A MARKOV RANDOM FIELD BASED MULTI-BAND MODEL

Guillaume Gravier, Marc Sigelle and Gérard Chollet

ENST/TSI and CNRS-URA 820
46, rue Barrault
75634 Paris Cedex 13
France
gravier@tsi.enst.fr

ABSTRACT

An extension of the multi-band model including inter-band control of time asynchrony is described. It is based on the framework of Markov random fields. The law of the speech process is given by a parametric Gibbs distribution and a maximum likelihood parameter estimation algorithm is developed. This random field model is applied to isolated word recognition. It is shown that similar performances are obtained with the new model and with standard HMM techniques in the mono-band case. In the multi-band case, it is shown that the recognition rate decreases when the number of band is increased but that modeling inter-band synchrony limits the performance decrease.

1. INTRODUCTION

Hidden Markov models (HMM) are extensively used in speech recognition for the computation of the likelihood of an observation knowing a sequence of words (see e.g. [1], chap. 2). Good results have been achieved with this statistical approach but there are limitations to this model. In particular, HMMs are not robust to additive or convolutive noises and distortions such as reverberation and clipping. Cepstral mean subtraction or RASTA processing is usually used to compensate for slowly varying convolutive noise such as telephone line distortions, but other techniques must be used for additive noises. The robustness to noise can be increased by a more stable representation of the speech signal than the cepstral ones or by a more accurate statistical model of the signal. Many efforts have been done to find out speech representations that are less sensitive to noise (as *e.g.* the modulation spectrum, TRAPS, LDA based analysis,...). Recently a multi-band approach to speech recognition has been proposed to deal with additive noises [2].

In the multi-band approach, the signal is divided into sub-bands and the technique relies on independent modeling of each sub-band with HMMs. The partial sub-band scores are merged at some point in the decoding process. Regardless of the recombination stage, the model implements asynchronous modeling of the sub-bands and it was shown in [3] that performances can be increased by allowing asynchrony between the sub-band. However, the independence hypothesis of the sub-bands seems unrealistic and the sub-bands are neither totally asynchronous nor synchronous. Moreover, part of the asynchrony may be due to the transmission channel and is irrelevant for speech recognition. Therefore, it may be interesting to add some interaction between the bands and to model the spectral synchrony. A model based on Markov random fields, in which modeling of the synchrony between frequency channels was implemented, was previously proposed [4] and applied to filter bank output features. The results obtained were not satisfying since the speech representation was too variable and also maybe because no real parameter estimation algorithm had been used. We propose to extend this model to a more standard sub-band approach, with cepstral representation of the signal in each band, using the maximum likelihood estimation algorithm defined in [5].

The paper is organised as follows: we first recall the definition of the parametric random field based model and of the related parameter estimation algorithms. Experiments on isolated word recognition with a multi-band approach are commented in section 3 and some concluding remarks are finally given.

2. RANDOM FIELD MODELING

2.1. Parametric model definition

2.1.1. Markov random field theory

A Markov random field X is defined by a neighbourhood system V on a lattice S so that the probability of a value at a point of the lattice, or site, s only depends on a finite number of neighbouring sites. This can be formally written as $P[x_s|x_r\ \forall r \neq s] = P[x_s|x_r\ \forall r \in V_s]$, where V_s denotes the neighbourhood of site s[1]. A given neighborhood system defines a set of cliques, a clique being a set of sites which are mutually neighbours. The Hammersley-Clifford theorem [6] ensures that, if no configuration has a null probability, a Markov field is equivalent to a Gibbs field and therefore there exists some clique potential functions $U_c(x)$ so that the probability of a configuration x of the Markov random field is given by the Gibbs distribution

$$P[x] = \frac{1}{Z} \exp\left(-\sum_{c \in \mathcal{C}} U_c(x)\right) \quad . \tag{1}$$

[1]For sake of simplicity, we denote by $P[x|y]$ the probability $P[X = x|Y = y]$.

In (1), the constant Z, called the partition function, is a normalisation term to define a probability measure.

As can be seen, the probability of a configuration is defined by a set of local probabilities on the cliques and a parametric model is completely defined by the neighbourhood system and the potential functions of the associated cliques. It is interesting to note that the higher the total energy $\sum_{c\in\mathcal{C}} U_c(x)$, the smaller the probability of the configuration x.

2.1.2. Potential functions

In the multi-band approach, the law of the hidden process (or field) X is given by the parallel HMMs and the probability $P[X_{t,k} = i]$ of being in state i in band k at time t only depends on $X_{t-1,k}$. It can be shown that this monolateral relation has a bilateral equivalence [7]. In this case, the neighbourhood of site (t,k) consists of the two sites $(t-1,k)$ and $(t+1,k)$. In order to model some interactions between the bands, the latter neighbourhood was extended to

$$V_{t,k} = \{(t-1,k),(t+1,k),(t,l) \quad \forall l \neq k\} \ . \tag{2}$$

Two types of cliques, namely horizontal and vertical cliques, are associated to such a neighbourhood definition and potential functions must be defined for both types of cliques.

