IMPROVED ACOUSTIC MODELING FOR SPEECH RECOGNITION USING 2D MARKOV RANDOM FIELDS

Helmut Lucke

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
E-mail: lucke@itl.atr.co.jp

ABSTRACT

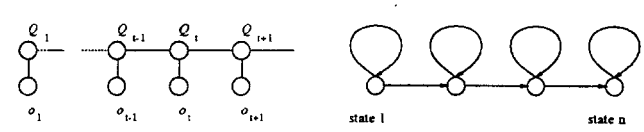
This paper argues that many HMM model inaccuracies are a direct consequence of the fact that the HMM is a one dimensional stochastic model applied to a two dimensional process. Thus we argue that a 2D stochastic process, known as a Random Markov Field (MRF) should perform better. We describe a training method for MRFs and analyze its convergence behavior.

1. INTRODUCTION

Speech recognition is a two dimensional pattern recognition process. All commonly used low level representations of a speech signal (filter bank, FFT, LPC, autocorrelation, cepstrum) are two dimensional; one dimension in the time the other in the 'frequency' domain. (For simplicity we will refer to the non-temporal domain as the 'frequency' domain, even though, depending on the representation, it may not really be frequency. Similarly the components of one observation vector will be referred to as frequency bins even though they may actually have arisen from some other transform such as cepstrum.)

On the other hand Hidden Markov models, which are commonly used as pattern recognizers, are essentially one-dimensional models. Fig. 1(a) makes this clear. Here the nodes labeled o_i are observation vectors and the nodes labeled Q_t are random variables that range over the set of all possible states. The horizontal and vertical lines express the statistical dependencies that are implied by the Markov assumption. This representation of a Markov model is to be clearly distinguished from the much more common one shown in Fig. 1(b) in which the states are thought of as sites to be visited by a stochastic finite state machine. We will not use Fig. 1(b) and only included it here to avoid confusion.

Going back to Fig. 1(a) we can therefore regard the HMM as a one dimensional sequence of random variables that is matched to a two dimensional array of



(a) (b)

Figure 1:

pixels (frequency bins). To perform this matching, a 2D to 1D mapping has to be performed at some level. In the early days this was achieved by vector quantization (VQ), nowadays this mapping is usually performed in a more subtle way using continuous HMMs. Nevertheless a number of modeling errors occur as a direct result of this dimensionality mis-match:

• **Relationship of neighboring components in the frequency domain:** A vector with high energy in bin i should be similar to a vector with high energy in bin $i+1$, all other bins being equal. However the VQ or continuous HMM formulation does not take the ordering of frequency bins into account. Any similarity of neighboring components thus has to be learned from examples during training. This greatly increases the amount of training data required to obtain robust models.

• **Relationship of neighboring components in the time domain:** Even if the neighborhood relationship in the frequency domain can be learned by presenting many examples, we loose modeling accuracy in the temporal domain. Suppose we have learned that for a given phoneme, say /a/, high energy occurs either in bin i or $i+1$. Now, if we observe an /a/ with high energy in bin i in frame t we ought to be able to predict that in frame $t+1$ the high energy also occurs in bin i rather than $i+1$. However with current HMMs this is impossible, for at the state level of the HMM bins i and $i+1$ are essentially 'mapped together' so discrimination between the two is no longer possible.

What is required is a two dimensional stochstic model that takes correlations in both dimension into account. Such models are known as Markov Random Fields (MRFs).

0-7803-2431-5/95 $4.00 © 1995 IEEE

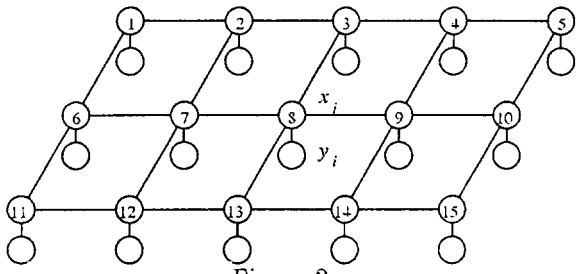


Figure 2:

2. PREVIOUS WORK ON MRFS

MRFs have mainly been studied for image restauration, where algorithms for finding maximum a posteriori probability (MAP) estimates were developed (eg. [4, 8]). These models are parameterized with only a few parameters and some algorithms, exploiting special cases of symmetry, have been proposed for their estimation [1, 5, 6]. These algorithms are not applicable for the pattern recognition tasks required for speech recognition. Zhao and Atlas [9] applied Gibbs distribuitons to speech recognition. However their work considers only a one dimensional MRF, which is equivalent to an HMM, so it does not address the modelling problems of HMMs described above.

