Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM

E- Hocine Bourouba, Mouldi Bedda and Rafik Djemili
Department of electronic
Faculty of Engineering, University of Annaba, Algeria
Automatic and Signals Laboratory
E-mail: {Bourouba2004, mouldi_bedda, djemili_r}@yahoo.fr

Keywords: speech recognition, hidden Markov models, dynamic time warping, hybrid system

Received: February 11, 2005

In this paper, we present a new hybrid approach for isolated spoken word recognition using Hidden Markov Model models (HMM) combined with Dynamic time warping (DTW). HMM have been shown to be robust in spoken recognition systems. We propose to extend the HMM method by combining it with the DTW algorithm in order to combine the advantages of these two powerful pattern recognition technique. In this work we do a comparative evaluation between traditional Continuous Hidden Markov Models (GHMM), and the new approach DTW/GHMM. This approach integrates the prototype (word reference template) for each word in the training phase of the Hybrid system. An iterative algorithm based on conventional DTW algorithm and on an averaging technique is used for determined the best prototype during the training phase in order to increase model discrimination. The test phase is identical for the GHMM and DTW/GHMM methods. We evaluate the performance of each system using several different test sets and observe that, the new approach models presented the best results in all cases

Povzetek: V sestavku so opisane hibridne metode za prepoznavanje besed.

1 Introduction

Automatic speech recognition has been an active research topic for more than four decades. With the advent of digital computing and signal processing, the problem of speech recognition was clearly posed and thoroughly studied. These developments were complemented with an increased awareness of the advantages of conversational systems. The range of the possible applications is wide and includes: voice-controlled appliances, fully featured speech-to-text software, automation of operator-assisted services, and voice recognition aids for the handicapped….

Different approaches in speech recognition have been adopted. They can be divided mainly in three trends namely Dynamic Time Warping (DTW), Hidden Markov Models (HMM), and Artificial Neural Networks (ANN).

The introducing of speech HMM has made an impact and has enabled great progress during these last few years. However, there is a lot to be accomplished in this area in order to improve their quality, i.e. the re-enforcing of the discrimination between different models, which seems to be very promising.

In the 1990's, a fourth technique called Hybrid Approach was introduced. The combination of the multiple methods produced a more precise final result because it exploited the advantages of each one. This Combination seems to constitute an interesting approach in speech recognition.

Most the new speech recognition systems are now based on hybrid approach HMM/ANN. HMM has a great capacity to treat events in time, while ANN is an expert in the classification of static forms.

The main solution s suggested to compensate the lack of discrimination in the Markov models come in the model training phase. An alternative approach consists in a local introducing of the discrimination in the model's definition. Among existing methods, the utilizing of ANN's as a discriminating probability estimator has proven to be efficient; nevertheless, it is costly and difficult to put into action. Re-enforcing discrimination techniques between models by a re-estimation of model parameters based on a ANN discriminating criteria are complex, and don't provide a guarantee of convergence for the learning procedure. The approach we propose relies on the principle that the global discrimination between Markov models can be obtained from a discrimination of the models training sequences, and that by a transformation the representing space using the time alignment. Thus, we have developed an iterative algorithm to extract a most suitable prototype favoring the discrimination among the data classes from the training set the derived criteria can be summarized as follows: after the alignment of the sequences of each class by its prototype, each class becomes the most

regrouped possible, and the of classes the most dispersed possible.

The work presented in this paper is an alternative hybrid approach DTW/HMM used in speech recognition using hidden Markov model with DTW algorithm. The goal of this work is to apply DTW to solve the lack of discrimination in the Markov models. A basic idea is that even if DTW has been proven successful in modeling the temporal structure of the speech signal, it is not capable of assimilating a wide variety of speaker dependent spectrum pattern variations; on the other hand, training HMMs for recognizing spoken words is not discriminate So, for example, combining the high time alignment capabilities of DTW with the flexible learning function of the HMM is expected to lead to an advanced recognition model suitable to isolated speech recognition problems.

The new approach GHMM/DTW is introduced, evaluated and compared with traditional approach GHMM for isolated word recognition system. Both these approaches apply the same principles of feature extraction and time-sequence modeling; the principal difference lies in the architecture used for training phases.

The rest of the paper is organized as follows. In the next section, we introduce the acoustic modeling used in our experiments. In section 3 and 4, we discuss some aspects of GHMM and DTW Section 5 then present the existing hybrid system. In section 6 , we will discuss more amply the hybrid approach with the iterative algorithm based the DTW technique . In the next section we presented the experiments examine the performance of GHMM and DTW/GHMM on Arabic and French isolated word. Finally, section IV gives a summary and conclusion.

2 Feature extraction

In this phase speech signal is converted into stream of feature vectors coefficients which contain only that information about given utterance that is important for its correct recognition. An important property of feature extraction is the suppression of information irrelevant for correct classification, such as information about speaker (e.g. fundamental frequency) and information about transmission channel (e.g. characteristic of a microphone). The feature measurements of speech signals are typically extracted using one of the following spectral analysis techniques: MFCC Mel frequency filter bank analyzer, LPC analysis or discrete Fourier transform analysis. Currently the most popular features are Mel frequency cepstral coefficients MFCC [3].

