International Journal of the Physical Sciences Vol. 6(22), pp. 5064-5074, 2 October, 2011
Available online at http://www.academicjournals.org/IJPS
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

Hidden Markov models (HMMs) isolated word recognizer with the optimization of acoustical analysis and modeling techniques

Mondher Frikha[1]*, Ahmed Ben Hamida[1] and Mongi Lahiani[2]

[1]Advanced Technologies for Medical and Signals (ATMS) Research Unit, National School of Engineering of Sfax, Route, B. P. W, Sfax, Tunisia.
[2]Laboratory of Electronics and Information Technologies (LETI), National School of Engineering of Sfax, B.P.W, Sfax, Tunisia.

Accepted 21 December, 2010

Most state of the art automatic speech recognition (ASR) systems are typically based on continuous Hidden Markov Models (HMMs) as acoustic modeling technique. It has been shown that the performance of HMM speech recognizers may be affected by a bad choice of the type of acoustic feature parameters in the acoustic front end module. For these reasons, we propose in this paper a dedicated isolated word recognition system based on HMMs which was carefully optimized specifically at the acoustic analysis and HMM acoustical modeling levels. Such conception was tested and valued on Hidden Markov model toolkit platform (HTK). Systems performances were evaluated using the TIMIT database. One comparative study was carried out using two types of speech analysis: The cepstral method referred to as Mel frequency cepstral coefficients (MFCC) and the perceptual linear predictive (PLP) coding are used for different tests so as to evaluate and reinforce our conception. The frame shift duration effect of the acoustic analysis as well as the addition of the dynamic coefficients of the acoustic parameters (MFCC and PLP) were carefully tested in order to look for high accuracy for our optimized isolated word recognition (IWR) system. Finally, various experiments related to the HMM topology have been carried out in order to get better recognition accuracies. In fact, the effect of some modeling parameters of HMM on the recognition accuracy of the IWR system such as the number of states as well as the number of Gaussian mixtures were analyzed in order to get the optimal HMM topology.

Key words: Isolated word recognition, perceptual linear predictive (PLP) coding, Mel frequency cepstral coefficients (MFCC) PLP, HMM, Hidden Markov model toolkit platform (HTK),

INTRODUCTION

Speech is the most natural form of communication for humans. Automatic Speech Recognition (ASR) systems generally assume that the speech signal is a realisation of some message encoded as a sequence of one or more symbols (Calliope, 1989; Boite et al., 2000). We learn to speak and recognise the speech of others at an early age and without instruction. Modern speech recognition systems typically classify speech into sub-units (word, phoneme, etc.). The art and science of speech recognition have been advanced to the state where it is now possible to communicate reliably with a computer by speaking to it in a disciplined manner using a vocabulary of moderate size. The scope and quality of automatic speech recognition (ASR) systems has increased considerably, moving from isolated word recognition with small vocabularies (Ben Messaoud et al., 2005; Frikha, 2008) to large vocabulary continuous speech recognition systems. During the past few years, several studies investigated the potential of directly measured speech production parameters to improve the accuracy of automatic speech recognition systems

*Corresponding author. E-mail: mondher_frikha05@yahoo.fr.

Abbreviations: ASR, automatic speech recognition; **HMM,** hidden Markov models; MFCC, mel frequency cepstral coefficient; **PLP,** perceptual linear predictive; IWR, Isolated word recognition; **FIR,** finite impulse response; **P(W/O),** posteriori probability.

(Jankowski et al., 1995; Hermansky, 1990; Makhoul, 1975; Davis and Mermelstein, 1980; Furui, 1986).

The goal of ASR systems is to transcribe human speech into text, which can be further processed by machines or displayed for humans for reading in various applications. Up to now, the most efficient approach in speech recognition is the Hidden Markov Model (HMM) (Rabiner, 1989; Rabiner and Juang, 1986; Magdi and Gader, 2000; Ephraim and Merhav, 2002, Frikha et al., 2005). Various laboratories use this technique as a potential tool for the conception of recognizing systems and commercial ASR systems are actually based on this theory (Young et al., 2002). In particular, HMM based on continuous probability distributions has better recognition rates than discrete HMM using vector quantization preprocessing.

