TOWARD MARKOV RANDOM FIELD MODELING OF SPEECH

G. Gravier *M. Sigelle* *G. Chollet*

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

ABSTRACT

In this paper, we present a new technique for statistical modeling of speech segments based on Markov random fields. Classical and multi-stream HMMs are particular cases of this more general family of models. However, the Random Field Model (RFM) proposed here can be seen as an extension of the multi-band HMM in which interactions between the frequency bands have been added. In a first experiment, samples are drawn from different models and compared to real observations. This experiment shows that the RFM is able to produce realistic samples but a single HMM still performs better. Isolated word recognition experiments stress the fact that more work must be done on the RFM in order to reach the performances of classical hidden Markov modeling techniques. For the moment, the RFM parameters are estimated using a heuristic. We believe that a real maximum likelihood parameter estimation algorithm should improve the results. The main advantage of this new model is that it can easily be extended since a model is defined by some local interactions and the Gibbs potential functions associated to those interactions.

1 INTRODUCTION

In speech recognition, many techniques have been proposed to compute the likelihood of an observation given a statistical model (or a sequence of words). The most popular approach is based on Hidden Markov Models (HMMs). In this approach, a hidden stochastic process (the Markov chain) is used to model the temporal structure of speech while the probability density functions associated to the Markov chain states model the frequency variability. More recently, an extension of this model to a multi-band approach has been proposed [1]. It consists in dividing the signal into frequency sub-bands and in modeling each sub-band independently by a HMM.

We see several limitations to those approaches. First, the HMM approach can be seen as the superposition of two stochastic models, one for the time domain and one for the frequency domain. To model both time and frequency variability simultaneously, a real 2D stochastic model seems more appropriate. Secondly, the multi-band HMM assumes the independency of the sub-bands. It seems clear that such an assumption is intrinsically limitative. Indeed, some interactions between the frequency bands obviously exist and this should be included in a model.

As a step toward a real two dimensional model of speech, a new model is proposed, derived from the multi-band approach, in which the interactions between frequency bands are taken into account. In the classical multi-band model, the hidden process defines a field $X = \{X_{t,k}\}$ where k is the band index and the law of the process X is defined by the HMMs in each band. To allow for frequency interactions, the law for the field is changed and it is assumed that X is a Markov random field and the hidden process model can be seen as parallel HMMs mutually interacting. In other words, the state in which we are at time t in band k, depends on the states in the same band at times $t-1$ and $t+1$ and on the states in the other bands at time t. Such a set of dependencies defines a neighborhood and, thanks to the Hammersley-Clifford theorem [3], the law of X can be expressed in terms of interaction potentials. Finally, the observations are modeled by Gaussian probability density functions associated to each state of the underlying HMMs using the classical hypothesis of conditional independence. The observations consist in filter-bank outputs. The formalism of this model, called Random Field Model (RFM), is presented in section 2. The expected advantages of the RFM are that it may be able to use some frequency information that cannot be captured by HMMs thanks to the modeling of frequency band interactions. The state space of the hidden process is also more complex than with HMMs and may be able to model more complex processes. The advantages of having a complex state space are clearly shown in [2] where factorial HMMs are used.

To validate the pertinence of the Random Field Model, random samples of isolated words, drawn according to the law defined by the model, are compared to real observations using a Dynamic Time Warping algorithm. Finally, isolated word recognition experiments are carried out. The experimental protocol and results are presented in section 3.

2 RANDOM FIELD MODEL

2.1 The model

As mentioned in the introduction, the hidden field X is modeled as a Markov random field since only local interactions exist. Namely,

$$P[X_{t,k} = x_{t,k} | X_{|t,k} = x_{|t,k}] = P[X_{t,k} = x_{t,k} | x(V_{t,k})]$$

where $X_{|t,k}$ denotes the field X without $X_{t,k}$, and $V_{t,k}$ the neighborhood of site (t,k), defined by:

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

$x(V_{t,k})$ denotes the configuration of field X for the neighborhood $V_{t,k}$. The *prior* probability for the field can be expressed in terms of Gibbs field and is given by equation 1 [3].

