Markov Random Field Modeling for Speech Recognition

Guillaume Gravier, Marc Sigelle and Gérard Chollet

ENST/TSI and CNRS-URA 820
46 rue Barrault, 75634 Paris Cedex 13, France
email: {gravier,sigelle,chollet}@tsi.enst.fr

Abstract

This paper presents a new framework for ***statistical modeling of speech*** *based on* ***Markov random fields.*** *Classical and* ***multi-stream HMM*** *approaches are particular cases of the new family of model proposed. As an illustration, a Random Field Model (RFM) is presented for the hidden process to replace HMMs. This model can be seen as parallel HMMs, one for each sub-band, interacting together, where the interaction consists in controlling the synchrony between the HMMs. Results presented on single-speaker* ***isolated word recognition*** *experiments show that this model does not perform as well as the classical HMMs. However, this imperfect model allows to review and define the algorithms related to Markov random field modeling of speech and, because of its flexibility, we believe that this technique is promising. Indeed, RFMs are simply defined by local interactions and the associated potential functions and, as long as the energy function for the field is linear w.r.t. to the parameters of the model, all the algorithms presented here are applicable. Such a flexibility opens interesting perspectives for statistical speech modeling.*

1 Introduction

The statistical approach of speech recognition is classically based on the *maximum a posteriori* criterion which is decomposed, thanks to the Bayes rule, in an acoustic and a linguistic score. More formally, for an observation $Y = y$, one searches for the sequence of words $\hat{W}$ given by

$$\hat{W} = \arg\max_{w} P[Y = y|W = w]P[W = w] \ , \qquad (1)$$

where $P[Y|W]$ corresponds to the acoustic score while $P[W]$ corresponds to the linguistic one. There are several ways for computing the acoustic scores but the most widely used technique is based on hidden Markov models (HMM). In this approach, a hidden process, the Markov chain, is used to model the temporal structure of speech and probability density functions (pdf) associated to each of the Markov chain states are used to model the frequency variability of speech [1, 13]. The sequence of feature vectors is then considered to be piecewise stationary. If we denote X the hidden process, the acoustic score for the word W is given by

$$P_W[Y = y] = \sum_x P_W[Y = y|X = x]\, P_W[X = x] \ ,$$

and the summation over all state sequences is commonly replaced by the predominant term of the sum which can be computed using the Viterbi algorithm [20]. Usually, the pdf associated to the states are Gaussian mixtures and a HMM is defined by its transition matrix A and the weights, means and covariance matrices of each pdf.

Recently, an extension of this model, the multi-stream approach, has been proposed [4]. It consists in representing speech with different streams and modeling independently each stream by a HMM, recombination of the different streams being done at some points in time. A common approach for this model is to divide the signal into frequency bands and to model independently each sub-band [5, 12]. The motivation for such a model is its ability to be robust to band limited noises.

HMMs can be seen as the superposition of two stochastic models, one in the time domain and one in the frequency domain. A real 2-dimensional process should be more appropriate than the superposition of two stochastic processes. In the multi-band approach, it is assumed that the bands are independent. It seems clear that this assumption is intrinsically limitative. Indeed, some interactions between the frequency bands obviously exist and we believe that modeling those interactions should help. Markov random fields [2] can be used for this purpose. In the multi-band approach, the hidden process defines a field $X = \{X_{t,k}\}$ where k is the band index, the law for this field being defined by the HMMs in each band. The process X is defined on a lattice $S = \{(t, k)\}$ and the probability of being in a state i at (t, k) only depends on the state in which the process is at $(t-1, k)$. As a first approach to Markov random field modeling of speech, we propose to model interactions between the bands by changing the law for the hidden process X. Interactions between each of the HMMs are added to the model by assuming that the probability for the hidden process to be in

a given state at (t,k) depends on the states observed at $(t-1,k)$, $(t+1,k)$ and (t,l) for each $l \neq k$. Such dependencies define a neighborhood system and, using the Hammersley-Clifford theorem [2], the law of X can be expressed in terms of interaction potentials. The observations are still modeled by Gaussian pdf associated to each state of the underlying HMMs using the classical assumption of conditional independence. The expected advantages of a random field based model are that it may be able to capture some frequency information that can not be handled by HMMs. The state space of the hidden process is also more complex than for the multi-band model and it therefore may be able to capture more information in the signal as clearly shown in [9]. Another advantage of this new formulation is that the model can easily be modified and extended under light assumptions. A new framework for statistical modeling of speech is therefore defined and a model is presented as an illustration.