2.1 MFCC Analysis

The Mel-Filter Cepstral Coefficients are extracted from the speech signal as shown in the block diagram of Figure 1. The speech signal is pre-emphasized, framed and then windowed, usually with a Hamming window. Mel-spaced filter banks are then utilized to get the Mel-spectrum. The natural Logarithm is then taken to transform into the Cepstral domain and the Discrete Cosine Transform is finally computed to get the MFCCs. Figure 2 shows the Mel-spaced filter banks that are used to get the Mel-spectrum.

$$C_k=\sum_{i=1}^{N}\log(E_i)\times\cos\left[\frac{\pi k}{N}(i-\frac{1}{2})\right] \quad (1)$$

The following denotes the acronyms used in the block diagram:

- W : Frame Blocking and Windowing
- FFT: Fast Fourier Transform
- LOG: Natural Logarithm
- DCT: Discrete Cosine Transform

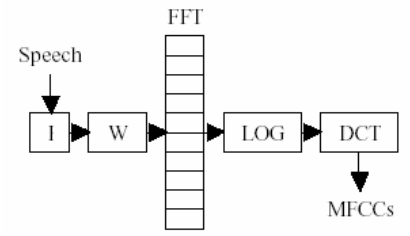

Figure 1: Mel-scale cepstral feature analysis.

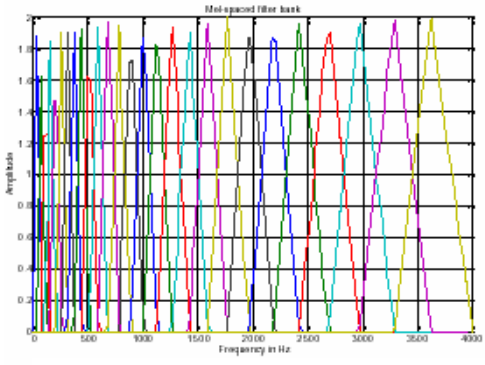

Figure 2: Mel-Spaced Filter Banks.

2.1.1 Pre-emphasis

In general, the digitized speech waveform has a high dynamic range. In order to reduce this range pre-emphasis is applied. By pre-emphasis [1], we imply the application of a high pass filter, which is usually a first - order FIR of the form $H(z)=1-a\times z^{-1}$.

The pre-emphasize is implemented as a fixed-coefficient filter or as an adaptive one, where the coefficient a is adjusted with time according to the autocorrelation values of the speech. The pre-emphasizer has the effect of spectral flattening which renders the signal less susceptible to finite precision effects (such as overflow and underflow) in any subsequent processing of the signal. The selected value for a in our work is 0.9375.

2.1.2 Frame blocking

Since the vocal tract moves mechanically slowly, speech can be assumed to be a random process with slowly varying properties [1]. Hence, the speech is divided into overlapping frames of 20ms every 10ms. The speech signal is assumed to be stationary over each frame and this property will prove useful in the following steps.

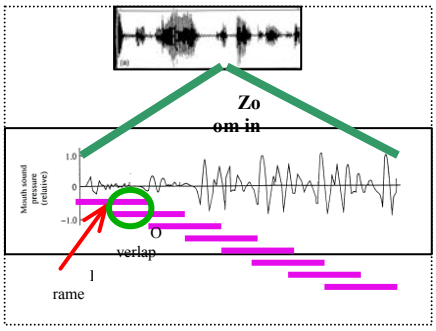

Figure 3 : Frame blocking Step.

2.1.3 Windowing

To minimize the discontinuity of a signal at the beginning and end of each frame, we window each frame frames [1]. The windowing tapers the signal to zero at the beginning and end of each frame. A typical window is the Hamming window of the form:

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad 0\le N\le N-1 \quad (2)$$

3 Hidden Markov model

3.1 Introduction

A Hidden Markov Model (HMM) is a type of stochastic model appropriate for non stationary stochastic sequences, with statistical properties that undergo distinct random transitions among a set of different stationary processes. In other; words, the HMM models a sequence of observations as a piecewise stationary process. Over the past years, Hidden Markov Models have been widely applied in several models like pattern [4,5], or speech recognition [6, 7]. The HMMs are suitable for the classification from one or two dimensional signals and can be used when the information is incomplete or uncertain. To use a HMM, we need a training phase and a test phase. For the training stage, we usually work with the Baum-Welch algorithm to estimate the parameters (Π_i ,A,B) for the HMM [8, 9]. This method is based on the maximum likelihood criterion. To compute the most probable state sequence, the Viterbi algorithm is the most suitable.

3.2 Basic HMM

A HMM model is basically a stochastic finite state automaton, which generates an observation string, that is, the sequence of observation vectors, $\boldsymbol{O}=\boldsymbol{O}_1,..\boldsymbol{O}_t$,… ,$\boldsymbol{O}_T$. Thus, a HMM model consists of a number of N states $S=\{S_i\}$ and of the observation string produced as a result of emitting a vector $\boldsymbol{O}_t$ for each successive transitions from one state S_i to a state S_j. $\boldsymbol{O}_t$ is d dimension and in the discrete case takes its values in a library of M symbols. The state transition probability distribution between state S_i to S_j is $A=\{a_{ij}\}$, and the observation probability distribution of emitting any vector $\boldsymbol{O}_t$ at state Sj is given by $B=\{b_j(\boldsymbol{O}_t)\}$. The probability distribution of initial state is $\Pi=\{\pi_i\}$.

$$a_{ij} = P(q_{t+1} = S_j \,|\, q_t = S_i) \quad (3)$$

$$b_j(\mathrm{O}_t) = P(\mathrm{O}_t \,|\, q_t = S_j) \quad (4)$$

$$\Pi_i = P(q_0 = S_i) \quad (5)$$

Then, given a observation sequence $\boldsymbol{O}$, and a HMM model $\lambda=(A,B,\Pi_i)$, we can compute $P(\boldsymbol{O}|\lambda)$ the probability of the observed sequence by means of the forward-backward procedure [10]. Concisely, the forward variable is defined as the probability of the partial observation sequence $\boldsymbol{O}_1,\boldsymbol{O}_2$,…,$\boldsymbol{O}_t$ (until time t) and state S_i at time t, with the model λ, as $\alpha_t(i)$. And the backward variable is defined as the probability of the partial observation sequence form $t+1$ to the end, given state S_i at time t and the model λ, as $\beta_t(i)$. The probability of the observation sequence is calculated as:

$$P(\mathrm{O}|\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) = \sum_{i=1}^{N} \alpha_T(i) \quad (6)$$

and the probability of being in state i at time t, given the observation sequence $\boldsymbol{O}$, and the model λ, as:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(\mathrm{O}|\lambda)} \quad (7)$$

