A Markov Random Field Model for Automatic Speech Recognition.

Guillaume Gravier, Marc Sigelle and Gérard Chollet
ENST-TSI and CNRS-URA 820
46, rue Barrault F-75634 Paris Cedex 13
{gravier,sigelle,chollet}@tsi.enst.fr

Abstract

Speech can be represented as a time/frequency distribution of energy using a multi-band filter bank. A Markov random field model, which takes into account the possible time asynchrony across the bands, is estimated for each segmental units to be recognized. The law of the speech process is given by a parametric Gibbs distribution and a maximum likelihood parameter estimation algorithm is developed. Experiments are conducted on an isolated word recognition problem. 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 modeling inter-band synchrony is an interesting approach to increase the performance when the number of bands increases.

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. 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 noise. 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 and recently a multi-band approach to speech recognition has been proposed to deal with additive noises [5].

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. 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 [3]. The motivation for this work is to investigate more descriptive statistical models of speech in the t/f domain to increase robustness to noise.

The paper is organized 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

In the multi-band approach of speech recognition, a hidden process (or field) $X = \{X_{t,k}\ t \in [1,T], k \in [1,K]\}$ is associated with the observation $Y = \{y_{t,k}\}$, where $X_{t,k} \in [1,N]$ if N states HMMs are used in each band. In the classical model, the law of X is given by the independent HMMs and the random variable $X_{t,k}$ only depends on $x_{t-1,k}$. It can be shown that this mono-lateral relation has a bilateral equivalence and, in order to model inter-band dependencies, the following neighborhood is considered

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

This neighborhood system defines X as a Markov random field whose distribution is given, according to the Hammersley-Clifford theorem [1], by

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

where $\mathcal{C}$ is the set of all the cliques defined by the neighborhood system and $U_c(x)$ is the potential function associated to the clique c. Two types of cliques, namely horizontal and vertical cliques, are associated to the neighborhood definition (1) and potential functions must be defined for both types of cliques.

Previous studies [6, 8] on random field modeling for speech recognition used the fact that a Markov chain is a particular Gibbs distribution where the potential function associated, in band k, to the clique $\{(t-1,k),(t,k)\}$ is given by

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

In this equation, the function $\delta(x_{t,k} = j)$ equals 1 if $x_{t,k} = j$ and 0 otherwise. The parameters $a^{(k)}_{ij}$, called *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 [7] that the performance is 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 $i \rightarrow j$ observed at the same time in the two bands), then a possible model of the synchrony is given by the clique potential function

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

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. The parameters f_{kl} are called "*synchronization weights*".

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 a Gibbs distribution whose total energy is

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

where

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

counts the number of $i \rightarrow j$ 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 in each band, the likelihood of an observation y knowing $X = x$ is given by a Gibbs distribution whose energy, denoted $U(y|x)$, is the sum over all sites of the opposite of the log-likelihood of $y_{t,k}$ knowing $x_{t,k}$. Finally, it can be shown that the energy of x under the posterior law is given by $U(x|y) = U(x) + U(y|x)$ [2] 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. Maximum likelihood parameter estimation

For K bands, the proposed random field model (RFM) is defined by the set of parameters θ which consists of the $(N \times N)$ matrices $A^{(k)}$ ($k = 1, \ldots, K$) gathering the transition weights and of the $(K \times K)$ synchronization matrix F. Direct maximum likelihood estimation of these parameters from examples is not tractable and we propose a generalized stochastic EM algorithm, where the maximization 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

$$\begin{aligned} Q(\theta,\theta^{(n)}) \ &= \ -\sum_k \sum_{i,j} a^{(k)}_{ij} E_{\theta^{(n)}}[\varphi^{(k)}_{ij}(x)|y] - \\ &\quad \sum_{k,l>k} f_{kl} E_{\theta^{(n)}}[\psi_{kl}(x)|y] - \ln Z_\theta - \\ &\quad \sum_{t,k} \gamma_{t,k}(i) \ \ln g(y_{t,k}; \mu^{(k)}_i, \sigma^{(k)}_i) \qquad (6) \end{aligned}$$

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 $\gamma_{t,k}(i) = P_{\theta^{(n)}}[X_{t,k} = i|y]$. The derivation of (6) w.r.t. $a_{ij}^{(k)}$ leads to the following maximization equation

$$E_{\theta^{(n+1)}}[\varphi_{ij}^{(k)}(x)] - E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)|y] = 0 \ . \qquad (7)$$

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. (7) and we propose to use a single step of a descent algorithm to calculate the new estimation which is then given by

$$a_{ij}^{(k)} \leftarrow a_{ij}^{(k)} + \frac{E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)] - E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)|y]}{V_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)]} \ . \qquad (8)$$

Since none of the expectations involved in (8) can be calculated, they are estimated from samples 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. (8) can not be applied as is 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. this vector. A similar updating equation as (8) is used with the Jacobi matrix to account for the parameter dependencies. 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 but the probabilities $\gamma_{t,k}(i)$ are estimated from samples under the posterior law rather than explicitly calculated.

Since the EM algorithm converges toward a local maximum of the likelihood, an initialization 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 configuration can be determined using the ICM algorithm [2] or simulated annealing. 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

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

The synchronization parameters are not initialized and the pdf parameters are estimated with the feature vectors associated to the considered state.

