SPEECH MODELING WITH STATE CONSTRAINED MARKOV FIELDS OVER FREQUENCY BANDS

V. Arsigny *G. Chollet* *G. Gravier** *M. Sigelle*

ENST-TSI and CNRS-URA 820
46, rue Barrault F-75634 Paris Cedex 13
arsigny@polytechnique.org {gravier,sigelle,chollet}@tsi.enst.fr

ABSTRACT

In this paper, we propose a model of speech segments in the time/frequency domain. This model is based on Markov random field (MRF) modeling and is an extension of our previous work on multi-band models with MRF. In this new approach, the time model in each band is defined on the constrained state space of strictly left-right Markov chains and a non-stationary synchrony model between the frequency band is added. We derive algorithms for parameter estimation and for segment scoring. Finally the model is tested and compared to HMM on an isolated word recognition task. Results show the interest of the synchrony model for test data corrupted with additive white noise but they also point out some weaknesses of the training algorithm which must be improved.

1 INTRODUCTION

Hidden Markov models (HMM) are widely used for speech modeling but are known to suffer some limitations, specially when the signal is corrupted by noise. A noise robust approach to speech modeling using the multi-band paradigm has been proposed in [1] and many studies have been carried out on this model. However very few models for the time/frequency (t/f) distribution of speech have been studied. Compensation for additive noises can hardly be carried out in the cepstral domain but could be more easy in the spectral domain if we had efficient speech models in this domain. Moreover, for indexing and segmentation purposes, it may be important to have the same signal representation for music and for speech. In this case, the t/f representation is suited for music and speech models in the t/f domain are necessary. These are the two main motivations for developing models in the time/frequency domain.

In a previous work, we studied a Markov random field based approach [1, 2] to speech modeling in the t/f domain. This approach was based on an extension of the multi-band paradigm where interactions between the Markov chains in the sub-bands were added. In the time domain, a hidden Markov chain was implemented in each band using the Markov field equivalence defined in [3]. The interaction in the frequency domain consisted in controlling the synchrony (or asynchrony) between the sub-bands, considering two bands synchronous if the spectrally stable zones are observed at the same instants. Promising results were obtained for this model, specially in the case of noisy test data in a multi-band approach with cepstral representation in each band. However, this first model has two major drawbacks. First, it is computationally expensive and second, the synchrony model is stationary, which means that the synchronization coefficients do not depend upon time. This assumption may be true for short segments but is clearly unrealistic for longer segments.

In this paper, we present a new Markov random field based speech segment model were modeling is done in a constrained state space to speed up computation. The constrained state space under consideration is the state space of strictly left to right Markov chains. This consideration allows some simplification in the representation of a field which is now specified solely by the variables D_i^k, the starting time of state i for the band k, the model being defined as a parametric Gibbs distribution (GM) of these variables. In addition to this, a state dependent synchrony model is added to the temporal model. The paper is organized as follows: sections 2 and 3 describes the model and the related parameter estimation procedures and results on isolated word recognition over the telephone are given section 4.

2 STATE CONSTRAINED MARKOV RANDOM FIELD MODEL

We follow here [4]. Let us first consider a N states Markov chain with strictly left-right topology (*i.e.* the only permitted transition are $i \rightarrow i$ with probability p_i and $i \rightarrow i+1$ with probability $1-p_i$). This chain is also submitted to the constraints that $x_1 = 1$ and $x_{T+1} = N+1$ where $N+1$

*G. Gravier is currently with IRISA/INRIA Rennes.

is an absorbing state. Such configurations are defined on a constrained state space denoted $\mathcal{E}^c$ and can be equivalently specified by one of the following variables:

- the state sequence $x = (x_t)_{t\in[1,T]}$ submitted to $x_1 = 1 \leq x_2 \ldots \leq x_{t-1} \leq x_t \ldots \leq x_{T+1} = N+1$
- the set of duration times (T_i) of each state i with $\forall i \in [1,N] \quad T_i \geq 1$ and $\sum_{i=1}^{N} T_i = T$
- the set of starting times (D_i) of each state i with $D_1 = 1$ and $D_{i-1} < D_i \ldots < D_{N+1} = T+1$.