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

In this equation, $\mathcal{C}$ is the set of cliques defined by the neighborhood system, $U_c(x)$ is a potential function associated to clique c and Z a partition function so that $P[X]$ is a probability measure. A clique is a set of sites mutually neighbors. There are two kinds of cliques associated to neighborhood $V_{t,k}$. The first one, $\{(t-1,k),(t,k)\}$ reflects the *horizontal* interactions, that is the temporal nature of the process, and the second one $\{(t,k),(t,l)\}$ models the interaction between frequency bands. To define a random field model, one must express the potential functions associated with both kind of cliques. It can be shown that a HMM is a particular random field [3, 4], and therefore, the horizontal potential functions $U^{(h)}_{t,k}(x)$ are defined to reflect a Markov chain behavior in each band. For the clique $\{(t-1,k),(t,k)\}$, assuming that $x_{t-1,k}=i$ and $x_{t,k}=j$, we have

$$U^{(h)}_{t,k}(x) = a^{(k)}_{i,j}$$

where $a^{(k)}_{i,j} = -\ln \tilde{a}_{i,j}$ if $\tilde{a}_{i,j}$ denotes the transition probability in Markov chain corresponding to band k. It should be noted that if a transition is not valid (*i.e.* $\tilde{a}_{i,j}=0$), then the energy is infinite and $P[X]$ is null. We therefore have a barrier energy for unauthorized transitions in the Markov chain so that the Hammersley-Clifford theorem is still applicable. Then, the vertical potential functions allow for a control of the synchronization between two bands. The underlying idea is that, if two bands have a synchronous behavior, then the stable spectral zones should occur at the same moments and, therefore, the states should change at about the same time in the corresponding Markov chains. To reflect this statement, the potential function $U^{(v)}_{t,k,l}(x)$ for the clique $\{(t,k),(t,l)\}$ is defined by

$$U^{(v)}_{t,k,l}(x) = f_{k,l}\,|i-j|$$

where $f_{k,l}$ is a synchronization weight between band k and l, if, as previously, $x_{t,k}=i$ and $x_{t,l}=j$. The bigger the weight is, the more synchronous is the behavior of the two sub-bands. The total potential for the *prior* law can now be written, using the horizontal and vertical potentials defined previously, in the following way

$$U(x) = \sum_{t,k} a^{(k)}_{x_{t-1,k},x_{t,k}} + \sum_{t,k,l>k} f_{k,l}\,|x_{t,k}-x_{t,l}| \quad (2)$$

As the RFM focuses on interactions between frequency bands, the observation space is defined by the output of a filter-bank and k is the filter index. We assume the conditional independence of the observations $Y_{t,k}$ which are modeled by a single monodimensional Gaussian.

2.2 Training procedure

A key problem in statistical modeling is the problem of parameter estimation. Unfortunately, no equivalent of the Baum-Welch algorithm is available for maximum likelihood parameter estimation of the RFM. Since the model is strongly related to HMM, heuristic criterions can be used for training purposes. Indeed, the model can be seen as parallel HMMs with the only difference that the behavior of a HMM is influenced by the behavior of the HMMs in the other bands because of the horizontal interactions defined in the RFM. Therefore, the parameters of the HMMs can be estimated independently for each band using the standard Baum-Welch algorithm. Due to the vertical potentials, the coupling between HMMs in bands k and l is based on the measure

$$d(k,l) = \frac{1}{T}\sum_{t=1}^{T} |x_{t,k}-x_{t,l}|$$

If two bands are synchronous, this measure should be small and the synchronization weight $f_{k,l}$ should be big to penalize configurations where $d(k,l)$ is not small. This means that $f_{k,l}$ is inversely proportional to $d(k,l)$. An idea of what $d(k,l)$ should be for a given model can be obtained by computing this measure along the Viterbi path on the training data. If such a measure is denoted $\hat{d}(k,l)$ then $f_{k,l}$ is heuristically defined as:

$$f_{k,l} = \frac{\gamma}{\hat{d}(k,l)}$$

The hyper-parameter γ controls the relative contribution of the horizontal and vertical interactions in the computation of $P[X]$. If γ is set to zero, then the model is simply a multi-band HMM.

