Optimizing Some HMM Model Parameters in an Isolated Speech Recognition System


1 Equipe Traitement du Signal et Electronique Medicale, Laboratoire d’Electronique et de la Technologie de l’Information ‘LETI’, ENIS, Universite de Sfax – Tunisie

E-mail: mondher_frikha05@yahoo.fr, bm_sajia@yahoo.fr

Abstract

Hidden Markov models (HMMs) are stochastic models. They have been applied with great success in the field of speech recognition during the last three decades. It has been shown that the performance of a recognizer based on HMM modeling may be affected by a bad choice of the type of acoustic feature parameters in the acoustic front end module. For these reasons, we proposed in this paper a speech recognition system based on word-level HMMs built on the platform HTK (Hidden Markov model Toolkit Ver. 3.2) and we investigated its performance using an acoustic front end module based on Mel Frequency Cepstral Coefficients (MFCC). For better recognition rates, we tried through our experiments to modify the number of state in each HMM. Some system’s recognition rates are evaluated with different kind of MFCC derived coefficients. Results showed that a best recognition rate of 99.77% is obtained with, MFCC appended with the 0th order cepstral parameter and the first and second order regression coefficients, 1 Gaussian mixture and 6 states.

Key words: MFCC, HMM, HTK, TIMIT database

1. Introduction

Statistical techniques based on Hidden Markov Model (HMM) with Gaussian emission densities have dominated signal processing and pattern recognition literature for the past three decades. The power of an HMM representation lies on its ability to model the temporal evolution of speech as well as the acoustic variability of speech observations. The widespread use of HMM modeling of speech is due to the availability of efficient algorithms for parameter estimation procedures that involve maximizing the likelihood (ML) given the model [1, 2, 3, 6]. The convergence of those algorithms in both training and testing phases is guaranteed. For these reasons, statistical techniques remain the sector of interest of many speech researchers who try to find more potential real-world applications related to that challenging domain. Also, it has been shown that the success of most existing systems for speech recognition is related to an explicit choice of the signal processing technique which is expected to increase its performance [4, 5].

In this paper, we proposed a speech recognition system based on word-level HMMs built on the platform HTK (Hidden Markov model Toolkit Ver. 3.2). Left to right topology with continuous observation density distribution is used. The control front end is a Mel filter bank based cepstral transformation (MFCC) which has been shown to outperform other conventional signal processing methods used in speech recognition [1]. The choice of MFCC analysis is also motivated by properties of the human auditory system.

In this work, and for optimization purposes, we have investigated the effect of changing the number of states in each HMM on the performance of the isolated word speech recognition system taking in the acoustic front end a combination of (MFCC) along with log energy, 0th order cepstral coefficient and the first and second derivative parameters.

This paper is organized as follows. Section 2 describes basic theory of the signal modeling procedures based on MFCC parameters and HMM modeling. Experimental details and result discussions are presented in section 3. Finally, conclusion and future work are drawn in section 4.

2. Basic theory

The block diagram of a typical speech recognition system is shown in Figure 1 below.
The isolated word speech recognizer is composed of two main modules: the acoustic front end module based on speech processing techniques and the classifier based on HMM modeling.

2.1 Signal Modeling

In most of today's speech recognition systems use parametric representation of speech rather than the waveform itself. When a speech waveform is put into a recognizer, it is first processed by a front end module which extracts from a raw signal a sequence of observations or features \( O = o_1, \ldots, o_N \) of overlapping frames. Each frame contains an acoustic feature vector \( o \) consisting of cepstral coefficients or filter-bank parameters, etc...

The aim of such parameterization is (i) to preserve the pertinent characteristics of speech signal, (ii) to lower the information rate as much as possible for further processing and (iii) to eliminate as much as possible external effects (Communication channels, speakers differences etc...).

The most popular features used for speech recognition are Mel Frequency Cepstral Coefficients (MFCC). These coefficients are obtained by taking the inverse Fourier transform of the log spectrum after it's warped according to the Mel scale.