Very few studies have addressed the problem of Markov random field modeling of speech. In [22], it is shown that a HMM is a particular random field. The authors redefine the hidden Markov model in terms of Gibbs distribution and extend the Baum-Welch algorithm to their case. In [18], the same equivalence between HMM and random fields is used to design a parallel Viterbi decoder using the Iterated Conditional Mode (ICM) algorithm. Finally, Lucke, in [16], defines a random field model to capture some information between the cepstral coefficients presenting a rather complex training procedure for his model.

In this paper, we first define the random field model proposed and then present the algorithms used for training and scoring speech segments in section 3. In section 4, we present some results on speaker-dependent isolated word recognition. We finally discuss the perspective of such a model and conclude.

2 A random field model

2.1 Markov random field theory

In this section, a random field based model, where synchronization between the underlying Markov chains in the sub-bands is added, is defined. A Markov random field X is defined by a neighborhood system on a lattice S so that

$$P[X_s = x_s | X_r = x_r \, \forall r \neq s] = P[X_s = x_s | X_r = x_r \, \forall r \in V_s] \ ,$$

where V_s denotes the neighborhood of the point s. A set of mutually neighbor points of the lattice on which the field is defined is called a clique. Potential functions must also be defined for all the cliques where the higher the potential, the less probable the configuration. A clique potential function is a function of all the points that belong to the considered clique. Then, the Hammersley-Clifford theorem states that the probability law for the process X can be written as a Gibbs distribution given by

$$P[X = x] = \frac{1}{Z} \exp - \sum_{c \in \mathcal{C}} U_c(x) \ , \qquad (2)$$

where $\mathcal{C}$ is the set of all the cliques defined on the lattice S, and $U_c()$ is the potential function associated with the clique c. As can be seen, the probability of a configuration is defined by a set of local probabilities on the cliques. The constant Z, called the partition function, ensures that equation (2) defines a probability measure and is therefore given by

$$Z = \sum_x \exp - \sum_{c \in \mathcal{C}} U_c(x) \ .$$

In the case under consideration, the lattice S is a two dimensional lattice indexed by (t,k) where $t \in [1,T]$ is the time index and $k \in [1,K]$ is the band index. In the following a point of the lattice will be denoted by its coordinates (*e.g.* (t,k)). The neighborhood of (t,k) is defined by $V_{t,k} = \{(t-1,k),(t+1,k),(t,l) \forall l \neq k\}$. This neighborhood system defines two kinds of cliques, the first ones $\{(t-1,k),(t,k)\}$ accounting for the time modeling while the second ones $\{(t,k),(t,l)\}$ model the frequency interactions (or at least the band interaction). The potential functions associated to both types of cliques have now to be defined. We will refer to the first kind of cliques as horizontal ones while the second ones will be referred to as vertical.

2.2 Potential definitions

2.2.1 Horizontal potentials

The horizontal cliques are designed to correspond to the cliques defined by a Markov chain. Indeed, for a Markov chain with transition matrix A_c, the probability of X is given by

$$\begin{aligned} P[X = x] &= \pi(x_1) \prod_{t=2}^{T} a_c(x_{t-1}, x_t) \\ &= \exp - \sum_{t=1}^{T} a(x_{t-1}, x_t) \ , \end{aligned}$$

assuming that $a(i,j) = -\ln a_c(i,j)$ and $a(0,j) = -\ln \pi(j)$. The second expression for $P[X = x]$ corresponds to the formulation of a Markov chain in terms of Gibbs distribution. For a transition that is disabled, that is $a_c(i,j) = 0$, the corresponding energy is theoretically infinite and $P[X] = 0$. Actually, the Hammersley-Clifford theorem (2) applies only if all the possible configurations have a non-zero probability and therefore, the infinite energy is replaced by a barrier

energy arbitrarily set to a high value so that forbidden configurations have a very low probability. For the clique $\{(t-1,k),(t,k)\}$, the potential function is therefore given by

$$U_{t,k}(x) = a^{(k)}(x_{t-1,k}, x_{t,k}) \ ,$$

where such functions are defined for each band k.