Previous studies [8, 9] on random field modeling for speech recognition used the fact that a Markov chain is a particular Gibbs distribution where the potential function associated to the clique $\{(t-1,k),(t,k)\}$ is given, in band k, by

$$U^{(h)}_{t,k} = \sum_{i,j} a^{(k)}_{ij}\ \delta(x_{t-1,k} = i)\,\delta(x_{t,k} = j) \ . \tag{3}$$

In this equation, the function $\delta(x_{t,k} = j)$ equals 1 if $x_{t,k} = j$ and 0 otherwise. The parameters $a^{(k)}_{ij}$, related to as transition weights, are given by $-\ln(P^{(k)}(i,j))$ if $P^{(k)}$ is the transition matrix of the HMM corresponding to band k. A barrier energy is used for forbidden transitions, the probability of such a transition therefore being small enough so that the transition is never observed.

In full-band HMMs, all the bands are synchronized by default. It was shown in [3] that the performances are increased when asynchrony between the bands is allowed. The multi-band model also relies on the asynchrony assumption. The idea of the proposed approach is to model explicitly the synchrony (or asynchrony) between the bands. If two bands are considered synchronous when the state transition occurs at the same instants (*i.e.* transition $i \to i+1$ observed at the same time in the two bands in a left-right HMM), then a possible model of the synchrony is given by the clique potential function

$$U^{(v)}_{k,l} = f_{kl}\ |x_{t,k} - x_{t,l}| \ . \tag{4}$$

In this equation, the potential is defined for the clique $\{(t,k),(t,l)\}$ and it is assumed that the same number of states is used in each band. If f_{kl} is given a high value, the two bands are synchronous since the difference $|x_{t,k} - x_{t,l}|$ will be small for likely configurations.

2.1.3. Prior and posterior laws

According to the previously defined clique potential functions (3) and (4), the prior probability of a configuration x for a K band model with N states in each band is given by (1) with the following total energy

$$U(x) = \sum_{k=1}^{K}\sum_{i,j=1}^{N} a^{(k)}_{ij}\ \varphi^{(k)}_{ij}(x) + \sum_{k,l>k}^{K} f_{kl}\ \psi_{kl}(x) \ , \tag{5}$$

where

$$\varphi^{(k)}_{ij}(x) = \sum_t \delta(x_{t-1,k} = i)\,\delta(x_{t,k} = j)$$

counts the number of $i \to j$ transitions in band k and

$$\psi_{kl}(x) = \sum_t |x_{t,k} - x_{t,l}|$$

is the cumulated "gap" between bands k and l.

If we assume conditional independence of the observations $y_{t,k}$ and a Gaussian law for the probability density function (pdf) associated to each state, the likelihood of an observation y knowing $X = x$ is given by a Gibbs distribution whose energy is

$$U(y|x) = \sum_{t,k}\left(-\sum_i \delta(x_{t,k} = i)\ \ln g(y_{t,k};\mu^{(k)}_i,\sigma^{(k)}_i)\right) \tag{6}$$

where $g()$ is the Gaussian density function and $\mu^{(k)}_i$ and $\sigma^{(k)}_i$ are the mean and standard deviation associated with state i in band k.

Finally, it can be shown that the energy of x under the posterior law is given by $U(x|y) = U(x) + U(y|x)$ and therefore, random field based techniques can also be applied to the posterior distribution for sampling or finding out the most likely configuration.

2.2. Parameter estimation

2.2.1. ML estimation

If the signal is divided into K bands, the proposed random field model (RFM) is defined by the set of parameters θ which consists of the (N,N) matrices $A^{(k)}$ $(k = 1,\ldots,K)$ gathering the transition weights and of the (K,K) synchronisation matrix F. To estimate these parameters from examples, a generalized stochastic EM algorithm is proposed, where the maximisation step is replaced by a gradient probabilistic descent step.