The MFCC coefficients can be generated by first filtering the segment through a mel triangular filter bank centered at Mel frequencies given by the equation:

\[
\text{Mel} \left( f \right) = 2595 \log_{10} \left( 1 + \frac{f}{700} \right)
\]  

(1)

The MFC coefficients can be then computed using a decorrelation transform (usually Discrete Cosine Transform DCT) of the log amplitude of the filter bank outputs \( m_j \) [5, 8]:

\[
C_{ij} = \sum_j m_j \cos \left( \frac{\pi j}{M} \right) \quad 1 \leq j \leq q
\]  

(2)

Where, 

- \( M \) : is the number of filters 
- \( q \) : is the number of MFCCs 
- \( m_j \) are obtained by applying triangular filters on the FFT spectrum on a Mel scale.

To better represent temporal variations in the speech signal, higher order time derivatives (delta for first derivatives and delta-delta or acceleration for second derivatives) of signal measurements are appended to the set of static parameters.

2.2 HMM modeling

Hidden Markov models (HMM) are generative models based on stochastic finite state network. That means that they are composed of states that generate the feature vectors over time. Those states are connected by transitions. Each transition carries two sets of probabilities:

- A transition probability which provides the probability of going from one state \( i \) to another state \( j \) within the model denoted by \( a_{ij} \).
- The output probability density function which defines the conditional probability of observing a speech feature when a particular transition takes place.

Most speech recognition systems use continuous observations HMM with diagonal covariance to model the temporal sequence of feature vectors. The HMM model the sequence of feature vectors and statically model the temporal variability in speech, as a piecewise stationary process. That means that an utterance \( O = o_1, \ldots, o_N \) is modeled as a succession of stationary states \( Q = q_1, \ldots, q_k \) (\( k < N \)) with instantaneous transition between these states. Each state in the HMM is supposed to model a part of the sequence (e.g.; word, subword, etc...). Therefore, each state acts as a probability density function (pdf) model governed by its parameters. The most pdf's typically used in stochastic speech modeling are the Gaussian Mixture (GM) density. The speech parameter vector \( o \) is generated from the output probability function \( b_j(o) \) which is computed using the following formula:

\[
b_j(o) = \sum_{i=1}^{M} c_{ij} N(o; \mu_j, \Sigma_j)
\]  

(3)

Where \( M \) is the number of mixture components, \( c_{ij} \) is the mixture weight for the \( i^{th} \) mixture in state \( j \) and 
\( N(o; \mu_j, \Sigma_j) \) is the output of the multivariate Gaussian with mean vector \( \mu_j \) and covariance matrix \( \Sigma_j \).

Each individual Gaussian component is given by [8]:

\[
N(o; \mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp \left( -\frac{1}{2} (o - \mu_j)^T \Sigma^{-1} (o - \mu_j) \right)
\]  

(4)

Where \( p \) is the dimensionality of \( o \), 
\( |\Sigma| \) is the determinant of \( \Sigma \).

In HMM, there are three problems of interest [3, 6]:

- The evaluation problem: Given a model and a sequence of observations, what is the probability that the model generated the observations? Solution: forward-backward algorithm.
- The decoding problem: Given a model and a sequence of observations what is the most likely state sequence in the model that produced the observation? Solution: Viterbi algorithm.
- The learning problem: Given a model and a sequence of observations, what should the model's parameters be so that it has the maximum probability of generation the observations? Solution: Baum-Welch algorithm (or the forward-backward algorithm).
3. Experiments and results

3.1 System Description

The recognizer based on word–level HMM is built using the Cambridge university toolkit HTK [8]. HTK is a collection of programs written in C language that allows building and testing HMM based recognizers in an efficient and flexible manner. It consists of a set of library modules and a set of tools (programs) which provide sophisticated facilities for speech analysis, HMM training, testing and results analysis.

The speech material that was used in this experiment is taken from the TIMIT database [9]. The recognizer should recognize 11 different words taken from the sentence 'sa1' of the TIMIT database "She had your dark suit in greasy wash water all year".

The training set is composed of 180 utterances (1980 words) taken from the TIMIT database which was recorded at 16 kHz in clean condition. The test set composed of 40 utterances (440 words) taken from the TIMIT core test database. Left to right topology with continuous observation density distribution with diagonal covariance matrices is used. Each of the training and test utterance is spoken by different speakers.