The ergodic or fully connected HMM is a HMM with all states linked all together (every state can be reached from any state). The left-right (also called Bakis) is an HMM with the matrix transition defined as:

$$a_{ij} = 0 \quad if \quad j < i$$

$$a_{ij} = 0 \quad if \quad j < i + \Delta \quad (8)$$

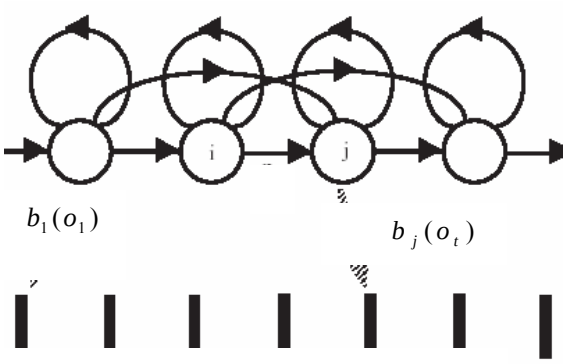

Figure 4: A Bakis (or left rigth) HMM.

We adjust the model parameter $\lambda=(A,B,\Pi_i)$ to maximize the probability of the observation sequence. Consequently, given W classes to recognize, we need to train λ^w for w=1...W HMM, one for each class, with the data set corresponding to the class w. We accomplish the above task using the iterative Baum-Welch method, which is equivalent to the EM (Expectation-Modification) procedure.

The Baum-Welch method, developed in this work as follows:

1. Estimate an initial HMM model as $\lambda=(A,B,\Pi)$.

2. Given λ and the observation sequence $\boldsymbol{O}$, we calculate a new model) , $\overline{\lambda}=(\overline{\mathrm{A}},\overline{\mathrm{B}},\overline{\Pi}_i)$ such as:

$$P(\mathrm{O}|\overline{\lambda})>P(\mathrm{O}|\lambda) \quad (9)$$

3. If the improvement

$$\frac{P(\mathrm{O}|\overline{\lambda})-P(\mathrm{O}|\lambda)}{P(\mathrm{O}|\overline{\lambda})}<threshold \quad (10)$$

then stop, otherwise put $\overline{\lambda}$ instead of λ and go to step 1.

In the GHMM case a gaussian mixtures density is a weighted sum of M component densities, given by the equation

$$b(0_t/\lambda)=\sum_{i=1}^{M} w_i b_j(0_t) \quad (11)$$

where $o_t\ (t=1.....T)$ is a D-dimensional random vector, $b_i(o_t), i=1.....,M$, are the component densities and $w_i, i=1.....,M$, are the mixture weights. Each component density is a D-variate gaussian function of the form

$$b_i(o_t)=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{\frac{1}{2}(o_t-\mu_i)^t\,\Sigma_i^{-1}(o_t-\mu_i)\right\} \quad (12)$$

ith mean vector μ_i and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} w_i=1$. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights form all component densities. These parameters are collectively represented by the notation $\lambda=\{w_i,\mu_i,\Sigma_i\}i=1.....,M$. The GMM can have the several different forms depending on the choice of covariance. The model can have full or diagonal matrix. In this paper the full and diagonal covariance matrix are used for word recognition. In the GHMM, the Baum-Welch algorithm estimates the means and variances for the mixture of Gaussians

$$\overline{c}_{jk}=\frac{\sum_{t=1}^{T-1}\gamma_t(j,k)}{\sum_{t=1}^{T}\sum_{t=1}^{M}\gamma_t(j,k)} \quad (13)$$

$$\overline{\mu}_{jk}=\frac{\sum_{t=1}^{T-1}\gamma_t(j,k)\times 0_t}{\sum_{i=1}^{T}\gamma_t(j,k)} \quad (14)$$

$$\overline{U}_{jk}=\frac{\sum_{t=1}^{T-1}\gamma_t(j,k)(0_t-\mu_{jk})(0_t-\mu_{jk})^t}{\sum_{i=1}^{T}\gamma_t(j,k)} \quad (15)$$

$$\gamma_t(j,k)=\left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{t=1}^{N}\alpha_t(j)\beta_t(j)}\right]\left[\frac{c_{jk}\mathrm{N}(0_t,\mu,U)}{\sum_{m=1}^{M}c_{jm}\mathrm{N}(0_t,\mu,U)}\right] \quad (16)$$

The Viterbi algorithm can be used to obtain the estimation of the most probable state sequence. Once all the HMMs $(\lambda^1,\lambda^2,.....\lambda^W)$ are correctly trained, to classify a sequence for the observation $\boldsymbol{O}$, $P_w=P(O|\lambda^w)$ is calculated for all the λ^w. The unknown observation $\boldsymbol{O}$ is then classified by the process:

$$w^*=\arg\max_{1\le w\le W}(p^w) \quad (17)$$

And so, w* is the optimum class for the observation $\boldsymbol{O}$.

The initialization and stop criteria must be chosen adequately for the HMM. It directly interacts on the relevancy of the HMM [11]. Equiprobable and equal occupancy methods for the initial models are provided as well as iteration and rate of the error for the stop criterion.

4 Dynamic time warping dynamic

The Dynamic Time Warping (DTW) distance measure is a technique that has long been known in speech recognition community. It allows a non-linear mapping of one signal to another by minimizing the distance between the two.