In the case of a single example, the auxiliary function of the EM algorithm is given by

$$\begin{aligned} Q(\theta,\theta^{(n)}) \quad = \quad & -\sum_k\sum_{i,j} a^{(k)}_{ij} E_{\theta^{(n)}}[\varphi^{(k)}_{ij}(x)|y] - \\ & \sum_{k,l>k} f_{kl} E_{\theta^{(n)}}[\psi_{kl}(x)|y] - \ln Z_\theta - \\ & \sum_{t,k} \gamma_{t,k}(i)\ \ln g(y_{t,k};\mu^{(k)}_i,\sigma^{(k)}_i) \end{aligned} \tag{7}$$

where Z_θ is the prior partition function associated to the Gibbs energy (5), $\theta^{(n)}$ is the current estimate of the parameters θ and $\gamma_{t,k}(i)$ is the posterior probability that $X_{t,k} = i$ for the parameters $\theta^{(n)}$. The derivation of (7) w.r.t. $a_{ij}^{(k)}$ gives the following equation for $\theta^{(n+1)}$

$$E_{\theta^{(n+1)}}[\varphi_{ij}^{(k)}(x)] - E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)|y] = 0 \quad . \tag{8}$$

Similar expressions are obtained for the f_{kl} parameters and only the updating of the transition weights will be illustrated in the paper. There is no analytic solution to Eq. (8) and, as in [10], we propose to use a single step of a descent algorithm to calculate the new estimation which is then given by

$$a_{ij}^{(k)} \leftarrow a_{ij}^{(k)} + \frac{E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)] - E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)|y]}{V_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)]} \quad . \tag{9}$$

The variance $V_\theta^{(n)}[\varphi_{ij}^{(k)}(x)]$ comes from the second order derivative of $\ln Z_\theta$ in (7) w.r.t. $a_{ij}^{(k)}$. Since none of the expectations involved in (9) can be calculated, they are estimated from samples of the field X drawn under the prior and posterior laws with the current set of parameters $\theta^{(n)}$.

Obviously, the $a_{ij}^{(k)}$ parameters are not independent and eq. (9) can not be applied directly for each parameter. To overcome the problem, the transition weights corresponding to the same starting state i are grouped in a vector and derivation is performed w.r.t. to that vector. A similar updating equation is used with the Jacobi matrix to account for the parameter dependencies and, in this case, the Jacobi matrix is the covariance matrix of the $\varphi_{ij}^{(k)}$ functions under the prior law.

For the Gaussian pdf parameters, the re-estimation formulae are the same as for the standard HMM approach (see [11]) but the expectations $E_{\theta^{(n)}}[\delta(x_{t,k} = i)|y]$ are estimated from samples under the posterior law rather than explicitly calculated.

2.2.2. Initialisation

Since it is well known that the EM algorithm converges to a local maximum of the likelihood, an initialisation strategy based on empirical estimators of the transition weights and of the pdf parameters is proposed. For an observation, a maximum a posteriori estimate x^* of the hidden field can be determined using the ICM algorithm [12] or a simulated annealing scheme [13]. The transition weights are estimated by counting the number of transitions in x^* and taking the opposite of the logarithm of the corresponding estimated transition probability, which formally gives

$$a_{ij}^{(k)} = -\left(\ln(\varphi_{ij}^{(k)}(x^*)) - \ln\left(\sum_j \varphi_{ij}^{(k)}(x^*)\right)\right) \quad .$$

The pdf parameters are estimated with the feature vectors associated to the considered state.

Good results were obtained when applying these procedures to simulated data.

2.3. Decoding strategies

Finally, a decoding strategy is proposed for isolated word recognition. Since it is not possible to compute the prior probability (1) of X because of the partition functions, some approximation must be done. The score of an observation y for a word w, $p_w(y)$, is approximated by

$$p_w(y) = pl_w(x^*)p_w(y|x^*) \tag{10}$$

where x^* is the most likely posterior configuration and pl denotes the pseudo-likelihood [14] of x^*. The pseudo-likelihood is the product on all sites of the local conditional probabilities. As for the parameter initialisation, x^* can be obtained with the ICM algorithm or with a simulated annealing algorithm, the former being faster but sub-optimal unless a good initial solution is available. As the ICM and simulated algorithms are iterative algorithms, they are initialised with a uniform segmentation.

3. EXPERIMENTAL RESULTS

3.1. Experimetal setup

Experiments are carried out on single-speaker isolated word recognition for telephone speech extracted from the PolyVar database [15]. The vocabulary consists of 10 common words. Fifty occurrences of each words are used for training while 50 other ones are used for the tests. Results are therefore reported for 500 tests and results must be taken with care since the confidence intervals are quite large.

The speech signal is divided into sub-bands regularly spread on a MEL scale and cepstral coefficients are computed in each band. The cepstral coefficients in a given band are computed as the inverse Fourier transform of the log module of the spectral coefficients corresponding to that band, after symmetrisation.

