DYNAMIC PLANAR WARPING FOR OPTICAL CHARACTER RECOGNITION

Esther Levin and Roberto Pieraccini

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974

ABSTRACT

In this paper we extend the dynamic time warping (DTW) algorithm, widely used in automatic speech recognition (ASR), to a dynamic plane warping (DPW) algorithm, for application in the field of optical character recognition (OCR) or similar applications. Although direct application of the optimality principle reduces the computational complexity somewhat, the dynamic plane warping (or image alignment) problem is exponential in the dimensions of the image. We show that by applying constraints to the image alignment problem, e.g., limiting the class of possible distortions, we can reduce the computational complexity dramatically, and find the optimal solution to the constrained problem in linear time. A statistical model (planar hidden Markov model - PHMM) describing statistical properties of images is then proposed. The PHMM approach was evaluated using a set of isolated, hand-written digits. An overall digit recognition accuracy of 95% was achieved. We expect that the advantage of this approach will be even more significant for harder tasks, such as cursive writing recognition and spotting.

1. INTRODUCTION

This work was motivated by the realization that many open problems in OCR are analogous to known problems in speech recognition. For example: the elastic distortions of hand-written characters are analogous to temporal distortions in speech signals; the segmentation problem in recognition of cursive handwriting is analogous to speech segmentation and endpoint detection in continuous speech recognition; the dependence of the written characters on their adjacent neighbors is similar to the coarticulation effect in speech; the use of basic units (characters) that compose the words is analogous to the use of sub-word phonetic unit modeling in large-vocabulary ASR; the use of grammars for constraining the possible sequences of characters and/or words is important in both applications.

The standard solutions to these alignment problems in ASR, including such methods as template matching[1] or hidden Markov modeling[2] (HMM), are all based on the application of dynamic programming principles. The purpose of dynamic programming time warping is to reduce the intra-class variability of spoken utterances, caused by temporal distortions, by optimally matching them to a template or a model. Generalizing this algorithm to the planar case allows us to approach the above class of OCR problems in a manner similar to the one used for ASR.

2. DPW PROBLEM FORMULATION

The goal of the DPW is to align a 2-dimensional reference image,

$$G_R=\{g_R(x,y): x\in \mathbf{Z}^+, y\in \mathbf{Z}^+, (x,y)\in L_{X_R,Y_R}, g_R(\cdot,\cdot)\in G\subset R^n\}$$

to an elastically distorted test image,

$$G=\{g(x,y): x\in \mathbf{Z}^+, y\in \mathbf{Z}^+, (x,y)\in L_{X,Y}, g(\cdot,\cdot)\in G\subset R^n\}.$$

Here an (x,y) pair describes pixel location by horizontal and vertical coordinates, and $L_{N,M}$ denotes a rectangular discrete lattice, i.e., a set of pixels $L_{N,M}=\{(x,y) \mid 1\le x\le N,\ 1\le y\le M\}$.

The idea of planar warping is to map the test lattice to the reference one through a mapping function F:

$$\begin{bmatrix}\tilde{x}\\ \tilde{y}\end{bmatrix}=F\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}F_x(x,y)\\ F_y(x,y)\end{bmatrix},$$

such that the distortion

$$D(G_R,G)=D=\sum_{x=1}^{X}\sum_{y=1}^{Y} d(g_R(\tilde{x},\tilde{y}),g(x,y)) \qquad \begin{bmatrix}\tilde{x}\\ \tilde{y}\end{bmatrix}=F\begin{bmatrix}x\\ y\end{bmatrix}$$

is minimal, subject to possible constraints like global boundary conditions:

$$F_x(1,y)=1;\ F_x(X,y)=X_R;\ F_y(x,1)=1;\ F_y(x,Y)=Y_R; \qquad (1)$$

and local monotonicity constraints, such as

$$\Delta F_{xx}=F_x(x+1,y)-F_x(x,y)\ge 0;\ \Delta F_{yy}=F_y(x,y+1)-F_y(x,y)\ge 0. \qquad (2)$$

We denote by F the set of all admissible mappings that satisfy the above conditions. Although we limit the discussion in this paper to constraints (1) and (2), the treatment of other kinds of constraints is similar.