Dynamic Time Warping is a pattern matching algorithm with a non-linear time normalization effect. It is based on Bellman's principle of optimality [12], which implies that, given an optimal path w from A to B and a point C lying somewhere on this path, the path segments AC and CB are optimal paths from A to C and from C to B respectively. The dynamic time warping algorithm [12] creates an alignment between two sequences of feature vectors, (T_1, T_2,....T_N) and (S_1, S_2,....,S_M).

A distance d(i, j) can be evaluated between any two feature vectors T_i and S_j . This distance is referred to as

the local distance. In DTW the global distance D(i,j) of any two feature vectors T_i and S_j is computed recursively by adding its local distance d(i,j) to the evaluated global distance for the best predecessor. The best predecessor is the one that gives the minimum global distance D(i,j) at row i and column j:

$$D(i,j) = \min_{m \le i, k \le j} [D(m,k)] + d(i,j) \quad (18)$$

The computational complexity can be reduced by imposing constraints that prevent the selection of sequences that cannot be optimal [13]. Global constraints affect the maximal overall stretching or compression. Local constraints affect the set of predecessors from which the best predecessor is chosen. Dynamic Time Warping (DTW) is used to establish a time scale alignment between two patterns. It results in a time warping vector w, describing the time alignment of segments of the two signals. assigns a certain segment of the source signal to each of a set of regularly spaced synthesis instants in the target signal.

5 Overview of hybrid system in speech recognition

In order to overcome the unsatisfying performance of speech recognition systems based DTW, HMM or ANN, researchers have attempted to combine these methods. The majority of the researchers combine the models of Markov hidden HMM with the networks of neurons ANN. Several researchers have explored some hybrid system of HMMs and neural networks, the majority are constructed by sending the output of a neural networks to a HMM post processor [14,17] , several others propose a NN architecture that can emulated a HMM[18], alternatively [19] uses the NN to restore the N-best hypotheses produced with a HMM. In [14,16,20,21] the outputs of the NN are not interpreted as probabilities, but rather are used as scores and generally combined with dynamic programming. In [21,24] a network per class or per state is trained to predict the next input frame given only a few previous frames. Amore recent hybrid predictive system is proposed in [25], where network per word vocabulary is created and trained to predict the next input frame given the previous one, the predicted errors summed over all frames are used as a recognition score. In [26] we propose a method which extends the VQ distortion method by combining it with the likelihood of the sequence of VQ indices against a discrete hidden Markov model (DHMM). The scores have to be combined in such a way that the coherence of the two sources is maximized and their differences minimized.

In [27] we combine Hidden Markov Models of various topologies and Nearest Neighbor classification techniques using DTW algorithm.

6 New system DTW/GHMM

We combine HMM and DTW in a modeling framework. HMM can capture the statistical characteristics of word and subword units among different speakers even in large vocabulary and thus is generally better than DTW in speaker independent large vocabulary speech recognition. However, there are useful applications of DTW in small vocabulary, isolated word, speaker dependent or multi-speaker speech recognition due to its relative simplicity and good recognition performance in these situations. DTW system can capture long-rang dependencies [1] in acoustic data, and can potentially adapt to differences in speaker, and accent [2]. According to the analysis above, DTW is effective in wide scale observation and HMM is suitable for solving the analysis of the detail. Hence it is feasible to devise a recognition system which combines these two methods.

The idea is to generate reference patterns for the words in the recognition vocabulary based on training data and then to aling all training data with them. The iterative algorithm was able to find a best reference templete that obtained over significant differences between training sets.

In the traditional HMM system each word is represented by a distinct HMM. In the training stage, each utterance is converted to the cepstral domain (MFCC features, energy, and first and second order deltas) which constitutes an observation sequence for the estimation of the HMM parameters associated to the respective word. The estimation is performed by optimizing the likelihood of the training vectors corresponding to each word in the vocabulary. Typically, the optimization is performed using the Baum-Welch algorithm or equivalently the EM (Expectation-Maximization) algorithm [1]. In the recognition stage, the observation sequence representing the word to be recognized is used to compute the likelihoods, for all possible models, that the sequence has been generated by these models. The recognized word corresponds to the one associated to the model with the highest likelihood. In this stage the Viterbi algorithm, is employed.

In the new system same steps are used as traditional system. In the training stage, feature vectors corresponding to the data samples are processed in order to generate a prototype pattern vector for each word. This is done by computing the centroid of the feature vectors associated to all the training occurrences of each word and then is used to normalize the training set using DTW algorithm to produce a stable set of clusters for which, σ, the ratio of average intercluster distance to average intracluster distance was maximized.

In the recognition stage don't need to align recognized utterance by prototypes, because it's time consuming and return the recognition step very complex.

In the hybrid approach, HMM model for each digit, is generated as follows:

1- Calculate the prototype template from the training data by iterative algorithm.

2- Use the Dynamic Time Warping technique to align all the training data with the prototype template.

3- Once he training data are aligned, then used to Train HMM model using Baum welsh algorithm.

According to description above, the architecture of the traditional and new system are shown on the following figures (5,6).

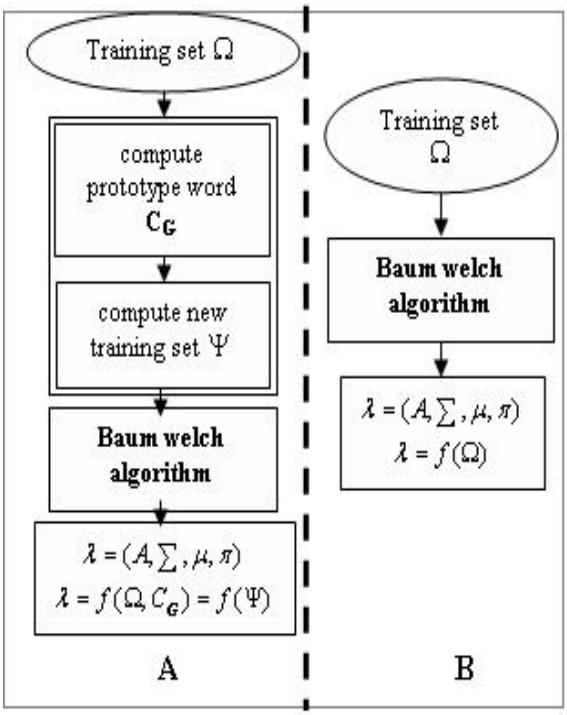

Figure 5: Training Step.
A- Hybrid system.
B- Baseline system.