2.2.2 Vertical potentials

The vertical interactions allow for a control of the synchrony between two bands, say k and l. The idea is that when two bands have their spectrally stable zones occurring at the same instants, then the states in the corresponding HMMs should change at about the same time because of the piecewise stationarity of the HMMs. To reflect this behavior, the vertical potential function for the clique $\{(t,k),(t,l)\}$ are defined as

$$U_{t,k,l}(x) = f_{k,l}\,|x_{t,k} - x_{t,l}| \ ,$$

assuming that the HMMs in each bands have the same number of states. The parameter $f_{k,l}$ is a synchronization weight since when it is high, the absolute difference $|x_{t,k} - x_{t,l}|$ tends to be small in order to have a low energy configuration. It must be noted that the synchronization does not depend on the time. This may be true for short segments but it is clear that this assumption is false for longer segments which may be synchronized during a certain time and unsynchronized the rest of the time. This point will be discussed in the last section. The choice of modeling the band synchrony also need to be discussed and validated.

2.2.3 Attachment to the data

The observations are classically modeled with Gaussian probability density functions associated with each states of each sub-band underlying HMMs, using the assumption of conditional independence of the observations. The likelihood of the observation Y knowing that $X = x$ can also be expressed as a Gibbs distribution

$$P[Y = y|X = x] = \exp - \sum_{t,k} - \ln g(y_{t,k}; \mu_i^{(k)}, \sigma_i^{(k)}) \ , \tag{3}$$

where $g()$ is the Gaussian pdf with mean $\mu_i^{(k)}$ and diagonal covariance $\sigma_i^{(k)}$, assuming that $x_{t,k} = i$.

2.3 Prior and posterior laws

As shown previously, a random field model is defined by

$$\Lambda = \{N, K, A^{(k)} k \in [1,K], F\} \ ,$$

where N is the number of states, K the number of bands, $A^{(k)}$ the (N,N) transition weight matrix in band k and F is the (K,K) matrix of synchronization weights. According to equation (2), the energy of the field under the prior law is given by

$$\begin{aligned} U(x) &= -\sum_{t,k} a^{(k)}(x_{t-1,k}, x_{t,k}) \\ &\quad - \sum_{t,k,l>k} f_{k,l}|x_{t,k} - x_{t,l}| \\ &= -\sum_{k}\sum_{i,j} a_{i,j}^{(k)} \varphi_{i,j}^{(k)}(x) \\ &\quad - \sum_{k,l>k} f_{k,l}\psi_{k,l}(x) \ , \end{aligned} \tag{4}$$

where $\varphi_{i,j}^{(k)}(x) = \sum_t \delta_i(x_{t-1,k})\delta_j(x_{t,k})$ and $\psi_{k,l}(x) = \sum_t |x_{t,k} - x_{t,l}|$. The function $\delta_i(x_s)$ takes a value of 1 if $x_s = i$ and is 0 otherwise. The second formulation has the advantage to clearly stress that the energy function $U(x)$ for the field is linear with respect to the parameters, which will be important for the parameter estimation procedure. For the posterior law, the energy of the field is

$$U(x|Y = y) = U(x) + U(y|x)$$

where $U(y|x)$ is the potential function defined by equation (3).

Now that the model is completely defined, we review the algorithms related to Markov random field modeling of speech.

3 Algorithms for random fields

In this section we review and define the algorithms that enable to solve the two main problems of speech recognition. The first problem is to find out the most likely hidden state sequence for a given observation. This is solved by the Viterbi algorithm in the HMM framework. The second problem consists in estimating the model parameters given a set of examples and is classically done using the Baum-Welch algorithm for hidden Markov models. None of the previously mentioned algorithm applies to random fields.

3.1 The decoding problem

As mentioned in the introduction, the computation of $P[Y = y|W = w]$ in equation (1), which requires a sum over all possible state sequences, is often approximated by finding out the most likely state sequence for a given observation. In the random field framework, two algorithms are available for computing *maximum a posteriori* configurations. The first algorithm is known as the Iterated Conditional Modes (ICM) algorithm [17] and its principle is to iteratively maximize

the local conditional probabilities at each point of the lattice. Starting from an initial configuration, the algorithm iterates over each site (t,k), assigning it the value defined by $X_{t,k} = \arg\max_i P[X_{t,k} = i | x(V_{t,k}), Y]$ until no changes are made throughout the lattice, $x(V_{t,k})$ being the configuration associated with the neighborhood of point (t,k) for the current estimate of the solution. The main advantage of the ICM algorithm is that it is very fast since it converges in very few iteration but unfortunately, it converges to a local maximum and it therefore requires a good initialization.