The input speech is first preemphasized and windowed at a frame of 25ms with a frame rate of 10ms. The MFCC cepstral vectors are then computed from 26 channels FFT based Mel warped log spectra filter bank. The HMM word prototype models are then configured with a pre-defined number of states, 1 Gaussian mixture and 12 MFCC feature vectors. After that, each HMM word models are initialized using the 'flat-start' procedure which uses the Baum-Welch algorithm to find the most likely state sequence that corresponds to each training sample. The isolated unit models were then refined using the Baum-Welch (forward-backward) algorithm for a fixed number of iterations. The HMM trained models are then used in the recognizer. The Viterbi decoder is used for recognition. For better recognition rates, we tried through our experiments to study the effect of the number of states inside any HMM when using different kind of MFCC feature coefficients taking into account the energy and the dynamic parameters (first and second order deltas).

3.2 Experimental Results

All our experimental results are gathered in Table 1. From these results, it can be seen that the system's highest accuracy of recognition 99.77% is obtained for an isolated word recognition system using HMM classifier with 6 states and an acoustic front end based on MFCC appended with C0, delta and acceleration coefficients. The same accuracy is reached with an acoustic front end based on MFCC appended with the energy (E), delta and acceleration coefficients but 8 states in each HMM. It can be noticed also that the recognizer performance is enhanced by adding the dynamic features to the static MFCC coefficients. In fact, the inter-frame modeling capacity is added when the dynamic information is included. This avoids a number of constraining assumptions in statistical speech recognition systems and particularly the fact that dealing with uncorrelated acoustic vectors (observations).

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>70.91</td>
<td>81.59</td>
<td>94.55</td>
<td>95.46</td>
<td>95.45</td>
<td>98.18</td>
<td>98.16</td>
</tr>
<tr>
<td>MFCC_0</td>
<td>76.82</td>
<td>89.08</td>
<td>96.72</td>
<td>90.31</td>
<td>94.32</td>
<td>98.86</td>
<td>99.08</td>
</tr>
<tr>
<td>MFCC_E</td>
<td>76.82</td>
<td>89.08</td>
<td>96.72</td>
<td>90.31</td>
<td>94.32</td>
<td>98.86</td>
<td>99.08</td>
</tr>
<tr>
<td>MFCC_D</td>
<td>74.09</td>
<td>93.28</td>
<td>98.39</td>
<td>99.59</td>
<td>99.55</td>
<td>99.55</td>
<td>99.55</td>
</tr>
<tr>
<td>MFCC_D_A</td>
<td>77.21</td>
<td>95.36</td>
<td>98.38</td>
<td>99.32</td>
<td>99.32</td>
<td>99.55</td>
<td></td>
</tr>
<tr>
<td>MFCC_D_A</td>
<td>64.32</td>
<td>90.10</td>
<td>98.41</td>
<td>99.54</td>
<td>99.32</td>
<td>98.77</td>
<td>98.77</td>
</tr>
<tr>
<td>MFCC_D_A</td>
<td>86.82</td>
<td>94.77</td>
<td>99.99</td>
<td>99.77</td>
<td>99.77</td>
<td>98.77</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: The effect of the number of states for different kind of MFCC feature coefficients on the speech recognition rate

Figure 2 shows the recognition accuracy of the isolated word recognition system based on HMM with 6 states for different kind of MFCC feature coefficients.

![Recognition accuracy of the isolated word recognition system based on 6 states HMM for different kind of MFCC feature coefficients.](image)

4. Conclusion and Future Work

An isolated word recognition system based on HMM models was built using HTK toolkit. The parameters of 11 HMMs were estimated using Baum Welch procedure. The recognizer was tested using the TIMIT database for the training and testing sets by modifying the number of states and taking only one type of speech parameterization based on the MFCC analysis. A high recognition accuracy of
99.7% is obtained with a number of states in each HMM of 6 and MFCC coefficients appended by cepstral coefficient of order 0 as well as deltas and acceleration coefficients. That means that the combination of dynamic and static features had proved the discriminability for speech pattern comparison and consequently improved the accuracy of the speech recognition process.

There are a number of experiments that we plan to do to extend this work in order to more optimize our recognizer:

- Augment the number of Gaussian mixture components
- Try other kind of speech analysis such as LPC and PLP
- Augment the size of the training and testing sets
- Check the performance of the recognizer in adverse conditions.

5. References