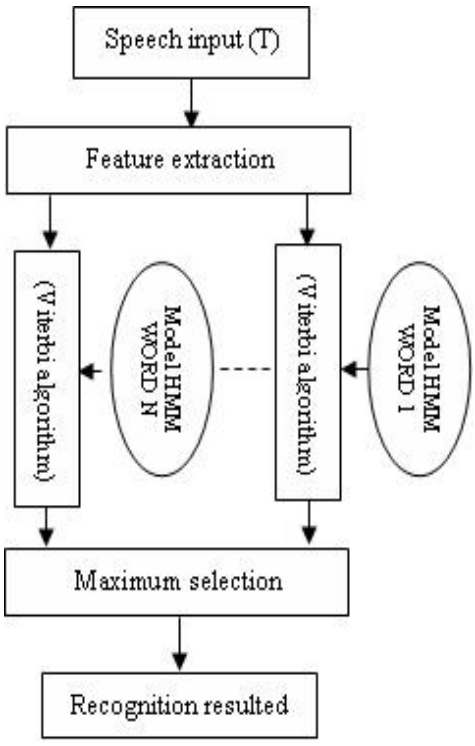

Figure 6 : Recognition Step.

7 New system DTW/GHMM

The choice of the prototype templates will affect the performance of the recognition process. Two methods commonly used are to choose the cluster member that minimizes the distance to all other members of the cluster, or to simply average the members of the cluster. The advantage of the latter method is that it smooths out noise that may be present in any individual data item. Unfortunately, it is only workable when the cluster elements are embedded in a metric space (e.g. Cartesian space). Although we cannot embed cluster elements in a metric space, DTW allows us to use a combination of the two methods. The details of the algorithm are now presented as an iterative algorithm:

1- First, we select the utterance from the training data that minimizes distance to all other utterances in a given cluster.

2- Then we warp all other patterns into that centroid, resulting in a set of patterns that are all on the same time scale.

3- It is then a simple matter to take the average value at each time point over all of the series and use the result as the cluster prototype.

7.1 Introduction

We assume that L finite sets Ω_ℓ are given, with N_ℓ patterns each (repetitions of the same word). The set of training data is $\Omega = \bigcup_{\ell=1}^{L} \Omega_\ell$.

Where $\Omega_\ell = \{y_{\ell,1}, y_{\ell,2}, \ldots\ldots\ldots y_{\ell,N_\ell}\}$, L is the number of words in vocabulary. $y_{\ell,i}$ is a pattern representing the i^{th} repetition of the ℓ^{th} word. A pattern y is assumed to consist of F frames with P features each. If we denote the i^{th} frame of y as y(i), then we can represent y as the set of vectors $y = \{y(1), y(2)\ldots\ldots\ldots, y(F)\}$.

The MFCC feature, y(i), are computed from the MFCC coefficients by the relation (1). Since the iterative algorithm is based on distance data, a distance d_{ij} between patterns x_i and y_j and warping function are computed by :

$$[d_{ij}, w(t)] = DTW(x_i, y_j) \quad (19)$$

Where $x_i(t) = y_j(w(t))$

The function w(t) is the warping function obtained from a dynamic time warping (DTW) match of pattern x_i to y_j, which minimizes the total distance over a constrained set of possible w(t).

A flow diagram of the iterative algorithm is given in figure 07

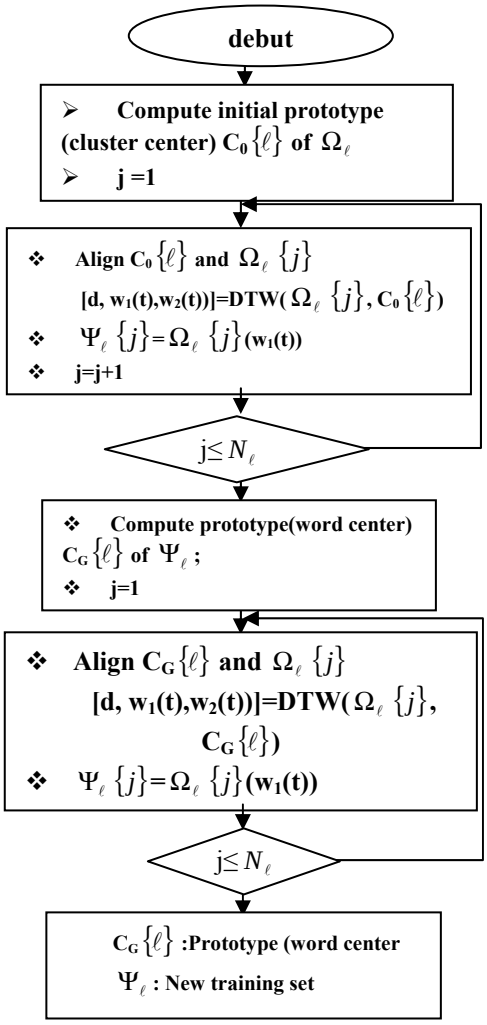

Figure 7: Steps of iterative algorithm.