The probability of a configuration $x \in \mathcal{E}^c$ is given by

$$P[X=x] = C \prod_{i=1}^{N} (1-p_i) p_i^{T_i - 1} = C^* \prod_{i=1}^{N} p_i^{T_i} \qquad (1)$$

where C and C^* are normalization constants. For the uniform distribution μ_0 on $\mathcal{E}^c$, *i.e.* when all the p_i are equal to some value p, this is written as

$$P[X=x] = C(1-p)^{N-1} p^{T-N} = \frac{1}{\text{Card } \mathcal{E}^c}$$

where Card $\mathcal{E}^c = C_{T-1}^{N-1}$ since it is the number of placements of $N-1$ different starting times in interval $[2,T]$.

We now consider the law of K independent Markov chains, in a multi-band framework. The process $x^k, k \in [1,K]$ is a MRF on the bidimensional lattice $S = [1,T] \times [1,K]$ and takes its values in configuration space $\mathcal{E}$. MRF have been widely used in Image Processing since pioneer work [5]. Given a neighborhood system (*i.e.* a set of neighborhood dependencies on S), the Hammersley-Clifford theorem states that every positive MRF is a Gibbs distribution $P[X=x] = Z^{-1} \exp -U(x)$, where the total energy function $U(x) = \sum_{c\in\mathcal{C}} U_c(x)$ is a sum of local clique energy functions, and the partition function is given by $Z = \sum_{x\in\mathcal{E}} \exp -U(x)$. It can easily be seen from (1) that for a K band Markov chain

$$P[X=x] = \prod_{k=1}^{K} P[X^k = x^k] = \frac{\exp -U_h(x)}{Z}$$

where the horizontal energy is given by

$$U_h(x) = \sum_{i=1,k=1}^{N,K} \alpha_i^k \; T_i^k = \sum_{i=2,k=1}^{N,K} A_i^k \; D_i^k = \bar{A}.\bar{D}$$

with $\bar{D} = (D_i^k)$, $\bar{A} = (A_i^k)$, $\alpha_i^k = -\ln p_i^k$ and $A_i^k = \alpha_{i-1}^k - \alpha_i^k$. The law of the hidden process is therefore defined as a MRF.

So far we have been considering independent Markov chains in each bands. We now introduce dependencies between the bands by adding a state-dependent synchronization model between two bands k and l. The model is related to the difference of starting times of state i in these two bands as $|D_i^k - D_i^l|^\gamma$. The total synchronization energy is given by

$$U_v = \sum_{i=1}^{N} \sum_{k,l=1}^{K} F_i^{k,l} |D_i^k - D_i^l|^\gamma = \bar{F}.\bar{\Psi}$$

where $\bar{F} = (F_i^{k,l})$ and $\bar{\Psi} = (\Psi_i^{k,l})$ with $\Psi_i^{k,l} = |D_i^k - D_i^l|^\gamma$. The constant γ is a positive exponent. The higher the $F_i^{k,l}$, the more stringent the synchronization condition between starting times D_i^k and D_i^l. Throughout the paper, we will consider $\gamma = 2$. We conclude that the *ad hoc* variables of this model are the (D_i^k).

The main differences with model proposed in [1] is thus the reduction of the MRF state space to the state space of strictly left-right Markov chains in each band while the synchronization model has been extended to be non stationary. The complete prior law of the SC-MRF model is given by

$$P[X=x] = P[\bar{D}] = \frac{\exp -\bar{A}.\bar{D} - \bar{F}.\bar{\Psi}}{Z_{\bar{A},\bar{F}}^{\text{prior}}} \qquad (2)$$

Assuming conditional independence of the observations and a Gaussian observation law, the conditional law of X can be written as

$$\begin{aligned} P[Y=y \,|\, X=x] &= \prod_{t,k} P[Y_{t,k} = y_{t,k} \,|\, X_{t,k} = x_{t,k}] \\ &= \exp -U_d(y\,|\,x) = \exp -U_d(y\,|\,\bar{D}) \end{aligned}$$

which can be expressed in a closed form since $P[Y_{t,k} \,|\, X_{t,k}] = \mathcal{N}(\mu_{x_{t,k}}, \sigma^2_{x_{t,k}})$ and because of the equivalent representation $\bar{D}$ of x in $\mathcal{E}^c$. From Bayes rule, the posterior law of the hidden process is given by

$$\begin{aligned} P[X=x \,|\, Y=y] &= \frac{P[Y=y \,|\, X=x] P[X=x]}{P[Y=y]} \\ &= \frac{\exp -U_d(y\,|\,\bar{D}) - \bar{A}.\bar{D} - \bar{F}.\bar{\Psi}}{Z^{\text{post}}_{\bar{\mu},\bar{\sigma},\bar{A},\bar{F}}} \qquad (3) \end{aligned}$$