3. MARKOV RANDOM FIELDS

In order to improve modeling of the 2 dimensional speech signal we require a two dimensional stochastic process in place of the HMM. Such processes are known as Markov Random Fields (MRFs). A simple MRF is shown in Fig. 2. We have a two-dimensional array of (unobserved) states X_i on top of a 2 dimensional array of observed variables Y_i. The Y_i can be considered as the low level speech representation. Each X_i takes as values one of N discrete states. An MRF requires the definition of a neighborhood relation between the unobserved nodes. In this paper we will just be concerned with the nearest neighbors, which are linked by edges in Fig. 2. Let i be the index of some variable and let N_i be the set of all nearest neighbor indices. The Markov assumption in two dimensions is formulated as:

$$P\left(X_i \middle| \begin{array}{cc} X_j & Y_j \\ \text{all } j \neq i & , \text{ all } j \end{array}\right) = P\left(X_i \middle| \begin{array}{cc} X_j & \\ j \in N_i & , Y_i \end{array}\right), \quad (1)$$

i.e. only the nearest neighbors of X_i affect X_i.

Unfortunately a MRF can not be parameterized by a set of conditional probabilities, as is the case with HMMs and the efficient Baum-Welch algorithm does not exist.

3.1. Parameterization of MRF

It is well known that the joint probability distribution of a MRF with a neighborhood relation as defined above is given by Gibbs distribution:

$$\mathrm{P(X,Y)} = \frac{1}{\mathrm{Z}} \exp\left(\sum_{\mathrm{j}\in\mathrm{N_i}} \mathrm{U_{ij}(X_i, X_j)} + \sum_{\mathrm{i}} \mathrm{V_i(X_i, Y_i)}\right) \quad (2)$$

Here Z is a normalization constant, known as the partition function and U_{ij}, V_i are potential functions. We will assume translational invariance in both the time and frequency domain and hence V_i is independent of i and this index will be dropped from now on. Similarly, under the invariance assumption, there are only two U functions one for horizontal and one for vertical neighbors which we will denote U_H and U_V. Thus the choice of the functions U_H, U_V and V defines the probability distribution of the MRF. Since the state-space is discrete we can parameterize U_H as a matrix of parameters one for each state combination (similarly for U_V). However since Y_i is a continuous variable we parameterize $\exp(V)$ by a set of N Gaussian distributions (one for each state that X_i can assume).

3.2. Calculation of posterior distributions for unobserved variables

This is the problem solved by the forward-backward or Viterbi calculations in HMMs. For MRFs these methods cannot be applied. However a number of well understood algorithms have been developed, the main ones being based on stochastic relaxation [4] and the mean field theory [8].

3.3. Parameter estimation for MRF

This problem is as yet unsolved for general MRFs. The contribution of this paper lies in the provision of a training algorithm for the MRF model described above.

We will describe a method based on the EM-algorithm [3] using ideas from the Re-normalization group theory of statistical physics [2, 7].

The EM algorithm is an iterative parameter estimation procedure in which we first estimate the unobserved data using some previous parameter estimates (E-step) and then calculate the new parameter estimates by finding the ones that maximize the likelihood of the estimated unobserved data (M-step). Denote $\bar{\Theta}$ the old parameters and Θ the new parameters. The E-step amounts to calculating $P_{\bar{\Theta}}(X|Y)$, which we can do using either the stochastic relaxation or mean field methods mentioned above. For the M-step, the new

set of parameterized Gaussians in the $V(\cdot,\cdot)$ potential function can be obtained just as in the continuous HMM case.