7.2 Proposed Iterative Algorithm

Assume that for Ω_ℓ ℓ=1,2….,L ,sets (word patterns), the raw data $y_{\ell,i}=\Omega_\ell\{i\}$, i=1,2,…….. N_ℓ are to be aligned with the center cluster C{ ℓ } to product new training set Ψ_ℓ . With the above definitions the proposed algorithm is described as follows:

7.2.1 Determination of the minmax center $C_0\{\ell\}$ of the observation set Ω_ℓ :

For each set Ω_ℓ :

Compute a matrix of distance D:

$$[D_\ell(i,j),\mathrm{w(t)}]=\mathbf{DTW}(\Omega_\ell\{i\},\Omega_\ell\{j\}) \quad (20)$$

Compute cluster center $C_0\{\ell\}$ using :

$$C_0\{\ell\}\equiv\Omega_\ell\{i\}\text{if} \max_{1\le m\le N_\ell} D(i,m)\text{is m}in \quad (21)$$

$C_0\{\ell\}$ is the word $y_{\ell,i}$ such that the maximum distance to any another word in Ω_ℓ is minimum. Since all distances of any word in Ω_ℓ are computed and stored in D, minimax computations of the type given in Eq. (21) are especially simple to implement. These steps are given in Figure 08

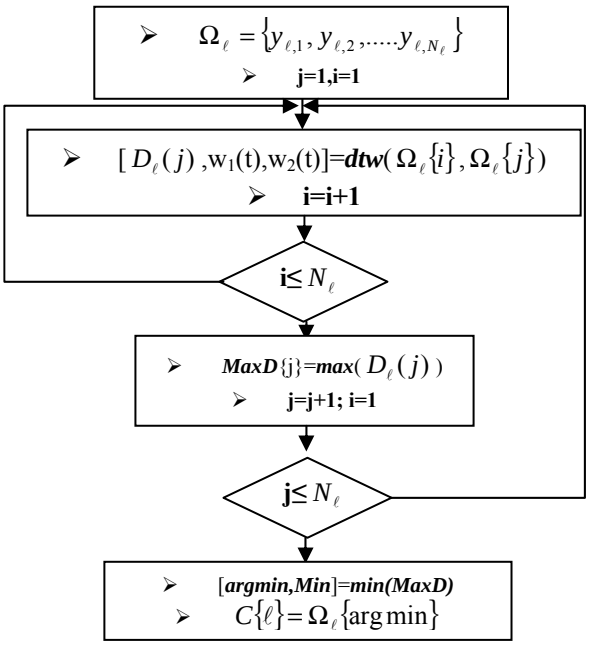

Figure 8: The first step in the iterative algorithm.

7.2.2 Compute the new training set Ψ_ℓ

Align the patterns in each cluster Ω_ℓ , to the length of the prototype $C_0\{\ell\}$. Replace all patterns $y_{\ell,i}$ i=1,2… N_ℓ , with the corresponding warped patterns $\tilde{y}_{\ell,i}$ using

$$\mathbf{[d, w(t)]=DTW(}\Omega_\ell\{j\}\mathbf{, C_0}\{\ell\}\mathbf{)} \quad \mathbf{(22)}$$

$\tilde{y}_{\ell,i}=y_{\ell,i}(w(t))$, t=1,…… F_ℓ (F_ℓ is the frames number of $C_0\{\ell\}$). Where $w(t)$ is vector contain the indices frame which $y_{\ell,i}=C_0\{\ell\}$. So that in each set, the time length (number of frames) of all patterns become equal. Therefore Ψ_ℓ is the new set:

$$\Psi_\ell=\{\tilde{y}_{\ell,1},\tilde{y}_{\ell,2},.....,\tilde{y}_{\ell,N_\ell}\}\ ,\ \Psi_\ell\{j\}=\tilde{y}_{\ell,i} \quad (23).$$

7.2.3 Compute the prototype cluster

Compute the Prototype cluster (cluster center) of the entire patters set Ψ_ℓ . The cluster center is computed by averaging:

$$C_G(\ell)=\frac{1}{N_\ell}\sum_{i=1}^{N_\ell}\tilde{y}_{li} \quad (24)$$

7.2.4 Recomputed the new training set Ψ_ℓ

Again align the patterns in each cluster Ω_ℓ, to the length of the prototype $C_G\{\ell\}$. Replace all pattern $y_{\ell,i}$ i=1,2… N_ℓ, with the corresponding warped patterns $x_{\ell,i}$ using

$$\mathbf{[d, w(t)]=DTW(}\, y_{\ell,i}\,\mathbf{,}\, C_G\{\ell\}\mathbf{)} \quad (25)$$

$x_{\ell,i}=y_{\ell,i}(w(t))$, t=1,…… F_ℓ (F_ℓ is the frames number of $C_G\{\ell\}$). Where $w(t)$ is vector contain the indices frame which $y_{\ell,i}=C_G\{\ell\}$. Therefore Ψ_ℓ is the new set:

$$\Psi_\ell=\{x_{\ell,1},x_{\ell,2},.....,x_{\ell,N_\ell}\} \quad (23).$$

8 Experimental Evaluation

This section presents the experimental evaluation of GHMM and DTW/GHMM approaches for spoken word recognition. Two databases (French and Arabic) were used for the training and testing. The first database is the Digits Corpus from the National Laboratory of Automatic and Signals in The University BADJI-MOKHTAR Annaba Algeria. The data is sampled at 10 KHZ sampling rate and digitized to 8-bit resolution. A subset of the database used in our experiments comprised a small vocabulary spoken by 10 speakers (8 males and 2 females) and test data spoken by 15 different speakers (11 males and 4 females). There are utterances 300 in the training sequence and 600(400 (4 for each training speaker) + 200(4 for another speaker)) in the testing sequence. The second database comprises 48 isolated Arabic words is sampled at 10 Hz and digitized to 8-bit resolution. Here we used only a subset of 10 words. There are 10 speakers (2 male and 8 female) in the database and each word was repeated 5 times by the Speakers. The three first one repetitions were used as the training set and the rest as the testing set. There are 1440 utterances in the training sequence and 960 in the testing sequence. The feature extraction procedure for both databases is the same.