3 EXPERIMENTS

3.1 Database and protocol

All the experiments are carried out on the Polyvar database[1] using data from a single speaker. Ten keywords of that database were used since the experiments are based on isolated words. The data were collected over telephone lines on a period of one year. The first fifty sessions are considered as the training corpus and used to estimate the parameters of all the models under consideration. The next twenty sessions are reserved as DTW reference templates. Finally, the next 50 ones are used to carry out some isolated word recognition experiment.

Various model can be trained for each word using the training corpus, each model having two states per phoneme. In order to compare random field modeling of speech with more usual techniques, two HMMs are trained. For the first one, denoted HMM_{cep}, the feature vector consists in 12 cepstral coefficients derived from a 24 channel filter-bank on a linear frequency scale. The second HMM, HMM_{fbk} directly takes as input the output of the filter-bank. In both cases, the probability density functions associated to the states are single diagonal covariance Gaussians. Finally, random field models are also trained for various factors

[1] Distributed by ELRA: *http://www.icp.grenet.fr/ELRA*

γ. It should be noted that the model HMM$_{fbk}$ is, in principle, similar to a fully synchronized RFM, except for the training procedure. As stated before, HMMs are a particular case of the random field model proposed and, therefore, the algorithms used for sampling data from the models and for isolated word recognition experiments are always the same, whatever the model.

3.2 Sampling and comparison

For each of the 10 words and each of the models, 50 samples are drawn using a Gibbs sampler [5, 3]. The principle of the Gibbs sampler is to iterate on every point of the lattice and to randomly choose a value for that point according to the local conditional law. The mean distance between the samples and the observations can be computed using a DTW algorithm and is normalized by the test pattern length. A sample is aligned with the 20 corresponding reference patterns using a Euclidean local distance and the final distance is the smallest of the distances to the reference patterns. Table 1 shows the mean distances for each word-model pair. The last row is the distance averaged over all the words.

	HMM		RFM (γ)			
	cep	fbk	0.0	0.005	0.02	0.05
annulation	4.02	5.69	7.89	6.77	6.60	6.80
casino	3.57	5.58	8.21	7.57	7.49	7.92
cinéma	3.91	5.53	7.06	6.82	6.81	7.05
concert	3.95	6.09	7.55	7.03	7.07	7.15
corso	3.60	5.69	7.71	6.87	6.53	7.24
guide	3.92	6.45	7.62	6.78	6.35	7.55
message	3.87	6.33	7.69	6.90	6.88	7.76
musée	3.84	5.36	7.38	6.78	6.16	7.16
quitter	3.66	5.79	7.39	6.95	6.58	6.03
suivant	3.50	5.49	7.37	6.86	6.74	6.66
average	3.78	5.80	7.59	6.93	6.72	7.13

Table 1: (word,model) DTW mean distance

The distances obtained with filter-bank output feature vectors is always greater than the one obtain with cepstral coefficients. This is due to the fact that we have 12 cepstral coefficients while we have 24 filter-bank outputs. Therefore, the local distance tends to be higher for filter-bank based feature vectors and one must be careful comparing the results obtained with HMM$_{cep}$ with the other results reported here. The figures in this table show that a single HMM is more able to produce realistic samples than a Random Field Model, whatever the value of the hyper-parameter γ. However, the mean distance of 6.72 obtained with $\gamma = 0.02$ indicates that such a model is not very far from *reality*. It is interesting to note that, with RFMs, the minimum of the mean DTW distance for a given word is not always for $\gamma = 0.02$. For example, for the word "*quitter*", the distance is 6.03 for $\gamma = 0.05$ but it is 6.72 for $\gamma = 0.02$. This may be explained by the fact that γ is an experimentally set parameter and it does not depend on the model. A real parameter estimation procedure should avoid such problems by directly finding out the best value of the synchronization parameters, thus giving more realistic samples. In order to be able to compare the HMM$_{cep}$ results with the filter-bank based models, one can normalize the distance by the feature vector dimension. In this case, the model HMM$_{fbk}$ outperforms all the other models and RFMs and HMM$_{cep}$ give similar performances.