The difficulty lies with the estimation of the U_H and U_V matrices. It can be shown that the maximization of the M-step is obtained if we can find Θ such that $P_\Theta(X) = \frac{1}{N_Y}\sum_Y P_{\tilde{\Theta}}(X|Y)$ where the sum is over all N_Y training patterns Y. A method for finding such a Θ will now be described.

Suppose the sites are numbered as in Fig. 2, i.e. such that even and odd numbers form a checkerboard pattern. Then by conditioning on the even numbered sites we can use the Markov assumption to obtain

$$\begin{aligned} P(X) &= P(X_1, X_2, \ldots) \\ &= P(X_2, X_4, \ldots)P(X_1, X_3, \ldots | X_2, X_4, \ldots) \\ &= P(X_2, X_4, \ldots) \times P(X_1|X_2, X_6) \\ &\quad \times P(X_7|X_6, X_2, X_8, X_{12}) \\ &\quad \times P(X_9|X_8, X_4, X_{10}, X_{14}) \ldots \end{aligned} \tag{3}$$

All the even numbered sites form a rectangular lattice (inclined at 45°) of a coarser resolution. We assume that it is Markov, translationally invariant and can be parameterized by the Gibbs distribution

$$P(X_{\text{even}}) = \frac{1}{Z'}\exp\Big(\sum_{\text{up}} U_U(X_i, X_j) + \sum_{\text{down}} U_D(X_i, X_j)\Big)$$

where the first sum is over all pairs of sites in which X_j is to the north-east of X_i and the second sum over all pairs where X_j is to the south-east of X_i.

Suppose that we have already managed to obtain parameter estimates for the U_U and U_D matrices at the coarser resolution. To obtain estimates for U_V and U_H we proceed as follows: Using the most likely state allocation obtained by the stochastic relaxation algorithm we find the statistic $S(k_0, k_W, k_N, k_E, k_S)$ describing the total number of times state k_0 occured surrounded by the states k_W, k_N, k_E, k_S at the nearest neighbor sites. From this after normalization and suitable smoothing we obtain the quantity $Q(k_0|k_W, k_N, k_E, k_S)$ describing the estimated conditional probability of state k_0 occuring in the given context. We now choose U_H and U_V such that

$$\begin{aligned} &\exp\big(\mathrm{U_H(k_W, k_0) + U_H(k_0, k_E) + U_V(k_S, k_0) + U_V(k_0, k_N)} \\ &\mathrm{- \tfrac{1}{2}(U_U(k_W, k_N) + U_U(k_S, k_E) + U_D(k_N, k_E) + U_D(k_W, k_S))\big)} \\ &\approx \mathrm{Q(k_0|k_W, k_N, k_E, k_S)} \end{aligned} \tag{4}$$

If we substitute this expression into equation 3 together with the Gibbs distribution for the even numbered sites, we obtain after some re-arrangement

$$P(X) = \exp\Big(\sum_{\text{hor}} U_H(X_i, X_j) + \sum_{\text{ver}} U_V(X_i, X_j)\Big)$$

i.e. we succeeded in finding suitable U_H and U_V parameters. Moreover these are chosen such that the partition function is unity. Eq. 4 is difficult to solve in closed form. Here we use gradient search technique which minimizes the square of the error.

We have thus reduced the problem of finding U_V and U_H to that of finding U_U and U_D at a coarser level. We can repeat this procedure and reduce to even coarser grids. At the coarsest level that we wish to consider, we have two possibilities to initiate this inductive procedure: (1) If the lattice is so coarse that it constitutes a singly connected line of sites, we may use the Baum-Welch algorithm, as this is a one dimensional model. (2) Alternatively we may assume that there is no longer any statistical dependence between neighboring sites. Under this assumption all sites are to be treated as independent and the parameters of the potential function can easily be found. In the experimental work we used the second approach.

4. PATTERN DISCRIMINATION USING MRFS

There are several ways in which MRFs can be used as pattern discriminators. In this paper, we use a small number of states which are shared by all phoneme models. Each phoneme model has its own U_H and U_V matrices. The model that explains the data with highest likelihood is selected.

5. EXPERIMENTAL WORK

We performed a number of experiments to verify that the parameter estimation procedure described herein works satisfactorily.

