BN. We do this by induction, the resulting junction tree is shown in Figure 2. We thus have a computationally optimal tree to propagate observations.

The complexity of the Dawid algorithm scales as the sum of clique state-spaces. Therefore, given the (a)synchrony assumptions, the left-to-right topology and our junction tree, the total complexity[2] of our decoding algorithm is $O(m^B T + |V|^2 T)$.

4. MODEL PARAMETERS ESTIMATION

So far, we have assumed that parameters of the B-band BN are known for each word. In this section, we present the algorithm of parameters estimation (in the case of a single Gaussian per state). In the experiments we carry out, we learn the model of each word independently of the others (i.e., we do not perform embedded training). Thus, in order to simplify the notation in the formulae below, we drop the reference to the word under consideration. Thus, all the quantities below correspond to the notations of Section 2 (not Section 3). Suppose that we have (for a given word v) an observation vector $o = (o_1^{(1)}, ..., o_T^{(1)}, ..., o_1^{(B)}, ..., o_T^{(B)})$. Using the EM algorithm, we obtain the re-estimation formulae as follows. Suppose that we have estimated the parameters at iteration l. Define $\psi_t^{(1)}(i,j) \triangleq P(Q_{t-1}^{(1)} = i, Q_t^{(1)} = j|o)$; $\psi_t^{(n)}(i,j,k) \triangleq P(Q_t^{(n-1)} = i, Q_{t-1}^{(n)} = j, Q_t^{(n)} = k|o)$;
$\psi^{(1)}(i,j) \triangleq \sum_{t=1}^{T} \psi_t^{(1)}(i,j)$; $\psi^{(n)}(i,j,k) \triangleq \sum_{t=1}^{T} \psi_t^{(n)}(i,j,k)$; $\gamma_t^{(1)}(j) \triangleq \sum_{i=1}^{m} \psi_t^{(1)}(i,j)$;
$\gamma_t^{(n)}(k) \triangleq \sum_{i,j=1}^{m} \psi_t^{(n)}(i,j,k)$. Then, the new parameters at iteration $l+1$ are given by[3]

$$a_{ij} = \frac{\psi^{(1)}(i,j)}{\sum_j \psi^{(1)}(i,j)} \; ; \; u_{ijk}^{(n)} = \frac{\psi^{(n)}(i,j,k)}{\sum_k \psi^{(n)}(i,j,k)} \; ;$$

[2] If the number of hidden states is the same for all words and equals some integer m ($card\,(I_v) = m, \forall v$).

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

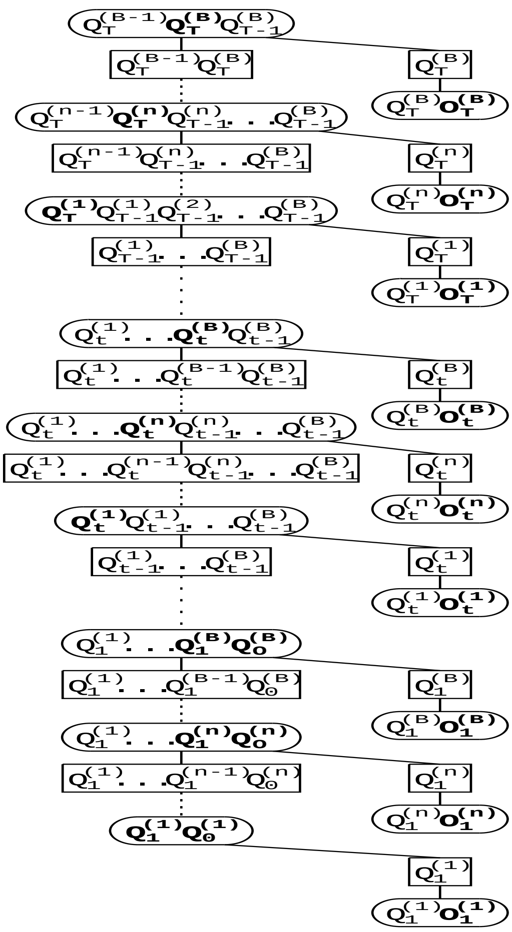

Fig. 2. Junction tree of the B-band Bayesian network. Cliques and separators are respectively represented by ellipsoids and rectangles.

$$\mu_i^{(n)} = \frac{\sum_{t=1}^{T} \gamma_t^{(n)}(i) o_t^{(n)}}{\sum_{t=1}^{T} \gamma_t^{(n)}(i)} \; ;$$

$$\Sigma_i^{(n)} = \frac{\sum_{t=1}^{T} \gamma_t^{(n)}(i)(o_t^{(n)} - \mu_i^{(n)})(o_t^{(n)} - \mu_i^{(n)})^*}{\sum_{t=1}^{T} \gamma_t^{(n)}(i)}.$$

Notice that, for each time t and each sub-band n, the subset of hidden variables involved in the posterior probability $\psi_t^{(n)}(i,j,k)$ is included in some clique of the junction tree. Therefore, all the quantities above can be efficiently computed using the JLO algorithm [8] which allows the computation of marginal and conditional probabilities of clique variables.