In this work, we were interested in the study of the HMM system topology regarding mainly the optimization of such recognizing system. We tried to modify the number of states of the HMM structure, the number of Gaussian components at each state and the frame shift duration analysis of one IWR system. Several experiments were performed using two kinds of acoustic features: the MFCC and PLP static coefficients eventually appended with their first and second differential derivatives. The ASR system was finally tested with HTK platform (Young et al., 2002). All words were extracted from TIMIT database (DARPA, 1990).

This paper is organized as follows. Subsequently, the state of the art of speech recognition systems is exposed and the feature extraction and acoustical modeling modules which are the basic components of those systems are described, after which the optimization strategies followed at the acoustical analysis and acoustical modeling levels in the IWR task are presented. This is followed by a description of the experimental details, after which the results were supplied when the optimization techniques described here are applied. Finally, the summary and conclusions of the study are given.

STATE OF THE ART IN SPEECH RECOGNITION

The task of a speech recognition system is to produce an estimate of the word sequence associated with a given speech waveform (Calliope, 1989; Picone, 1990; Rabiner, 1989). A diagram of speech recognition process is shown in Figure 1.

Produced speech is first converted into a specific form suitable for recognition processing, a process known as feature extraction.

Feature extraction

The purpose of feature extraction in speech recognition is to transform speech signals into a set of vectors relevant

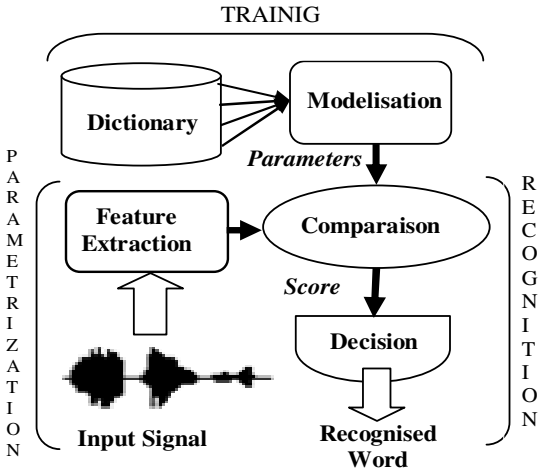

Figure 1. Diagram speech recognition process.

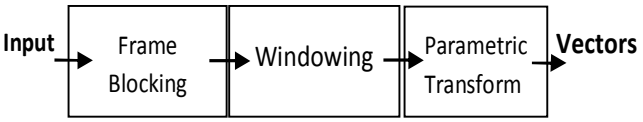

Figure 2. Features' extraction principle.

for speech recognition while discarding unreliable parts. Almost all speech recognition systems use a parametric representation of speech rather than the waveform itself as the basis for pattern recognition. The parameters usually carry the information about the short-time spectrum of the signal. Firstly, a front-end para-meterisation is needed which can extract from the speech waveform all of the necessary acoustic information in a compact form, compatible with the mathematical acoustical modeling tool. Hence, the feature extraction module would be one front end module which extracts from the input signal a sequence of observations or features referred to as $O=o_1,\ldots,o_m$ of an overlapping frames. Each frame contains an acoustic feature vector O. Figure 2 shows the general process for feature vectors extraction.

The majority of feature extraction techniques commonly used today is based on simplified vocal tract models. We used in our experiments the two dominant approaches of feature extraction, the Mel Frequency Cepstral Coefficient Analysis (MFCC) (Davis and Mermelstein, 1980), and then perceptual linear predictive (PLP) Coding (Hermansky, 1990). The common approach for those representations is to modify the front end module to mimic the frequency map of the ear. The next subsection indicates the main steps followed in the extraction's process.