3. THE GENERAL APPROACH

The complexity of the problem of finding the optimal warping function is exponential, namely $O((X_RY_R)^{XY})$. This complexity can be reduced, as in the one-dimensional case, by generalizing the optimality principle[3]. We will use the following definitions:

1. Define Θ to be a set of N_T test sub-shapes $\{\theta_n\}$, where each test sub-shape is a set of pixels $\{(x,y)\}$ satisfying the following conditions:

- N_T is polynomial in $\{X,Y\}$,
- $\theta_n \supset \theta_{n-1}$, $1\le n\le N_T$
- $\Delta\theta_n \equiv \{(x,y) \mid (x,y)\in \theta_n \text{ and } (x,y)\notin \theta_{n-1}\}$ has a natural mono-dimensional parametrization,

- $\theta_{N_T} = L_{X,Y}$

where θ_0 is the empty set. In particular, we choose θ to be a set of Y rectangles, $\theta_n = L_{X,n}$, $n=1, \cdots, Y$. In this case $\Delta\theta_n$ are pixels of the n-th row.

2. Define Φ to be a set of admissible warping sequences $\Phi = \{\phi_i \mid 1 \le i \le N_\Phi\}$, where ϕ_i is a sequence of X reference pixels $\phi_i = \{(\tilde{x}_1^i, \tilde{y}_1^i), \ldots, (\tilde{x}_x^i, \tilde{y}_x^i), \ldots, (\tilde{x}_X^i, \tilde{y}_X^i)\}$ that meets the following conditions:

$$\phi_i \subset L_{X_R,Y_R} ; \tilde{x}_1^i = 1 , \tilde{x}_X^i = X_R ; \tilde{x}_{x+1}^i \ge \tilde{x}_x^i , \ 1 \le x \le X .$$

This definition of the set Φ depends on the particular choice of the set Θ and the constraints (1) and (2). Φ is constructed to contain all possible warping sequences of each $\Delta\theta_n$ that satisfy the constraints.

From this definition it is clear that for each i, $1 \le i \le N_\Phi$, and for each n, $2 \le n \le Y-1$, there exists $F \in \mathbf{F}$ such that $\begin{bmatrix} \tilde{x}_x^i \\ \tilde{y}_x^i \end{bmatrix} = F \begin{bmatrix} x \\ n \end{bmatrix}$ for $1 \le x \le X$. Also for each $F \in \mathbf{F}$ and any n, $1 \le n \le Y$, there exists $\phi_i \in \Phi$ such that $\begin{bmatrix} \tilde{x}_x^i \\ \tilde{y}_x^i \end{bmatrix} = F \begin{bmatrix} x \\ n \end{bmatrix}$ for $1 \le x \le X$. The cardinality of Φ is $N_\Phi = O((X_R Y_R)^X)$.

3. Each sequence $\phi_i \in \Phi$ determines a subset $\Lambda_i \subset \Phi$ of sequences

$$\Lambda_i = \{\phi_k : \tilde{y}_x^k \le \tilde{y}_x^i , 1 \le x \le X\}.$$

Whenever we consider ϕ_i to be a candidate warping sequence for the n-th row of the test image, the preceding $(n-1)$-th row can be matched only with a warping sequence in Λ_i in order to meet the vertical monotonicity condition.

4. Denote by $\mathbf{F}_{i,n}$ a set of sub-mapping functions from the n-th test rectangle θ_n, $1 \le n \le N_T$, that satisfy the monotonicity conditions, boundary conditions, and match the n-th row of the test $\Delta\theta_n$ with ϕ_i:

$$\text{for any } F \in \mathbf{F}_{i,n} \quad \begin{bmatrix} \tilde{x}_x^i \\ \tilde{y}_x^i \end{bmatrix} = F \begin{bmatrix} x \\ n \end{bmatrix} \ 1 \le x \le X .$$

Using these definitions we are ready to describe the DPW algorithm.