Form (2) and (3) it can be seen that the prior and posterior laws of X are also Gibbs distributions. More generally, Gibbs distributions with inverse temperature $P_\beta[X=x] = Z_\beta^{-1} \exp -\beta U(x)$ are suited for simulation and optimization. Simulation at inverse temperature β is performed by the Metropolis dynamics: choose a site (t,k) at random and draw a candidate starting time in the range $[D_{i-1}^k + 1, D_{i+1}^k - 1]$. Denoting ΔU the energy variation corresponding to the candidate value, the value of $X_{t,k}$ is changed with probability 1 if $\Delta U < 0$ or with probability $\exp -\beta \; \Delta U$ otherwise. Simulated annealing (SA) enables to detect the global minimum of the energy function $U(x)$ as $\beta \to \infty$ [5].

3 PARAMETER ESTIMATION

Consider a word with overall parameters $\bar{\lambda} = (\bar{\mu}, \bar{\sigma})$ and $\bar{\theta} = (\bar{A}, \bar{F})$ to be estimated from M independent samples $y^m\{y_{t,k}^m\}$, each of them with duration T_m.

3.1 Maximum likelihood estimation

The likelihood of word w to be maximized *w.r.t.* parameters $(\bar{\lambda}, \bar{\theta})$ is defined by

$$\begin{aligned} P_w[Y=y] &= \sum_{x \in \mathcal{E}^c} P_w[Y=y|X=x] P_w[X=x] \\ &= \frac{Z^{\text{post}}_{\bar{\mu},\bar{\sigma},\bar{A},\bar{F}}}{Z^{\text{prior}}_{\bar{A},\bar{F}}} \end{aligned}$$

Deriving the logarithm of this expression *w.r.t.* each parameter gives for the prior parameters the following set of self-coherent equations

$$\frac{1}{M}\sum_{m=1}^{M} \mathbf{E}^{\text{prior}}_m[\bar{D}] = \frac{1}{M}\sum_{m=1}^{M} \mathbf{E}^{\text{post}}_m[\bar{D}] \quad (4)$$

$$\frac{1}{M}\sum_{m=1}^{M} \mathbf{E}^{\text{prior}}_m[\bar{\Psi}] = \frac{1}{M}\sum_{m=1}^{M} \mathbf{E}^{\text{post}}_m[\bar{\Psi}] \quad (5)$$

The operator $\mathbf{E}$ refers to the expectation, m is the training token index and superscripts *prior* and *post* refer to (2) and (3) respectively. For the observation parameters $\bar{\lambda}$, we have the classical EM formulae (see *e.g.* [6]). The MRF-EM algorithm consists in (*i*) computing, in the E step, the posterior expectations involved in (4)-(5) and $P[X_{t,k} = i|y_{t,k}]$, and (*ii*) solving *w.r.t.* the prior the left-member part of these equations. Several stochastic approximations must be performed to solve this set of equations. First, the current posterior probabilities and expectations cannot be computed exactly (as it is the case with the HMM) and are estimated empirically from samples drawn under the current posterior law (3). This enables the estimation of new values for $\bar{\lambda}$. Second, a stochastic gradient (SG) algorithm [7] is used to solve the prior parameters equations (4) and (5). Let us denote $(\theta^{(p)}, \lambda^{(p)})$ the parameters after p iterations of the EM algorithm. If $\theta^{(p)}_n$ are the parameters at iteration n of the SG inside iteration p of the EM, then the parameters are updated as follows, assuming diagonal approximation of the Hessian

$$A^k_i \leftarrow A^k_i + \frac{1}{n^\delta} \frac{\mathbf{E}^{\text{prior}}_{\theta^{(p)}_n,m}[D^k_i] - \mathbf{E}^{\text{post}}_{\theta^{(p)},\lambda^{(p)},m}[D^k_i]}{\sum_{m=1}^{M} \mathbf{V}^{\text{prior}}_{\theta^{(p)}_n,m}[D^k_i]} \quad (6)$$

$$F^{k,l}_i \leftarrow F^{k,l}_i + \frac{1}{n^\delta} \frac{\mathbf{E}^{\text{prior}}_{\theta^{(p)}_n,m}[\Psi^{k,l}_i] - \mathbf{E}^{\text{post}}_{\theta^{(p)},\lambda^{(p)},m}[\Psi^{k,l}_i]}{\sum_{m=1}^{M} \mathbf{V}^{\text{prior}}_{\theta^{(p)}_n,m}[D^k_i]} \quad (7)$$

with $0.5 < \delta \leq 1$.