Acoustical modelling

After speech has been encoded with a small number of

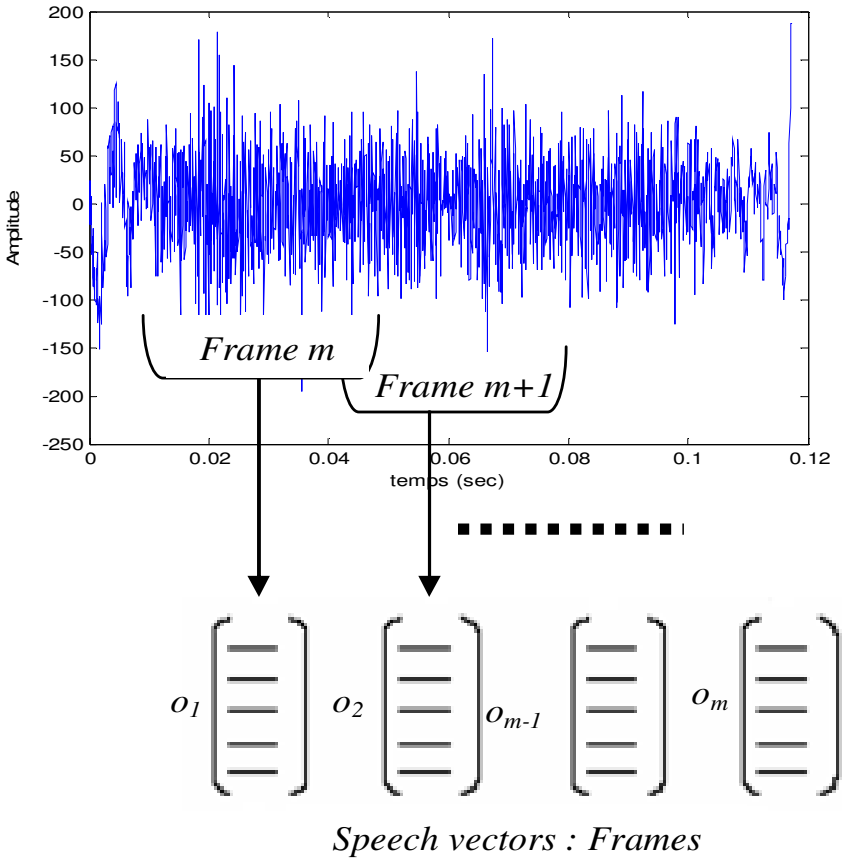

Figure 3. Feature extraction process.

parameters called features, a speech model has to be established in order to model characteristics and variations of speech and to further reduce the number of parameters. During the past few years, several acoustical modeling methods were used in speech recognition process. The most popular ones are the Hidden Markov Models referred to as 'HMM' and the artificial neural networks known as 'ANN'. Experiments conducted in this work were obtained using only HMM based recognizers. HMM modeling tool supposed that speech is a piecewise stationary process, that is, a unit is modelled as a succession of discrete stationary states, with instantaneous transitions between these states (Picone, 1990; Rabiner, 1989; Rabiner and Juang, 1986). The widespread use of HMMs is due to the existence of efficient and powerful training and recognition algorithms (Jelinek, 1976; Baum, 1972). HMM may be due to the existence of an efficient training algorithm, the Baum-Welch or forward-backward algorithm (Dempster et al., 1977). Given the structure of HMM and training data, the algorithm finds the parameter values of the HMM according to the maximum likelihood (ML) criterion. The convergence of this iterative procedure to a local maximum of the objective function is guaranteed by an inequality discovered by Baum (1972).

OPTIMIZATION IN ISOLATED WORD RECOGNITION

Acoustical analysis

Almost all speech recognition systems use a parametric representation of speech rather than the waveform itself as the basis for pattern recognition. The parameters usually carry the information about the short-time spectrum of the signal. Firstly, a front-end parameterization is needed which can extract from the speech waveform all of the necessary acoustic information in a compact form compatible with the HMM based acoustic models. 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. Figure 3 shows the general process for feature vectors extraction.