In the n-th iteration of the algorithm, $2 \le n \le Y$, we assume that the optimal warpings of the $(n-1)$-th rectangle of the test image $g(x,y), (x,y) \in \theta_{n-1}$, that match the $(n-1)$-th test image row $g(x,y), (x,y) \in \Delta\theta_{n-1}$ with the warping sequence ϕ_i are known for $1 \le i \le N_\Phi$. Each optimal warping is defined by a mapping $F_{i,n-1} \in \mathbf{F}_{i,n-1}$ and a distortion $D_{i,n-1}$, such that:

$$D_{i,n-1} = \min_{F \in \mathbf{F}_{i,n-1}} \sum_{x=1}^{X} \sum_{\substack{y=1 \\ \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = F \begin{bmatrix} x \\ y \end{bmatrix}}}^{n-1} d(g_R(\tilde{x},\tilde{y}), g(x,y)) ;$$

$$F_{i,n-1} = arg \min_{F \in \mathbf{F}_{i,n-1}} \sum_{x=1}^{X} \sum_{\substack{y=1 \\ \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = F \begin{bmatrix} x \\ y \end{bmatrix}}}^{n-1} d(g_R(\tilde{x},\tilde{y}), g(x,y)) .$$

Now we can find the optimal warping of the n-th test rectangle, $g(x,y), (x,y) \in \theta_n$, that matches the n-th test image row to to the j-th warping sequence, $g_R(x,y), (x,y) \in \phi_j$:

$$D_{j,n} = \min_{F \in \mathbf{F}_{j,n}} \sum_{x=1}^{X} \sum_{\substack{y=1 \\ \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = F \begin{bmatrix} x \\ y \end{bmatrix}}}^{n} d(g_R(\tilde{x},\tilde{y}), g(x,y)) =$$

$$= \min_{\substack{i \\ \phi_i \in \Lambda_j}} \min_{F \in \mathbf{F}_{i,n-1}} \sum_{x=1}^{X} \sum_{\substack{y=1 \\ \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = F \begin{bmatrix} x \\ y \end{bmatrix}}}^{n-1} d(g_R(\tilde{x},\tilde{y}), g(x,y)) + \sum_{x=1}^{X} d(g_R(\tilde{x}_x^j, \tilde{y}_x^j), g(x,n)) =$$

$$= \min_{\substack{i \\ \phi_i \in \Lambda_j}} D_{i,n-1} + \sum_{x=1}^{X} d(g_R(\tilde{x}_x^j, \tilde{y}_x^j), g(x,n)). \qquad (3)$$

The optimal mapping $F_{j,n}$ is

$$F_{j,n}(x,y) = \begin{cases} F_{i,n-1}(x,y) & \text{for } (x,y) \in \theta_{n-1} \\ (\tilde{x}_x^j, \tilde{y}_x^j) & \text{for } (x,y) \in \Delta\theta_n \end{cases},$$

where i is the argument minimizing (3). Constraining the minimization in (3) only to those i such that $\phi_i \in \Lambda_j$, guarantees that the vertical monotonicity condition is satisfied.

To complete the n-th iteration, the optimal warping of the n-th test rectangle has to be found for every warping sequence $\phi_j \in \Phi$, thus requiring $N_\Phi X$ operations.

The algorithm is initialized for $n=1$ by setting

$$D_{1,i} = \sum_{x=1}^{X} d(g(x,1) g_R(\tilde{x}_x^i, \tilde{y}_x^i)) \, \delta^{-1}(\tilde{y}_x^i - 1),$$

where $\delta(\cdot)$ is the Kronecker delta function. The algorithm is stopped after $n=Y$, when the optimal warpings $F_{i,Y}$ are found for all i for which $\Lambda_i = \Phi$, thereby requiring a total of $O(YXN_\Phi)$ computations. The global optimal warping function $F_{optimal}$ is chosen among these warpings as the one that produces the minimal distortion:

$$F_{optimal} = F_{j,Y} , \ j = arg \min_{i:\Lambda_i = \Phi} D_{i,Y} .$$

4. CONSTRAINING THE WARPING PROBLEM

Even though applying the optimality principle reduces the complexity of the planar warping, the computation is still exponential. Therefore the algorithm is impractical for real-size images Further reduction of the computational complexity can be achieved in two different ways:

1. Finding a sub-optimal solution to the warping problem. Examples of sub-optimal procedures can be found in [4] and [5], where the images are divided into small sub-images for which finding the optimal warping function is possible.