For the experiments we used 16 cepstral and delta cepstral coefficients of speech data at a frame rate of 100Hz as the observed data Y. The 'frequency' domain hence consisted of 16 vector valued components, each component being a cepstrum and corresponding delta cepstrum coefficient. $\exp(V)$ was modeled by a 2-variate Gaussians for each state with full covariance matrix.

5.1. Accuracy of equation 4

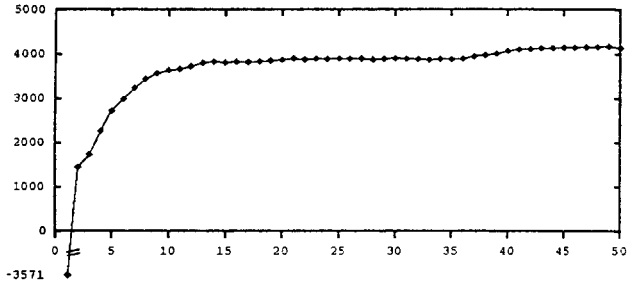
We first investigated how accurately Q could be approximated using the gradient search estimation procedure. For this purpose we trained models of the vowel /a/ with 2,3,4,5,6 states. The average estimation errors for these approximation at various levels of coarseness are given below (the coarseness-level is indicated by the Euclidean distance between the nearest neighbor sites):

	2	3	4	5	6
4	15%	9.2%	4.3%	2.2%	1.2%
2.82	17%	11%	5.0%	2.7%	1.5%
2	15%	12%	7.1%	4.1%	2.4%
1.41	17%	14%	7.7%	4.2%	2.7%
1	16%	12%	9.4%	7.1%	4.6%

The numbers suggest that the Q tensor can indeed be accurately approximated using the gradient search, if a sufficient number of states is used. Since in real applications more than 10 states are likely to be used the estimation error should not be significant.

5.2. Convergence of the parameter estimation procedure

The EM algorithm is guaranteed to converge monotonically. However we are approximating the algorithm in three places: (1) use of stochastic relaxation with a finite cooling schedule, (2) estimation of Q tensor, (3) boundary effects of MRF. Thus the convergence of the algorithm needs to be checked empirically. The following graph shows a plot of the total Gibbs potential of the training data (equivalent to log likelihood) during a training run with 6 states and 50 parameter updates



Even though convergence is not strictly monotonic, due to the approximations mentioned above, we observe a 'fairly monotonic' behavior.

5.3. Discrimination experiments

As further evidence for the utility of the parameter estimation procedure rudimentary discrimination experiments have been carried out with two phoneme models (/a/ and /i/). The two models were trained on 10 examples each and tested on 5. Perfect discrimination was observed in the test.

6. CONCLUSION

A 2 dimensional stochastic process instead of an HMM has the advantage of much more accurately modelling the signal at the expense of more complicated training and decoding schemes. We have provided a workable MRF training algorithm and demonstrated its stability and convergence. Many more engineering choices will have to be made before a working speech recognition system can be realized which utilizes this technology.

7. REFERENCES

[1] N. Balram and J. Moura. Noncausal Gauss Markov random fields: Parameter structure and estimation. *IEEE Transactions on Information Theory*, 39(4), July 1993.

[2] D. Chandler. *Introduction to modern statistical Mechanics.* Oxford University Press, 1987.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Royal Statistical Society Ser. B*, 39:1–38, 1977.

[4] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[5] M. I. Gurelli and L. Onural. On a parameter estimation method for Gibbs-Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(4), Apr. 1993.

[6] G. G. Potamianos and J. K. Goutsias. Partition function estimation of Gibbs random field images using Monte Carlo simulations. *IEEE Transactions on Information Theory*, 39(4), July 1993.

[7] K. G. Wilson. The renormalization group and critical phenomena. *Reviews of Modern Physics*, 55(3), July 1983.

[8] J. Zhang. The mean field theory in EM prodedures for Markov random fields. *IEEE Transactions on Signal Processing*, 40(10), Oct. 1992.

[9] Y. Zhao and L. E. Atlas. Application of the Gibbs distribution to Hidden Markov modelling in speaker independent isolated word recognition. *IEEE Transactions on Signal Processing*, 39(6), June 1991.