Mel frequency cepstrum coefficients

MFCC analysis is considered as the standard method for feature extraction in speech recognition tasks. First, speech waveform is pre-emphasized with one coefficient finite impulse response (FIR) digital filter (Picone, 1993):

$$H_{pre}(z)=1+a_{pre}z^{-1} \quad (1)$$

a_{pre} is the pre-emphasis coefficient which value is close to -1.

The resulting signal is then windowed with a specified window function. A commonly used window is the Hamming. It is calculated as (Boite et al., 2000):

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

where N is the length of the Hamming window.

For each frame, the fast Fourier transform (FFT) is performed to estimate its power spectrum which is then fitted to a bank of Mel-filters modeling the hair spacing along the basilar membrane of the ear.

Mel-scale cepstral coefficients are extracted in two stages. First, narrow-band filter energies are determined using a Mel-scale filter bank centred at Mel frequencies given by the equation (Boite et al., 2000):

$$Mel(f) = 2595x\log_{10}\left(1+\frac{f}{700}\right) \quad (3)$$

Next, these filter energies are coded using a Fourier transform. This is the cepstral analysis stage of processing (Kammoun et al., 2006). A specified number of overlapping triangular filters with center frequencies equally spaced in the corresponding mel-frequency scale are placed in a limited predefined frequency range (Davis and Mermelstein, 1980). The Mel frequency scale is based on results from psychophysical studies on humans. Each interval on the Mel-scale corresponds to the perceived relative pitch of a reference tone. The Mel-scale filter bank can be considered to be one of a set of possible spectral estimation techniques.

The MFC coefficients can be then computed using a decorrelation transform usually discrete cosine transform (DCT) of the log amplitude of the filter bank amplitudes $\{m_j\}$ (Kammoun et al., 2006):

$$C_i=\sqrt{\frac{2}{N}}\sum_{j=1}^{N} m_j\cos\left(\frac{\pi i}{N}(j\text{-}0.5)\right); \text{ For } 1 \le i \le q \quad (4)$$

where, N is the number of filter banks and q is the number of MFCCs

Perceptual linear predictive

The PLP analysis method is more adapted to human hearing, in comparison to the classic linear prediction coding (LPC) (Makhoul, 1975). The procedure for computing the PLP coefficients is largely the same as the procedure for determining the LP cepstral coefficients (Hermansky, 1990; Morgan et al., 1995). The primary difference is that, before performing the all-pole modeling, the power spectrum is warped, smoothed and compressed based on concepts from auditory perception. In the PLP analysis, the speech processing is based on some biological analogies and physiological characteristics of human hearing (Hermansky, 1990). The method is inspired of the principle of the combined linear prediction by a representation of the word signal that follows a ladder of the human audition. A conversion from frequency to Bark, which represents a better representation of the human hearing resolution in frequency, the Bark frequency corresponding to an audio frequency is given by:

$$F_{Bark} = 6\ln\left[\frac{f}{600}+\sqrt{\left(\frac{f}{600}\right)^2+1}\right] \quad (5)$$

The resulting spectrum is convoluted with the power spectrum of the critical band-masking curve, which act like a bank of filters centred at Bark frequencies given by the last equation. The last operation prior to the all-pole modeling is the cubic-root amplitude compression (intensity–loudness conversion), which simulates the non-linear relation between the intensity of sound and its perceived loudness. Autoregressive modeling is the final stage of the PLP analysis, which consists of approximating the spectrum by an all-pole model, using the autocorrelation method. An inverse discrete fourier transformation is applied to the spectrum samples, resulting in the dual autocorrelation function.

Spectral dynamics

Dynamic features capture the rate of change (speed) and acceleration of spectral components (delta Δ and double delta ΔΔ parameters). It was shown that those features greatly improve speech recognition rates when they are taken together with traditional features (Jankowski et al., 1995; Furui, 1986). Dynamic features need not be extracted from the data itself, they can rather be determined from a set of static features such as LP cepstral coefficients or MFCCs. Since static features generally give information on the time-localized spectral envelope, changes in static features give information about the spectral dynamics of a signal.