5. EXPERIMENTS

In this section, we evaluate the performance of our B-band BN on a connected digits recognition task. Our experiments are carried out on the Tidigits database. In learning we only use the isolated part of the training database where each speaker utters 11 digits twice. Also, we do not remove the initial and final pauses, thus we do not have a silence model. In test, we use the full (test) database in which 8636 sentences are uttered, each sentence contains between 1 and 7 digits.

We compare[4] the performances of our B-band BN to those of 2 models: HMMs and "synchronous" BNs. In all the experiments, for every digit and all models, the number of hidden states is six ($m = 6$) and we have a single Gaussian per state with a diagonal covariance matrix. We use a uniform language model, i.e., $P(v|v') = \frac{1}{11}$ $(|V| = 11)$.

The parameterization of 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 $\Delta\Delta$. For our B-band model, we present experiments for $B = 2, 3$. The parameterization of the 2-band BN is done as follows: each frame is passed through the 14 first (resp. last 10) 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 $[0..1467Hz]$ and $[1211Hz..10000Hz]$ respectively. For the 3-band BN, each frame is passed through the first 8, second 8 and last 8 filters resulting in the acoustic vector of sub-band 1, 2 and 3 respectively. Each vector contains 11 features: 3 static MFCC, 4 Δ and 4 $\Delta\Delta$. The resulting bandwidths of sub-bands 1, 2 and 3 are $[0..692Hz]$, $[615Hz..2152Hz]$ and $[1777Hz..10000Hz]$ respectively. The parameterization of the third model is done as follows: for each frame, we concatenate the acoustic vectors of sub-band 1 and 2 (resp. 1, 2 and 3) and use the resulting vector of 34 (resp. 33) features as an input for the HMM-based system. We refer to these as Sync2b and Sync3b respectively. The behavior of this third model is very interesting to analyze. This is because it is exactly equivalent to a B-band BN ($B = 2, 3$) where a complete inside-word (or frame) synchrony between the sub-bands is imposed (since we use diagonal covariances). Therefore, the comparison between Sync2b (resp. Sync3b) and our 2-band (resp. 3-band) BN is a good indication about the importance of asynchrony.

Table 1 shows the word accuracy scores obtained using the 3 models. The comparison between our B-band BN and the frame-synchronous one indicates clearly that asynchrony between sub-bands is very important in the modeling. Regarding the comparison between our B-band BN and HMMs, let us point out that (to the best of our knowledge) the only multi-band systems which out-perform HMMs in clean conditions use the full-band parameterization as an additional "sub-band". Our multi-band system do not use such (conceptually disturbing) procedure, still it does considerably outperform HMMs. This also shows that, indeed, taking into account some of the frequency dynamics (asynchrony and dependency between sub-bands) leads to a higher fidelity speech modeling than HMMs.

Model	HMM	Sync2b	Sync3b	2-band	3-band
Score	61.4%	60.5%	55.1%	**73.0**%	**68.9**%

Table 1. Word accuracy scores on clean speech.

[4]Very good scores can be obtained on this database using multi-Gaussians HMMs and adjusted parameters. Our goal here is not to tune the parameters in order to achieve the highest performances. Rather, we want to provide a fair comparison using a baseline system for all the models we consider. We believe that this way we have a fair initial judgment on the capacities of each system.

6. CONCLUSION

Using the Bayesian networks framework, we developed a new multi-band approach for continuous speech recognition. We carried out preliminary experiments in clean speech conditions. Our system does not only outperform the standard multi-band ones, but also outperforms the HMM-based one. This shows that our approach is very promising in the field speech recognition. Although, in the time of writing, we do not have results in noisy speech conditions, we are very confident on this matter. Indeed, in our previous work [4] on isolated speech recognition, tests in noisy conditions has been conducted and the results were very promising. We expect the same behavior to hold in the continuous setting.

7. REFERENCES

[1] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. ICSLP'96.

[2] H. Hermansky et al. Towards ASR on partially corrupted speech. ICSLP'96.

[3] B. C. J. Moore. *An introduction to the psychology of hearing*. Academic Press, London, 1989.

[4] K. Daoudi, D. Fohr, and C. Antoine. A new approach for multi-band speech recognition based on probabilistic graphical models. ICSLP'2000.

[5] G. Zweig. Speech recognition with dynamic Bayesian networks. *PhD thesis, Univ. California, Berkeley*, 1998.

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

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

[8] 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.