3.1.1 Initialization

Of utmost importance is the initialization for this class of stochastic algorithms. We proceed as follows:

1. estimate empiric observation parameters $(\bar{\mu}, \bar{\sigma})$ using uniform *segmentation* in each frequency band. Re-estimate them using a segmentation obtained by SA for these values, assuming uniform *distribution* in each band and no synchronization $(\bar{A}_0 = \bar{0}, \bar{F}_0 = \bar{0})$.

2. estimate $(\bar{A}, \bar{F})$ with SG starting from initial uniform distribution in each independent band, *i.e.* $\bar{\theta}_0 = (\bar{0}, \bar{0})$. Equations (6) and (7) yield and exact value of parameter $\bar{\theta}_1$ when $\gamma = 2$ since the moments of the starting times variables D^k_i can be computed in a closed form for a uniform distribution μ_0 in each band [4]. An alternative is also to leave $\bar{F} = \bar{0}$.

4 ISOLATED WORD RECOGNITION

The SC-RFM is applied to single speaker isolated word recognition over the telephone. The vocabulary is consists of ten command words and 100 occurrences of each words are available. The first 50 occurrences are used for training while the remaining one are used for the tests. The results are therefore given for a total of 500 tests. The score of a test segment y for a given word w is the likelihood of the complete statistics, that is

$$P_w(x^*, y) = \frac{\exp\left(-U^{(w)}_d(y|x^*) - U^{(w)}_h(x^*) - U^{(w)}_v(x^*)\right)}{Z^{\text{prior}}_{\bar{A},\bar{F}}}$$

The hidden configuration x^* is obtained by simulated annealing to find out the most likely configuration using a MAP criterion. The complete statistic probability cannot be directly computed because of the partition function $Z^{\text{prior}}_{A,F}$ which must be approximated. In practice, the partition function is approximated by

$$\begin{aligned} Z^{\text{prior}}_{\bar{A},\bar{F}} &= Z^{\text{prior}}_{\bar{A}} \; \mathbf{E}^{\text{prior}}_{\bar{A}}[\exp -U^{(w)}_v(x)] \\ &\simeq Z^{\text{prior}}_{\bar{A}} \exp\left(-\bar{F} \; \mathbf{E}^{\text{prior}}_{\bar{A}}[\bar{\Psi}]\right) \end{aligned}$$

where $Z^{\text{prior}}_{\bar{A}}$ and $\mathbf{E}_{\bar{A}}$ refers to the GD defined by the horizontal potential function only. The partition function $Z^{\text{prior}}_{\bar{A}}$ can be explicitly calculated in some particular cases [4] but it is generally approximated by

$$Z^{\text{prior}}_{\bar{A}} \simeq Z_0 \exp\left(-\bar{A} \; \mathbf{E}^{\text{prior}}_0[\bar{D}]\right)$$

where Z^{prior}_0 and $\mathbf{E}^{\text{prior}}_0$ refer to the uniform law on $\mathcal{E}^c$.

In a first experiment, we compare different strategies for initializing the parameters. Results are given table 1. The first row corresponds to an initialization where only step 1 of the procedure proposed in section 3.1.1 is implemented. The second row is obtained by initializing the means and variances of the Gaussians and the intra-band parameters $\bar{A}$ according to (4). Finally, the third row corresponds to the complete procedure where either $\bar{A}$ and $\bar{F}$ are initialized. These results point out the importance of the initialization. The proposed initialization of the inter-band potential parameters $\bar{F}$ degrades the results compared to the case where these parameters are set to null. However, this gives good initial values for the EM algorithm. The initialization of the intra-band potential parameters seems valid and gives good results after the initialization step. It also can be seen that when the initialization is not done properly, the stochastic approximation of the EM algorithm may deteriorate the recognition rate.

algorithm	$\bar{A}=0, \bar{F}=0$	$\bar{F}=0$	SG
init.	94.2	96.4	74.6
init. + MLE	95.0	98.2	93.8

Table 1: Recognition rate for various parameter initialization strategies (vertical range = 2).