The vocabulary to be recognized is composed by the ten French utterances of the digits from zero to nine and ten Arabic utterances of the states name in Algeria as following:

	Vocabulary
V1	(1-10) digits in French
V2	*(Adrar, Chlef, Laghouat, Oum el bouaghi, Batna, Béjaia, Biskra, Béchar, Blida, Bouira)*
V3	(*Tamanrasset, Tébéssa, Tlemcen, Tiaret, Tizi ozou, Alger, Djelfa, Jijel, Sétif, Saida*
V4	(*Skikda, Sidi Bel Abbes, Annaba, Geulma, Constantine, Médéa, Mostaghanem, Msila, Mascara, Ouargla)*

Notice: The words signals were recorded in a room without any special acoustic protection. Repetitions from one speaker were done in different days with a different type of microphone).

For comparison purposes, we have been using systems based on different kinds of acoustic feature:

D1=(12 MFCC)

D2=(12 MFCC + E)

D3 = ((12 MFCC + E)+ Δ

D4 = ((12 MFCC + E)+ Δ+ ΔΔ

8.1 Results and Discussion

The tables below shows the various results obtained for the two developed systems of traditional GHMM recognition and hybrid (DTW/GHMM) applied to the different vocabularies:

Table1: the tests results with 1 mixture and 5 stats

	Traditional system			
	D1	D2	D3	D4
V1	75.00	80.25	85.00	86.25
V2	56.66	58.66	65.33	67.33
V3	75.33	81.33	84.00	84.66
V4	70.00	76.00	79.33	82.00
V5	78.00	79.33	84.60	86.66

Table2: the tests results with 1 mixture and 5 stats				
	Hybrid system			
	D1	D2	D3	D4
V1	77.00	91.25	92.25	93.50
V2	60.00	65.33	76.66	83.33
V3	83.33	86.00	86.66	88.33
V4	76.66	80.00	84.60	85.33
V5	82.00	85.33	86.00	90.00

Table 1 and 2 report respectively test results of conventional HMMs and DTW/HMM algorithm, where HMMs have 5 states and 1 mixture components for different vector coefficients (D1…D4) and vocabulary's(V1….V5).

The variation of performance raised about 2-10 % between the system GHMM and GHMM/DTW are observed for the registered test set. From the experiments above, we know that DTW/GHMM has better performance than conventional HMMs.

The main advantages of our method are the following:

❖ Our experiments show, that the alignment of two sequences of the same word with respect to its class prototype result in a decrease of the distance between the two sequences before being aligned (figure 09).

Let x_1 and $x_2 \in \Omega_\ell$

d1<d2

d1=DTW(x_1 , x_2 **)**

d2=DTW($\tilde{x}_1$, $\tilde{x}_2$ **)** with $\{[d, w_1(t)] = dtw(x_1, C_G(\ell)\}$

$\tilde{x}_1 = x_1(w_1(t))$

$\{[d, w_2(t)] = dtw(x_2, C_G(\ell)\}$

$\tilde{x}_2 = x_2(w_2(t))$

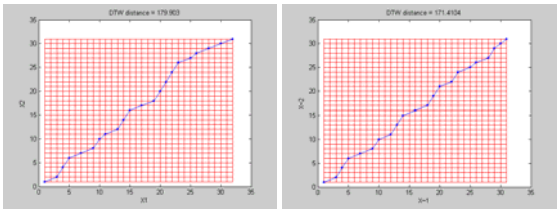

Figure 9: distance and time warping function of two utterances of the same word before and after alignment

❖ On the other hand the fact of aligning two different word sequences with their respective class prototype has the effect of increasing the distance between the two sequences after being aligned (figure 10).

Let $x \in \Omega_{\ell 1}$ and $y \in \Omega_{\ell 2}$

d1>d2

d1=DTW(x , y **)**

d2=DTW($\tilde{x}$, $\tilde{y}$ **)** with $\{[d, w_1(t)] = dtw(x, C_G(\ell_1)\}$

$\tilde{x} = x(w_1(t))$

$\{[d, w_2(t)] = dtw(y, C_G(\ell_2)\}$

$\tilde{y} = y(w_2(t))$

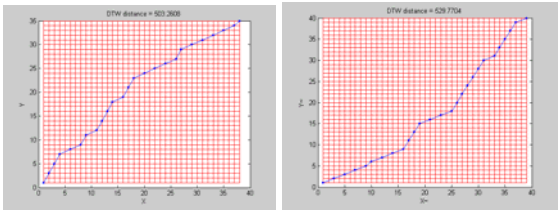

Figure 10: distance and time warping function of two different patterns before and after alignment

❖ Combing the two previous properties results increase of intra-cluster correlation and a good inter cluster discrimination. In addition the time alignment of two training sets X1 and X2 representing two different words produces two new training sets Y1 and Y2 that are easily discriminated. As a consequence Y1 and Y2 are more effective differentiating HMM models λ_1, λ_2 .

Time-warping all the utterances in the training set (cluster) to the same duration as a central template is used to improve the training process. The time-normalized utterances improve the ability of the baum welsh algorithms to learn the data, because the average intercluster distance to the average intracluster distance is maximized after alignment of the training sequences with respect to the prototype, this favors the discrimination of models training sequences which result a discrimination of the models. On the other hand, the fact the classes have their own sizes after the alignment will increase the discrimination of the models particularly at the transition matrix A level. This is especially important for words which are phonetically close to each other.