The second algorithm is based on simulated annealing [14]. Given a field X and its associated energy function $U(x)$, the principle of the simulated annealing algorithm is to draw some samples of the field with a potential function given by $\frac{U(x)}{T}$ using a Gibbs sampler [8], with a decreasing temperature T. The algorithm starts with a high temperature and, as the temperature decreases according to a geometric law, the solution (*i.e.* the samples drawn) converges towards the minimum energy configuration. It can be shown that if the temperature decreases sufficiently slowly, this algorithm converges to the global maximum of the posterior probability. The drawback of this algorithm is that it can be very slow ...

3.2 Parameter estimation

3.2.1 Generalized EM procedure

The maximum likelihood estimation of the parameters of a hidden field still is an open issue in the random field literature. When closed form solutions of the parameter estimates are available, algorithms such as the stochastic EM or the ICE [19] can be used. On the contrary, a gradient probabilistic descent algorithm can be used [21] and we propose here a generalized EM algorithm. The principle of the EM algorithm [7] is to first compute the expectation of the complete statistics (*i.e.* observation and hidden process) log-likelihood with the true parameters under the posterior law for the current estimate of the parameters and to then maximize this auxiliary function. Usually, this maximization step gives a closed form estimate of the parameters which is not the case with Markov random fields. In this case, as in [15], the maximization can be replaced by a gradient probabilistic descent step.

The auxiliary function $Q(\Theta, \Theta^{(n)})$, where Θ denotes the model parameters (*i.e.* $A^{(k)} k \in [1,K], F$) and $\Theta^{(n)}$ the current estimate, associated with the potential functions defined in the previous sections is given by

$$\begin{aligned} Q(\Theta,\Theta^{(n)}) &= -\sum_k \sum_{i,j} a_{i,j}^{(k)} E[\varphi_{i,j}^{(k)}(x) \mid Y, \Theta^{(n)}] \\ &\quad - \sum_{k,l>k} f_{k,l} E[\psi_{k,l}(x) \mid Y, \Theta^{(n)}] \\ &\quad - \ln Z_\Theta \\ &\quad - \sum_{t,k} \sum_i \gamma_{t,k}(i) \ln g(Y_{t,k}; \mu_i^{(k)}, \sigma_i^{(k)}) \ , \end{aligned}$$

where $\gamma_{t,k}(i) = p[X_{t,k} = i]$ and Z_Θ is the partition function associated to the Gibbs distribution (2) for the prior law with parameter Θ. The derivation of $Q(\Theta, \Theta^{(n)})$ with respect to $a_{i,j}^{(k)}$ leads to

$$\frac{\partial Q(\Theta,\Theta^{(n)})}{\partial a_{i,j}^{(k)}} = E[\varphi_{i,j}^{(k)}(x)|\Theta] - E[\varphi_{i,j}^{(k)}(x)|Y,\Theta^{(n)}]$$

and similar formulae are obtained for the $f_{k,l}$ parameters. As mentioned previously, this means that no closed form formulae are available for computing the new parameter estimates. In order to maximize the auxiliary function, a gradient probabilistic descent step is applied. Before giving the re-estimation formulae, it must be noted that the $a_{i,j}^{(k)}$ parameters are not independent. They are considered as independent for different values of i and k, but not with respect to j for i and k fixed. In this case, the Hessian matrix must be computed and we consider the vector of parameters $a_{i,(k)}$ whose j'th element is $a_{i,j}^{(k)}$. The Hessian matrix $H_{i,(k)}$ is defined by

$$\begin{aligned} H_{i,(k)}(m,n) &= \frac{\partial Q(\Theta,\Theta^{(n)})}{\partial a_{i,m}^{(k)} \partial a_{i,n}^{(k)}} \\ &= -\mathrm{cov}_\Theta(\varphi_{i,m}^{(k)}(x), \varphi_{i,n}^{(k)}(x)) \ , \end{aligned}$$

where $\mathrm{cov}_\Theta()$ denotes the covariance under the prior law with the true parameters. If we denote, $a'_{i,(k)}$ the vector containing the partial derivatives of $a_{i,(k)}$ w.r.t. $a_{i,j}^{(k)}$, the following re-estimation formula is obtained

$$a_{i,(k)}^{(n+1)} = a_{i,(k)}^{(n)} - H_{i,(k)}^{-1} a'_{i,(k)} \ .$$

As the structure of the Hessian can be very particular, specially for left-right HMM topologies, the pseudo-inverse of the matrix is used instead of the traditional inverse. The $f_{k,l}$ parameters are considered as independent and the re-estimation formula is therefore

$$f_{k,l}^{(n+1)} = f_{k,l}^{(n)} + \frac{E[\psi_{k,l}(x)|\Theta^{(n)}] - E[\psi_{k,l}(x)|Y,\Theta^{(n)}]}{\mathrm{var}(\psi_{k,l}(x)|\Theta^{(n)})}$$