2.3. Decoding strategies

Finally, a decoding strategy is proposed for isolated word recognition. Since explicit computation of the observation likelihood is impossible because of the sums over all configurations, some approximation must be done. The score of an observation y for a word w, $p_w(y)$, is approximated by

$$p_w(y) = pl_w(x^*) p_w(y|x^*) \qquad (9)$$

where x^* is the most likely posterior configuration and *pl* denotes the pseudo-likelihood [1] of x^*. As for the parameter initialization, x^* can be obtained by the ICM algorithm or by simulated annealing, 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 initialized with a uniform segmentation.

3. Experimental results

3.1. Experimental setup

Experiments are carried out on single-speaker isolated word recognition for telephone speech with a 10 word vocabulary. 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 must be taken with care since the confidence intervals are quite large. However these experiments offer the opportunity to study the proposed model.

When multiple bands are used, the speech signal is divided into sub-bands regularly spread on a MEL scale. 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

The model was first tested using a conventional full-band approach and different decoding strategies based on the ICM and simulated annealing (SA) algorithms were studied. As expected, the ICM based decoder turned to be very sensitive to the initialization and gave poor results when initialized with a uniform segmentation. Comparable results were obtained with the classical Viterbi algorithm and with a simulated annealing based decoder. However, the performance of the simulated annealing decoder highly depends on the choice of the initial temperature and of the speed of the cooling scheme. Part of these results can be seen from table 1 where column b1c12 corresponds to the mono-band case.

Several sub-band divisions were then tested and results for 1, 3, 5 and 7 bands (b1 – b7), with respectively 12, 5, 3 and 2 cepstral coefficients in each band (c12 – c2), are given in table 1. Different training and decoding algorithms were

		sub-band architecture			
decoding	training	b1c12	b3c5	b5c3	b7c2
ICM	heuristic	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
SA	ICM	99.6	97.8	92.6	88.2
	ICM-GEM	–	97.8	95.0	94.2

Table 1. Recognition rate (in %) for different sub-band architectures.

used. The heuristic training corresponds to independently trained parallel HMMs and the estimation of x^* in the decoding stage is obtained by applying the Viterbi algorithm independently in each band as proposed in [4]. It therefore correspond to a standard Viterbi approach (except for the score computation given by (9)) and defines the baseline system. In all other cases, ICM training corresponds to the initialization procedure with x^* obtained using the ICM algorithm while in the ICM-GEM case, re-estimation of the model parameters is performed using the proposed EM procedure.

Those results show 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 heuristic training, compared to 95% with only 2 coefficients. This points out the fact that the recombination is the crucial point of multi-band models. Another explanation is that narrow bands are more variable than wide bands and therefore the statistical model degrades with narrow bands.

With the ICM based decoder, the EM re-estimation of the parameters seems to degrade the results. Except for the mono-band case, the recognition rates for ICM-GEM training along with an ICM based decoder are less than the results obtained solely with the ICM initialization procedure. However, the opposite is observed when a SA based decoder is used. These results stress the weaknesses of the ICM based decoder, in particular when a more complex prior (or regularization) model is used. Indeed, the ICM-GEM trained model is more complex since the synchronization weights have been estimated while they are arbitrarily set to zero in the ICM training procedure. For the SA based decoder, it is observed that re-estimation of the parameters with the EM algorithm (and estimation of the synchronization weights) improves the results when the number of bands increases. Performances comparable (though slightly lower) to the baseline system ones are obtained in this case. This result shows that a good model of the prior process X is needed when the observation becomes more variable. Adding a synchrony model, even if the model is unrealistic, allows for a better regularization of the segmentation and, as a matter of fact, for a higher recognition rate.

In the experiments reported here, standard HMM techniques give slightly better results than the best RFM technique with less computation. However, it must be recalled that the multi-band approach was designed to deal with noisy test data. Preliminary results with noisy test data show that in a 3-band case, the RFM gives better results than a linear combination of the HMM scores in each band.

4. Conclusion

A Markov random field model for speech recognition, based on an extension of the multi-band HMMs, is proposed and studied on an isolated word recognition experiment. When the observation becomes too variable, which is the case when the number of bands is increased, there is a need for regularization of the segmentation and the experiments showed that inter-band synchrony is an interesting and valuable regularization model. To conclude, we should stress the fact that the synchrony modeling used in the current random field model assumes that the same synchronization weight applies for all the duration of a word. This assumption is certainly wrong and may partially explain that independent HMMs perform better. A better modeling of the inter-band synchrony should be therefore be envisaged.

References

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

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

[3] G. Gravier, M. Sigelle, and G. Chollet. Markov random field modeling for speech recognition. *Australian J. Intelligent Inf. Proc. Sys.*, 5(4), 1998.

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

[5] H. Hermansky, M. Pavel, and S. Tibrewala. Towards ASR using partially corrupted speech. In *ICSLP*, 1996.

[6] H. Noda *et al.* A MRF-based parallel processing algorithm for speech recognition using linear predictive HMM. In *ICASSP*, 1994.

[7] M. J. Tomlinson, *et al.* Modelling asynchrony in speech using elementary single-signal decomposition. In *ICASSP*, 1997.

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