2. Redefining and simplifying the original warping problem. The idea here is to limit the number of admissible warping sequences in Φ in such a way that an optimal solution to the constrained problem can be found in polynomial time. The additional constraints used are not arbitrary, but instead reflect the geometric properties of the specific set of images being compared. For example, we can constrain the possible mappings to only to those for which the vertical distortion is independent of the horizontal position. In this case $N_\Phi = O(X_R^X Y)$, and the admissible warping sequences $\phi \in \Phi$ are naturally grouped into Y_R subsets. The m-th subset, λ_m contains all those sequences ϕ_i for which $\tilde{y}_x^i = m$, $1 \le x \le X$. For all $\phi_i \in \lambda_m$, $\Lambda_i = \bigcup_{k=1}^{m} \lambda_k$, i.e., the satisfaction of the vertical monotonicity condition is independent of a particular

horizontal warping. This allows further reduction of computational complexity as follows. We define

$$\hat{D}_{m,n} \equiv \min_{j:\phi_j \in \lambda_m} D_{j,n}, \quad \hat{F}_{m,n} = F_{i,n}, \tag{4}$$

where i is the argument that minimizes (4). The recursion relation (3) can now be rewritten in terms of these quantities as:

$$\hat{D}_{m,n} \equiv \min_{j:\phi_j \in \lambda_m} \min_{\substack{i \\ \phi_i \in \Lambda_j}} \min_{F \in \mathbb{F}_{i,n-1}} \sum_{x=1}^{X} \sum_{\substack{y=1 \\ \left[\tilde{x} \atop \tilde{y}\right] = F\left[x \atop y\right]}}^{n-1} d(g_R(\tilde{x},\tilde{y}), g(x,y)) +$$

$$+ \min_{j:\phi_j \in \lambda_m} \sum_{x=1}^{X} d(g_R(\tilde{x}_x^j, \tilde{y}_x^j), g(x,n)) = \tag{5}$$

$$= \min_{1 \le k \le m} \hat{D}_{k,n-1} + \Delta D_{m,n}$$

The minimization of the second term in (5), corresponds to a single-dimensional warping (i.e. an algorithm similar to DTW can be used) and requires $O(XX_R)$ calculations. Denote by $f_{m,n}$ the optimal mapping that alignes the n-th row of the test image to the m-th row of the reference image constrained by (1), (2) and minimizing $\Delta Dm,n$. Then

$$\hat{F}_{m,n} = \begin{cases} \hat{F}_{i,n-1} & \text{for } (x,y) \in \theta_{n-1} \\ f_{m,n} & \text{for } (x,y) \in \Delta\theta_n \end{cases},$$

where $i = arg \min_{1 \le k \le m} \hat{D}_{k,n-1}$.

The optimal mapping is $F_{optimal} = \hat{F}_{Y_R,Y}$, and the complexity of its computation is only $O(Y_R Y X_R X)$!

5. PLANAR HMM

In this section we provide a statistical interpretation and generalization of the DPW approach.

The *planar HMM* (PHMM) is a composite source, comprising a set, s, of $N = X_R Y_R$ states $s = \{(\tilde{x},\tilde{y}), 1 \le \tilde{x} \le X_R, 1 \le \tilde{y} \le Y_R\}$. Each state in s is a stochastic source characterized by its probability density $P_{\tilde{x}\tilde{y}}(g)$ over the space of observations $g \in G$. It is convenient to think of the states of the model as being located on a rectangular lattice $L_{X_R Y_R}$, corresponding to the reference lattice of DPW. Similarly to the conventional HMM, only one state is active in the generation of the (x,y)-th image pixel $g(x,y)$. We denote by $s(x,y) \in s$ the active state of the model that generates $g(x,y)$. The joint distribution governing the choice of active states and image values has the following Markovian property:

$$P(g(x,y), s(x,y) \mid g(1{:}X, 1{:}y-1), g(1{:}x-1, y), s(1{:}X, 1{:}y-1), s(1{:}x-1, y)) =$$

$$= P(g(x,y) \mid s(x,y))\, P(s(x,y) \mid s(x-1,y), s(x,y-1)) =$$

$$= P_{s(x,y)}(g(x,y))\, P(s(x,y) \mid s(x-1,y), s(x,y-1)) =$$

where:

$$g(1{:}X, y-1) \equiv \{g(x,y) : (x,y) \in R_{X,y-1}\},$$

$$g(1{:}x-1, y) \equiv \{g(1,y), \cdots, g(x-1,y)\},$$

and $s(1{:}X, 1{:}y-1)$, $s(1{:}x-1, y)$ are the active states involved in generating $g(1{:}X, y-1)$, $g(1{:}x-1, y)$, respectively. The joint likelihood of the image $G = g(1{:}X, 1{:}Y)$ and the state image $S = s(1{:}X, 1{:}Y)$ can be written as

$$P(G,S) = \prod_{x=1}^{X} \prod_{y=1}^{Y} P_{s(x,y)}(g(x,y))$$

$$\pi_{s(1,1)} \prod_{x=2}^{X} a^H_{s(x-1,1),s(x,1)} \prod_{y=2}^{Y} a^V_{s(1,y-1),s(1,y)} \prod_{y=2}^{Y} \prod_{x=2}^{X} A_{s(x-1,y),s(x,y-1),s(x,y)}$$

where

$$A_{(i,j),(k,l),(m,n)} \equiv P(s(x,y) = (m,n) \mid s(x-1,y) = (i,j), s(x,y-1) = (k,l)),$$

$$a^H_{(i,j),(m,n)} \equiv P(s(x,1) = (m,n) \mid s(x-1,1) = (i,j)),$$

$$a^V_{(k,l),(m,n)} \equiv P(s(1,y) = (m,n) \mid s(1,y) = (k,l)),$$

$$\pi_{ij} \equiv P(s(1,1) = (i,j)).$$

The state matrix $\hat{S}$ that best explains the observable G can be estimated by $\hat{S} = arg\max_S P(G,S)$, and then observation likelihood $P(G)$ is approximated as $\hat{P}(G) = P(G,\hat{S})$.

Therefore, the problem of finding $\hat{S}$ and $\hat{P}(G)$ is that of minimizing:

$$L \equiv \sum_{x=1}^{X} \sum_{y=1}^{Y} -\log P_{s(x,y)}(g(x,y)) - \log \pi_{s(1,1)}$$

$$- \sum_{x=2}^{X} \log a^H_{s(x-1,1),s(x,1)} - \sum_{y=2}^{Y} \log a^V_{s(1,y-1),s(1,y)} +$$

$$- \sum_{x=2}^{X} \sum_{y=2}^{Y} A_{s(x-1,y),s(x,y-1),s(x,y)} = D + C$$

over all possible state matrices S. Again, the problem is of exponential complexity, since there are $(X_R Y_R)^{XY}$ different state matrices. This complexity can be reduced by applying the optimality principle and by restricting the model parameters analogously to the way we constrained the DPW problem, as explained in the provious sections.

6. EXPERIMENTAL RESULTS

The PHMM approach was tested on a writer-independent isolated handwritten digit recognition application. The data we used in our experiments was collected from 12 subjects (6 for training and 6 for test). The subjects were each asked to write 10 samples of each digit. Each sample was written in a fixed-size box, therefore the samples were naturally size-normalized and centered. Each sample in the database was represented by a 16×16 binary image.

Each character class (digit) was represented by a single PHMM where the state matrix S was restricted to contain every state of the model, i.e., states could not be skipped. All models had the same number of states. Each state was represented by its own binary probability distribution, i.e., the probability of a pixel being 1 (black) or 0 (white). We estimated these probabilities from the training data with a generalization of the Viterbi training algorithm.[6] The recognition was performed by assigning the test sample to the class k for which $P_k(G)$ was maximal.

Table 1 shows the number of errors in the recognition of the training set and the test set for different sizes of the models. It is worth noting the following two points. First, the test error shows a minimum for $X_R = Y_R = 10$ of 5%. By increasing or decreasing the number of states this error increases. This phenomenon is due to the following:

1. The typical under/over parametrization behavior.
2. Increasing the number of states closer to the size of the

modeled images reduces the flexibility of the alignment procedure, making this a trivial uniform alignment when $X_R = Y_R = 16$.

Also, the training error decreases monotonically with increasing number of states up to $X_R = Y_R = 16$. This is again typical behavior for such systems, since by increasing the number of states, the number of model parameters grows, improving the fit to the training data. But when the number of states equals the dimensions of the sample images, $X_R = Y_R = 16$, there is a sudden significant increase in the training error. This behavior is consistent with point (2) above.