Finally, the re-estimation formulae for the pdf parameters, based on the $\gamma_{t,k}(i)$ quantities, are exactly the same as in the Baum-Welch algorithm. In all the re-estimation formulae, the expectations involved are estimated using samples drawn under the prior and the posterior laws using the current estimate of the parameters.

3.2.2 Initialization

The EM algorithm is known to converge to a local maximum of the likelihood and therefore a good initialization procedure must be defined before running it. To initialize the model, we iteratively estimate the MAP fields for all the training sequences, and update the transition weights counting the number of transitions for a given state to all the other states, and the pdf parameters with the vectors associated to the considered state.

3.2.3 Heuristic training

Since the horizontal potential functions were defined to reflect a Markov chain behavior, one can train those weights as well as the pdf parameters using the Baum-Welch algorithm independently in each band. The synchronization weights are then defined as

$$f_{k,l} = \frac{\gamma T}{\sum_t |x_{t,k} - x_{t,l}|}$$

where the sum over t is computed along the Viterbi path. The factor γ controls the relative importance of the vertical potential w.r.t. the horizontal ones. More details for this procedure can be found in [10].

4 Experiments

4.1 Protocol

In this section, experiments on speaker-dependent isolated word recognition using telephone speech are presented. A male speaker from the PolyVar database [6] and 10 keywords are used. The data were collected over the telephone over a period of one year. One hundred occurrences of each keyword are considered, the first 50 being used as the training corpus while the remaining ones are used to carry out the tests. The keywords have between 8 and 3 phonemes in length and the models have two states per phoneme in the word. In all the models, the attachment to the data is modeled by a single Gaussian pdf. A reference system is developed using classical hidden Markov modeling techniques with two kinds of feature vectors. The system HMM_{cep} is based on feature vectors of 12 cepstral coefficients on a linear frequency scale while the system HMM_{fbk} operates directly on the output of a 24 channel filter-bank. It may be surprising to use the output of a filter-bank as a feature vector but, since the RFM models frequency interactions, it should be able to operate directly on the output of a filter-bank. This representation also has the advantage of making no assumption on speech production.

4.2 RFM and filter-bank features

First, experiments with models trained using the heuristic procedure defined previously are presented. It must be noted that for full-band models, the heuristic is equivalent to the standard Baum-Welch training of HMMs. A random field model with 24 bands was trained with the filter-bank features. Scoring of the test segments was done using the ICM algorithm with two different initialization strategies. The first strategy consists in starting with a uniform segmentation of the hidden field. In the second strategy, the Viterbi solution is computed independently in each band and used to initialize the ICM solution. Results for various values of γ are summarized in table I. These results

	HMM		RFM (γ)			
	cep	fbk	0.0	0.005	0.02	0.05
uniform	84.4	59.6	43.2	43.8	43.6	47.2
Viterbi	99.8	94.6	69.0	69.6	70.0	69.2

Table I: Recognition rates using the ICM algorithm (in %)

show that the HMMs clearly outperform the random field model and that cepstral coefficients are the best way of representing the speech signal. The weakness of the ICM algorithm is also clearly pointed out. The synchronization parameter γ seems to have some influence on the results, specially for the uniform ICM case. It was observed that, as γ increases, more changes are done by the Viterbi-ICM algorithm but the hidden process energy variations are small compared to the attachment to the data. In other words, as γ increases, the first term of the energy under the posterior law, given by $U(x) + U(y|x)$, tends to have more influence but still is small compared to $U(y|x)$. A *regularisation hyper-parameter* β was therefore added so that the posterior energy is defined as $\beta U(x) + U(y|x)$. In the case of HMMs, this would correspond to multiplying by β the transition log-probabilities during the Viterbi search. The Viterbi-ICM algorithm was used with models trained for $\gamma = 0$ and $\gamma = 0.02$ for various values of the hyper-parameter β and the results are reported in figure 1. These results clearly show that the performances of the system can be increased by according more importance to the hidden process. This result is also true for full band models though the increase of performance is less important.