In a second experiment, we investigate the effect of the synchronization between the bands by varying the vertical neighborhood system. Let us define the vertical range (denoted v-range) as being the maximum gap between two synchronized bands. For example, if the v-range is two, band k will depend on band l if $|k-l| <= 2$, otherwise the band will be considered independent (*i.e.* $F_i^{k,l} = 0 \quad \forall i$). A v-range of 0 corresponds to parallel independent HMM while a v-range of 24 corresponds to the case where the synchrony model considers that one band may depend on all the other bands. State constrained random field models (SC-RFM) with various vertical range were tested on clean and noisy test data. Training is always performed on clean data and the test data were corrupted with additive white noise at various signal to noise ratios (SNR). Results are given table 2. The last row of the table correspond to results obtained with a standard single band HMM applied to the filter-bank output. It can be seen that for clean test data, the recognition rate is not improved by adding interactions between the bands and is comparable to the recognition rate obtained with a HMM. For noise corrupted test data at high SNR (30 and 20 dB), the recognition rate increases along with the vertical range. This shows the interest of adding vertical dependencies between the frequency bands in a t/f model. However, there is a drastic decrease of performances with SC-RFM when test data are corrupted by noise. In this case, a single band HMM performs significantly better. One explanation to the poor behavior of SC-RFM on noisy data is a problem of over-fitting in the parameter estimation procedure. In SC-RFM, it has been observed that for some words, the Gaussian variances can be either very small or very large. Some cases of large variances were also found with HMM but most of the variances are of the same magnitude order. This partially explains the recognition rate decrease when noise is added since words having small variances are no more recognized while the words having high variances *absorbs* all the decisions. This result shows that more work is to be done on the parameter estimation procedure.

	vertical range						
SNR	0	1	2	3	5	24	HMM
∞	98.2	98.2	98.2	98.2	98.2	98.2	98.8
30	68.8	69.2	69.6	70.0	70.0	73.0	96.8
20	16.2	16.2	16.2	16.6	16.6	17.0	73.2
10	10.0	10.0	10.0	10.0	10.0	10.0	17

Table 2: Recognition rate for different v-range as a function of the test data SNR.

5 DISCUSSION

We presented a new model of the time/frequency distribution of speech. The model is conveniently represented by a Markov Random Field on the state space of multi-band left-right Markov chains. A synchronization model between the Markov chains in each frequency band is added. This work is an extension of [1] to the case of non-stationary synchronization. The use of the constrained left-right state space was motivated to show the capabilities of MRF in term of computation time. Experiments on isolated word recognition showed the interest of the synchrony model for test data corrupted with white noise. However, it was observed that the main limitation was the parameter estimation procedure which sometimes leads to over-fitting problems. For instance, the burn-in simulation period before beginning of the stochastic gradient algorithm as well as the number of simulations needed to estimate the prior and posterior statistics must be properly set. This could result, in particular, in a better estimation of the Gaussian variances.

In the future, we plan to explore several extensions of the SC-RFM. First, we envisage to study other synchronization potentials based, for example, on recall measures. Indeed, the current potential function is based on the starting times D_i^k which are highly variable. Second, the formalism of MRF could be exploited to infer a neighborhood system or to model long range time interactions. Finally, this model could be used for speaker recognition or speech enhancement.

References

[1] G. Gravier, M. Sigelle, and G. Chollet. Toward Markov random field modeling of speech. In *Intl. Conf. Speech and Language Processing*, December 1998.

[2] G. Gravier, M. Sigelle, and G. Chollet. A Markov random field based multi-band model. In *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, 2000.

[3] Y. Zhao, L.A. Atlas, and X. Zhuang. Application of the Gibbs distribution to hidden Markov modeling in speaker independant isolated word recognition. *IEEE Trans. Signal Processing*, 39(6):1291–1298, 1991.

[4] V. Arsigny. Modélisation par champ de Markov du signal de parole et application à la reconnaissance vocale. Technical report, stage d'Option Scientifique - Polytechnique, 2000.

[5] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE trans. on PAMI*, 6(6):721–741, 1984.

[6] L. Rabiner. A tutorial to hidden Markov model etc.. *IEEE Trans. Signal Processing*, 0(0):0–0, 1985.

[7] Laurent Younes. Estimation and annealing for Gibbsian fields. *Ann. Inst. Henri Poincaré*, 24(2):269–294, 1988.