9 Conclusion

This paper presents the new DTW/GHMM system in isolated speech, where classical DTW and HMM is combined. In the training stage we define the prototype set and introduce iterative algorithm as the solution to build the best prototypes which favors the discrimination between the training sets to give discriminates models in the vocabulary space. The experiments show that the DTW/GHMM system increases the average recognition rate by 2-10% more than the HMM-based recognition method. Though the methods proposed in this paper got better performance, there are still some issues to be further investigated. If explicit effective features can be extracted, the recognition may have a better performance. It is a challenging issue that deserves further study.

References

[1] L. Rabiner, and B.H.Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[2] M. De Wachter, K. Demuynck, "Data driven example based continuous speech recognition" in Proc. Eurospeech, Geneva, Switzerland, Semptember 2003

[3] Q. Zhu, A. Alwan, "On the use of variable frame rate analysis in speech recognition", *Proc. IEEE ICASSP, Turkey, Vol. III, p. 1783-1786, June 2000.*

[4] J. A. Sánchez, C. M. Travieso, I. G. Alonso, M. A.Ferrer, *Handwritten recognizer by its envelope and strokes layout using HMM's*, 35rd *Annual 2001 IEEE Internacional Carnahan Conference on Security Technology, (IEEE ICCST'01)*, London, UK, 2001, 267-271.

[5] M. A. Ferrer, J. L. Camino, C. M. Travieso, C. Morales, *Signature Classification by Hidden Markov Model*, *33rd Anual 1999 IEEE Internacional Carnahan Conference on Security Technology, (IEEE ICCST'99), Comisaría General de Policía Científica, Ministerio del Interior, IEEE Spain Section, COIT, SSR-UPM, Seguritas Seguridad España S.A*, Madrid, Spain, Oct. 1999, 481-484.

[6] Renals, S., Morgan, N., Bourlard, H., Cohen, M. & Franco, H. (1994), Connectionist probability estimators in HMM speech recognition, *IEEE Transactions on Speech and Audio Processing 2(1)*,1994, 161-174.

[7] L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer, Maximum mutual information estimation of HMM parameters for speech recognition,. *In Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing,* , Tokyo, Japan, December 1986,49-52

[8] L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probablistic functions of Markov chains. *The Annals of Mathematical* Statistics, *41(1),* 1970, 164-171.

[9] L. Baum, An inequality and associated maximization technique in statistical estimation for probalistic functions of Markov processes. *Inequalities*, *3*, 1972, 1-8.

[10] L. R. Rabiner. Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition *Readings in Speech Recognition, chapter A*, 1989,267-295.

[11] M.A. Ferrer, I. Alonso, C. Travieso, "Influence of initialization and Stop Criteria on HMM based recognizers" , Electronics letters of IEE, *Vol. 36*, June 2000, 1165-1166.

[12] R. Bellman and S. Dreyfus, "Applied Dynamic Programming". Princeton, NJ: Princeton University Press, 1962.

[13] H. Silverman and D. Morgan, "The application of dynamic programming to connected speech recognition" IEEE ASSP Magazine, vol. 7, no. 3, pp. 6-25, 1990.

[14] Y. Bengio, "Artificial Neural Networks And Their Application To Sequence Recognition" PhD Thesis, McGill University, Montreal, cannada, 1991

[15] Bourlard H. and Wellekens C. J., "Speech Pattern Discrimination and Multilayer Perceptrons," *Computer Speech and Language*, vol. 3, pp. 1-19, 1989.

[16] P.Haffer M. Franzini A.waibel "Integrating Time Alignment And Neural Networks For High Performance Continuous Speech Recognition" Proc of the ICASSP' 91, pp.105-108, Torento, 1991.

[17] Morgan N. and Bourlard H., "Continuous Speech Recognition Using Multilayer Perceptrons with Hidden Markov Models," *in Proceedings of IEEE ICASSP*, vol. 2, pp. 26-30, Albuquerque, 1990.

[18] J.S. Bridle "Training Stochastic Model Recognition Algorithms As Networks Can Lead To Maximum Mutual Information Estimation Of Parameters" Advances in Nips (ed. D.s. Toniesky), Morgan Kaufmann Publ., pp.211-217,1990.

[19] G. Zavaliagkos, Y. Zhao, R. Schwartz and J Makhoul, "A Hybrid Segmental Neural Net/Hidden Markov Model System for Continuous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 151-160, 1994.

[20] X. Driancourt, L. Bottou, and P. Gallinari, "Learning Vector Quantization Multilayer Perceptron and Dynamic Programming: Comparison and Cooperation," *in Proceedings of the International Joint Conference on Neural Networks*, *IJCNN*, vol. 2, pp. 815-819, 1991.

[21] J. Tebelskis, A. Waibel, B.Petek, O.Schmidbauer, "Continuous Speech Recognition Using Linkeed Predictive Network" Advances in Neural Information Processing Systems 3, Eds Lippman, Moody and Touretsky, Publ. Morgan Kaufman, pp.199-205,1991.

[22] M .Franzini, K.F. Lee , and A. Waibel, "Connectionist Viterbi Training: A New Hybrid Method for Continuous Speech Recognition," *in Proceedings of ICASSP*, Albuquerque, NM, pp. 425-428, 1990.

[23] L. T. Niles , H. F. Silverman, "Combining Hidden Markov Models and Neural Networks classifiers," *in Proceedings ICASSP*, pp. 417- 420, Albuquerque, NM, 1990.

[24] E. Levin, "Word Recognition Using Hidden Control Neural Architecture," in Proceedings ICASSP, Albuquerque, NM, pp. 433-436, 1990.

[25] R. Djemili, M. Bedda, H. Bourouba "Recognition Of Spoken Arabic Digits Using Neural Predictive Hidden Markov Models" International Arab Journal on Information Technology, IAJIT, Vol.1, N°2, pp. 226-233, July 2004.

[26] M. N. Do and M. Wagner, "Speaker recognition with small training requirements using a combination of VQ and DHMM" , Proc. of Speaker Recognition an