3.3 Isolated word recognition

For isolated word recognition, one has to compute $P[Y|W]$, the probability of an observation Y knowing the word W. In the framework of hidden Markov modeling, this probability is approximated with the Viterbi algorithm. With random field models, this probability is approximated using the Iterated Conditional Mode (ICM) algorithm [3]. This algorithm finds out the lattice $\hat{X}$ which maximizes the *posterior* probability $P[X|Y, W]$. Noda *et al.* used this algorithm to make a parallel decoder [4]. The ICM algorithm is somewhat similar to the Gibbs sampler with the difference that, rather than randomly selecting a value for a point of the lattice, it chooses the value for which the local conditional probability is maximum. As it is an iterative algorithm, it converges to a local maximum and strongly depends on the initial conditions. Experiments are carried out with two different strategies for the initialization. The first strategy consists in starting with a uniformly segmented field while the second one consists in running a Viterbi decoding independently in each band and using the Viterbi paths to initialize the field before running the ICM algorithm. For a given utterance Y and a word W, the pseudo log-likelihood of the observation can be computed. The pseudo log-likelihood is the sum of the local log-likelihoods at each point of the lattice and is used instead of the log-likelihood which is intractable because of the partition function Z.

Table 2 gives the recognition rates for the various models. The first row shows the recognition rates when using a uniformly segmented initial field in the ICM algorithm while the second row is obtained using a Viterbi decoder to initialize the field.

	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 2: Isolated word recognition rates (in %)

In both cases (*i.e.* uniform or Viterbi initialization), the recognition error rate increases for RFMs. The results obtained using a uniform initial field are poor, even for the standard HMMs. The ICM algorithm leads to 84.4 % for HMM$_{cep}$ in the first case while the classical Viterbi algorithm gives a rate of 99.8 %. It is to be noted that the Viterbi initialized ICM and the Viterbi algorithm are equivalent for the HMMs and for the RFM with $\gamma = 0.0$. However, it can be seen that the synchronization parameters have some influence on the results. This is clearly shown in the uniform case where the recognition rate for $\gamma = 0.5$ is better than for other values of γ. This trend is less obvious in the Viterbi ICM case. As γ increases, more changes are done by the ICM

algorithm but the energy variations for the hidden field are small compared to the log likelihood of the observations and the solution does not improve a lot.

One interesting point to note is that for RFMs, about half of the errors are due to the same model which is often recognized in place of the correct word. When using a uniform initial field, it is the word "*corso*" which is responsible for most of the errors while, with the Viterbi initialization strategy, it is the word "*annulation*". Moreover, in the latter case, the word "*guide*" is rarely recognized (about 5 times out of 50 occurrences) which was not the case in the former case.

4 DISCUSSION

The model presented in this paper is a preliminary approach toward Markov random field modeling of speech. The experiments using simulations showed that this very simple model is able to generate realistic filter-bank output samples, in terms of distance to real observations. The experiments on isolated word recognition pointed out the weaknesses of this kind of model. The results clearly show that the ICM algorithm highly depends on the initial solution. To get rid of that problem, other decoding strategies, based on simulated annealing [6], which is known to converge to global minima of the field potential, must be investigated.

The interesting point of such a category of models is the simplicity by which the model can be extended, since many potential functions can be envisaged using the same formalism. For example, the understanding of the errors in the isolated word recognition experiments should help the design of more discriminant potential functions. We also believe that the poor recognition results obtained are mainly due to the lack of a real maximum-likelihood parameter estimation algorithm. Indeed, the heuristic used for training the model is based on HMMs independently trained for each band and therefore, the temporal structure of the model does not take into account the frequency interactions. We are currently developing an algorithm for parameter estimation which can be used in any case where the potential function is linear with respect to the parameters to be estimated. The algorithm is a combination of the EM algorithm and of a probabilistic descent to maximize the intermediate quantity [7, 8] of the EM algorithm.

5 CONCLUSION

A new category of statistical models of speech segments, based on Markov random fields, is presented and compared to classical hidden Markov modeling. Experiments on isolated word recognition showed that the current Random Field Model does not yet perform as well as standard HMMs but those preliminary experiments are promising. More work has to be done in order to define a real parameter estimation algorithm and more accurate models. The interesting point of such an approach is that the design of a Random Field Model is rather simple and intuitive. It also defines a large framework for statistical modeling of speech in which current models, such as single and multi-band HMMs, are particular cases.

References

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

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

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

[4] 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.

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

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

[7] 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.

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