Finally, the results points out the fact that the ICM algorithm is very sensitive to the initialization procedure and, even if we have an equivalence between HMMs and RFMs in the case of the heuristic training, it may be convenient to get rid of the Viterbi initialization. The ICM-based scoring algorithm is therefore replaced by a simulated annealing procedure based on

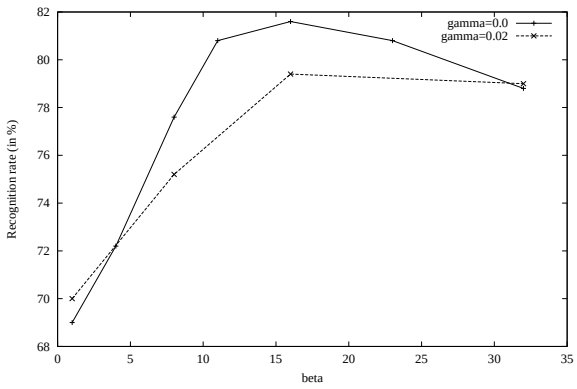

Figure 1: Recognition for the non-synchronized 24-channel filter-bank features as a function of β. The solid line is for $\gamma = 0$ and the dashed one is for $\gamma = 0.02$.

a Gibbs sampler as explained in section 3.1, starting with a temperature T_0 of 150 with a decrease law given by $T_n = T_0\, 0.997^n$. These parameters were set experimentally, the algorithm being initialized by the uniform field, and results are summarized in table II. In the case of pure hidden Markov models, the Viterbi algorithm performs better than the simulated annealing one but it can be seen that in the case of random field models, even without interactions between the bands, the simulated annealing algorithm performs better than the Viterbi initialized ICM.

	HMM_{cep}	HMM_{fbk}	RFM $\gamma = 0.0$
annealing	90.0	92.6	78.4

Table II: Recognition rates using the simulated annealing algorithm (in %)

4.3 Band architecture

It seems clear that the choice of the filter-bank output for the feature vector may not be an optimal solution, even if RFMs may be able to capture more information in such a process than HMMs. Several band divisions were tested in order to find out an optimal division of the frequency band. We divided the whole frequency band in 3, 5 and 7 bands according to the MEL scale and cepstral coefficients are computed in each band using the inverse DFT of the spectral coefficients corresponding to a band. Again in these experiments, the models are trained using the heuristic defined and results are reported in table III for $\gamma = 0$ and $\beta = 1$. The digit following b in the experiment names gives the number of sub-bands while the one following c gives the number of cepstral coefficients in each band (additional e stands for the energy). These results leads to the conclusion "the less bands, the better"! In [12], it

b1c12	b3c5	b5c3	b7c2	b7c3	b7c2e	b7c5
99.8	99.4	97.6	95.0	96.0	95.6	96.4

Table III: Recognition rate for various topologies of RFMs

is shown that the multi-band model with a MLP-based recombination procedure performs better than the full-band model. The results in table III shows the inverse but it must be stressed that no real recombination of the partial scores is done in our case. This shows that one of the key point of the multi-band model is the score recombination. It should also be mentioned that when there are few bands, the regularisation hyper-parameter β has basically no effect and that, adding some synchronization by increasing γ in the heuristic procedure tends to deteriorate the results. This means that the more bands, the more the *a priori* form of the hidden process is meaningful. But these conclusions must be taken with care since the heuristic training procedure is far from being optimal. It also must be stated that experiments are done in a simple case and they should be extended to the speaker-independent case and/or noisy case to definitely conclude on this point.

4.4 Training experiments

Since the likelihood of the data is not computable, it is not easy to define a convergence criterion for the generalized EM algorithm. The convergence of the MAP initialization procedure could be monitored using the average pseudo-likelihood of the data but not the EM procedure. We therefore simply limited the number of iterations for both algorithms. Another point is that there is no convergence proof for this algorithm but it was experimentally observed on artificial data that the algorithm converges towards the true values of the pdf parameters.