3.2. Results

Several sub-band divisions were tested and results for 1, 3, 5 and 7 bands, with respectively 12, 5, 3 and 2 cepstral coefficients in each band, are given in table 1. These results were obtained using an ICM based decoder. The first line corresponds to independent parallel HMMs. These models are independently trained and the estimation of x^* in the decoding stage (10) is obtained by the Viterbi algorithm applied independently in each band as proposed in [4]. The two following lines corresponds to random field approaches. In the ICM case, models were simply initialised using an ICM segmentation while for the last line, EM re-estimation of the parameters was used. Similar results are given in table 3.2 for a simulated annealing based decoder.

training	b1c12	b3c5	b5c3	b7c2
heuristique	99.8	99.4	97.6	95.0
ICM	87.2	84.6	78.8	78.2
ICM-GEM	88.6	80.6	75.4	76.0

Table 1: Recognition rate (in %) with an ICM based decoder for different sub-band architectures.

training	b1c12	b3c5	b5c3	b7c2
ICM	99.6	97.8	92.6	88.2
ICM-GEM	–	97.8	95.0	94.2

Table 2: Recognition rate (in %) with a simulated annealing based decoder for different sub-band architectures.

The comparison of the two tables points out the weaknesses of the ICM based decoder compared to the simulated annealing based one which is not surprising. In the full-band case (b1c12), the proposed model along with the simulated annealing decoding strategy performs as well as the standard HMM approach given by the heuristique case. It can also be seen that the recognition rate decreases when the number of bands increases. This could be explained by the fact that the representation of speech in each band is poorer since less coefficients are used. However, when the number of cepstral coefficients is increased in the 7 band case, the recognition rate only marginally improves. For example, a recognition rate of 96.4% is achieved with 7 bands and 5 cepstral coefficients using the heuristique training, compared to 95% with only 2 coefficients. This points out the fact that the recombination is the crucial point of multi-band models.

In the case of a ICM based decoder, the modeling of inter-band synchrony seems to deteriorate the results. Indeed, in the ICM training, the synchronisation weights are set to zero and are not updated. Therefore, there is no inter-band modeling in this case. Except for the full-band case, the recognition rates for the ICM-GEM training case in table 1 are less than the ones obtained with ICM training. On the other hand, the opposite is observed in the case of a simulated annealing based decoding strategy. In table 3.2, no decrease of the recognition rate is observed between the ICM and ICM-GEM cases, better performances being obtained in the latter case with 5 bands. This fact shows that a good model of the prior process X is needed when the observation becomes more variable. Indeed, the more bands, the more variable the observations in each band. Therefore, adding a synchrony model, even if the model is not perfect, allows for a better regularisation of the segmentation and, as a matter of fact, for a higher recognition rate.

4. CONCLUSION

In this paper, a multi-band model with modeling of the inter-band time synchrony is proposed. A maximum likelihood parameter estimation procedure is defined and experiments on isolated word recognition are presented. These experiments showed that, in any cases, increasing the number of bands goes along with a decrease of the recognition rate. However, with a simulated annealing based scoring algorithm, better results are achieved when inter-band synchrony is modeled. When the observation becomes too variable, which is the case when the number of bands is increased, there is a need for regularisation of the segmentation and the experiments showed that inter-band synchrony is an interesting and valuable regularisation model. To conclude, we should stress the fact that the synchrony modeling used in the current random field model assumes that the same synchronisation weight applies for all the duration of a word. This assumption is certainly wrong and a better modeling of the synchrony should be envisaged.

References

[1] F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, 1998.

[2] H. Hermansky, M. Pavel, and S. Tibrewala. Towards ASR using partially corrupted speech. In *ICSLP*, pages 458–461, Oct. 1996.

[3] M. J. Tomlinson, M. J. Russel, R. K. Moore, A. P. Buckland, and M. A. Fawley. Modelling asynchrony in speech using elementary single-signal decomposition. In *ICASSP*, volume 2, pages 1247–1250, 1997.

[4] G. Gravier, M. Sigelle, and G. Chollet. Toward Markov random field modeling of speech. In *ICSLP*, December 1998.

[5] G. Gravier, M. Sigelle, and G. Chollet. Markov random field modeling for speech recognition. *Australian Journal for Intelligent Information Processing Systems*, 5(4), 1998.

[6] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statistical Soc.*, B-48:192–236, 1974.

[7] H.-O. Georgii. *Gibbs measures and phase transitions*, volume 9 of *Studies in Mathematics*. de Gruyter, 1988.

[8] H. Noda, M. N. Shirazi . A MRF-based parallel processing algorithm for speech recognition using linear predictive HMM. In *ICASSP*, volume 1, pages 597–600, 1994.

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

[10] K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *J. Royal Statistical Soc.*, 57(2):425–437, 1993.

[11] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993.

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

[13] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[14] B. Chalmond. An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition*, 22(6):747–761, 1989.

[15] G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaboulet, and P. Langlais. Swiss French PolyPhone and PolyVar: telephone speech databases to model inter- and intra-speaker variability. Research Report 96-01, IDIAP, April 1996.