Number of states	Recognition Errors	
$X_R = Y_R$	Training	Test
6	78	82
8	36	50
9	35	48
10	26	32
11	21	38
12	18	42
16	36	64

TABLE 1. Number of errors in the recognition of the training set and the test set for different size of the models (out of 600 trials in both cases)

Figure 1 shows three sets of models with different numbers of states. The states of the models in this figure are represented by squares, where the grey level of the square encodes the probability $P(g=1)$. The (6×6) state models have a very coarse representation of the digits, because the number of states is so small. The (10×10) state models appear much sharper than the (16×16) state models, due to their ability to align the training samples.

7. SUMMARY AND DISCUSSION

In this paper we demonstrated how the DTW algorithm and HMMs, extensively used for speech recognition, can be generalized to OCR. We found two key problems in this generalization:

1. In order to use the optimality principle here, the set of all possible warping sequences satisfying horizontal constraints must be defined. For the n-th row of the test image every such sequence has to be considered as a candidate warping. The vertical constraints are taken into account by limiting the set of possible warping sequences of the previous $(n-1)$-th row. In this way the complexity of computation was reduced from $O(X_R Y_R^{XY})$ to $O(YX\,(Y_R X_R)^X)$.
2. We show that by restricting the original warping problem by limiting the class of possible distortions (for example, assuming that the vertical distortion is independent of the horizontal position), we can reduce the computational complexity dramatically, and find the optimal solution to the restricted problem in linear time. A statistical model (the planar hidden Markov model - PHMM) was developed to provide a probabilistic formulation to the planar warping problem. The restricted formulation of the warping problem corresponds to PHMM with constrained transition probabilities. The PHMM approach was tested on an isolated, hand-written digit recognition application, yielding 95% digit recognition. Further analysis of the results indicate that even in a simple case of isolated characters, the elimination of planar distortions enhances recognition performance significantly. We expect that the advantage of this approach will be even more valuable in harder tasks, such as cursive writing recognition/spotting, for which an effective solution using the current available techniques has not yet been found.

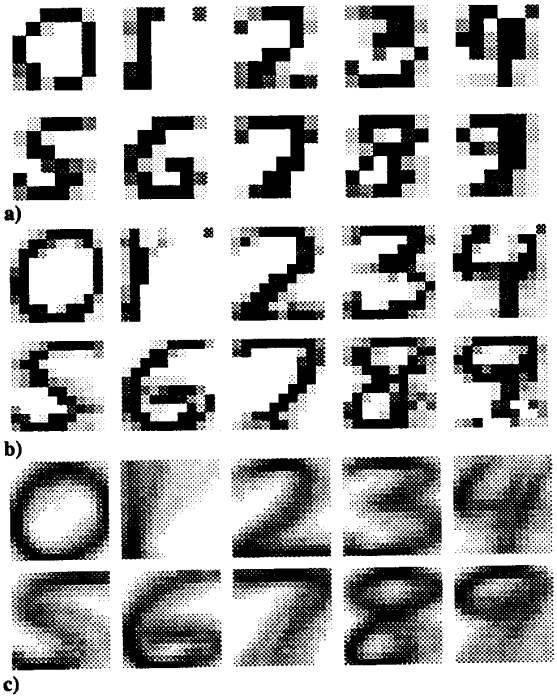

Figure 1. Digit models with a different number of states (a-6, b-10, c-16). The grey level encodes the value of $P(g=1)$ for each state.

REFERENCES

1. L.R. Rabiner, S.E. Levinson, 'Isolated and Connected Word Recognition - Theory and Selected Applications,' *The IEEE Trans. on Communications*, May 1981.
2. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, Feb. 1989.
3. R. Bellman, *Dynamic Programming*," Princeton, NJ: Princeton University Press, 1957
4. R. Chellappa, S. Chatterjee, "Classification of textures Using Gaussian Markov Random Fields," *IEEE Transactions on ASSP* , Vol. 33, No. 4, pp. 959 -963, August 1985.
5. H. Derin, H. Elliot, "Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields," *IEEE Transactions on PAMI* , Vol. 9, No. 1 pp. 39-55, January 1987.
6. F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proceedings of IEEE*, vol. 64, pp. 532-556, April 1976.