Using the generalized EM algorithm for the cepstral-based features and the simulated annealing algorithm for scoring, a recognition rate of 98.0 % is achieved. This rate still is not as good as the results obtained with a pure Viterbi algorithm but it outperforms the results obtained with the heuristic training, thus suggesting that a maximum likelihood training procedure adapted to the model is important. Another interesting point is the importance of the initialization procedure. In a second experiment, only the pdf parameters were initialized, leaving the transition weights unchanged. In that case, the transition weight matrices are poorly estimated with the EM algorithm and the recognition rate falls down to 51.6 %. This result shows that the transition weights are important in the random field formulation of speech recognition, while it is known that the transition probabilities have little influence in the hidden Markov model framework [3].

5 Discussion

We have presented a new statistical model for speech recognition based on Markov random fields to model the hidden process. The choice of the synchronized-HMM model presented here may be discussed since it relies on some assumptions that are basically wrong. Indeed, synchronization between the bands were added thanks to the $f_{k,l}$ parameters but these parameters are considered as time-invariant (or state independent which is the same in that case). As mentioned previously, this may be true for short segments where the synchronization is *uniform* across the segment but for longer segments, such as words as used in the experiments presented in this paper, this assumption is no longer valid. We plan to extend the synchrony model to state-dependent synchronization weights. This would also allow to force a full synchronization between two models in order to extend this technique to the connected word case.

Referring to Greenberg's work [11] in which it is shown that spectral asynchrony has no influence on the intelligibility of speech, the choice of modeling such a synchrony (or asynchrony) can be discussed. We believe that such a synchrony is an important factor for machine-based recognition of speech It could specially help in the case of corrupted speech (*e.g.* reverberant speech) to help keeping the synchronization in the model.

Whatever the weaknesses of the model proposed, the main purpose of the paper is to present a new technique for modeling speech and the model is used to test the algorithms related to such a technique. The main advantage of Markov random field based models is that, as long as the energy function of the field is linear with respect to the model parameters, the generalized EM algorithm presented here is applicable and new interactions (or longer range interactions) can easily be integrated in such a model by defining another neighborhood system and the related potential functions.

References

[1] J. K. Baker. *Stochastic modeling for automatic speech understanding.* R. Reddy ed., New York Academic Press, 1975.

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

[3] H. Bourlard. Reconnaissance de la parole: modélisation ou description? In *XXIes Journée d'Etude sur la Parole*, pages 263–272, 1996.

[4] H. Bourlard, S. Dupont, and C. Ris. Multi-stream speech recognition. Research Report RR 96-07, IDIAP, Dec. 1996.

[5] H. Bourlard et al. Towards subband-based speech recognition. In *EUSIPCO*, 1996.

[6] Gérard Chollet, Jean-Luc Cochard, Andrei Constantinescu, Cédric Jaboulet, and Philippe Langlais. Swiss French PolyPhone and PolyVar: telephone speech databases to model inter- and intra-speaker variability. Research Report RR 96-01, IDIAP, April 1996.

[7] A. P. Dempster, N. M. Laid, and D. B. Durbin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Soc.*, 39(39):1–38, 1977.

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

[9] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden Markov models. Technical report, Computational Cognitive Science Technical Report 9502, July 1996.

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

[11] Steven Greenberg and Takayuki Arai. Speech intelligibility is highly tolerant of cross-channel spectral asynchrony. In *Joint proceedings of the Acoustical Society of America and the International Congress on Acoustics*, Seattle, 1998.

[12] H. Hermansky, M. Pavel, and S. Tibrewala. Towards ASR using partially corrupted speech. In *Int. Conf. on Spoken Language Processing*, pages 458–461, Oct. 1996.

[13] F. Jelinek. Continuous speech recognition by statistical methods. *IEEE Proc.*, 64:532–556, April 1976.

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

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

[16] H. Lucke. Improved acoustic modeling for speech recognition using 2D Markov random fields. In *Int. Conf. on ASSP*, 1995.

[17] N. Metropolis, A. W. Rosenbluth, A. H. Teller, M. R. Rosenbluth, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 13(5):1087–1091, 1953.

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

[19] Wojciech Pieczynski. Champs de Markov cachés et estimation itérative. *Traitement du Signal*, 11(2):141–153, 1994.

[20] A. J. Viterbi. Error bounds for convolutional codes and asymptotically optimal decoding algorithm. *IEEE Trans. on Information Theory*, 13:260–269, April 1967.

[21] Laurent Younes. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82:625–645, 1989.

[22] Y. Zhao et al. Application of the Gibbs ditribution to hidden Markov modeling in speaker independent isolated word recognition. *IEEE Trans. on Signal Processing*, 39(6):1291–1298, 1991.