Dynamic features are determined using a simple first difference between static feature vectors or by applying regression analysis. The delta coefficients are computed using the following regression formula (Young et al., 2002):

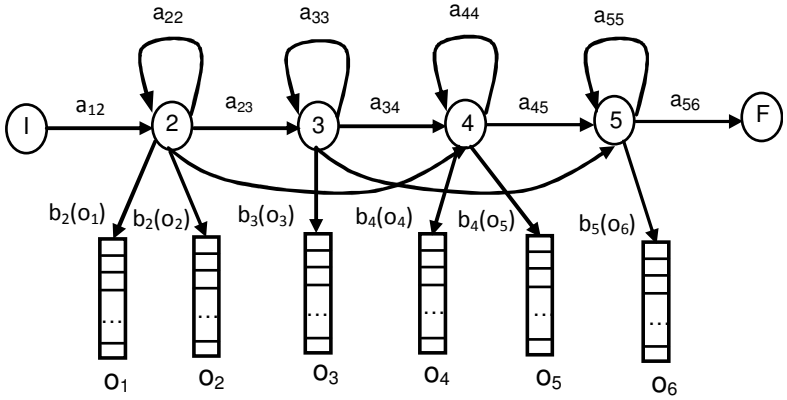

Figure 5. A schematic of a six states, left-to-right HMM. a_{ij} = Transition probability from state i to state j; b_i = Output probability density function for state i,

$$D_t = \frac{\sum_{k=1}^{M} k(c_{t+k} - c_{t-k})}{2\sum_{k=1}^{M} k^2} \quad (6)$$

where D_t is the delta coefficient at frame t, and c_{t-w} and c_{t+w} are static parameters before and next to the current frame c_t, respectively. The dynamic feature calculation represents then a finite impulse response (FIR) filtering of the time trajectories of static feature coefficients.

Acoustical model

Hidden Markov modeling of speech assumes that speech is a piecewise stationary process, that is, a unit is modelled as a succession of discrete stationary states, with instantaneous transitions between these states (Rabiner, 1989; Ephraim and Merhav, 2002). Those states are connected by transitions. Each transition carries two sets of probabilities:

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

The optimal decoder (speech recognizer) which achieves expected minimum word recognition error rate is the following maximum *a* posteriori MAP decoder which is referred to as optimal MAP decision rule:

$$W^* = \arg\max_W P(W/O) \quad (7)$$

where, P(W/O) is known as the *a* posterior probability since it represents the probability of occurrence of a sequence of words W after observing the acoustic signal O. W* is then the recognised word sequence.

Because of the complexity of the speech production mechanisms, there is no simple parametric representation of P(W/O) that involves both acoustic and linguistic information. The basic approach is to first divide the problem into acoustic and linguistic components that can be handled separately. This is achieved using a Bayesian reformulation:

$$W^* = \arg\max_W \frac{P(O/W).P(W)}{P(O)} \quad (8)$$

P(O/W) encodes the statistical distribution of speech acoustics given the linguistic labelling. P(W) is the probability assigned by the language model which encodes the *a* priori linguistic information. P(O/W) is provided by an acoustic model as HMMs. Equation 7 indicates that to find the most likely sequence of words W = $\{w_1, w_2, \ldots\ldots, w_k\}$, the word sequence which maximises the likelihood P(O/W) must be found.

A simple HMM is illustrated in Figure 5. Essentially, an HMM is a stochastic automaton, with a stochastic output process attached to each state. Thus we have two concurrent stochastic processes: a Markov process modeling the temporal structure of speech and a set of state output process, modeling the stationary character of the speech signal. Most speech recognition systems use continuous observations, HMM with diagonal covariance to model the temporal sequence of feature vectors.

An input feature vector o_t at time frame t is associated with state i with a probability which can be calculated from the transition and the output probabilities. The next input feature vector at time frame t+1 may be associated with the same state i again (with self-transition probability a_{ii}) or state j (with transition probability a_{ij}, $i \neq j$). In this way a sequence of input feature vectors is associated with the

Table 1. Isolated recognition system vocabularies for the training and testing corpus.

Number of words	Words to recognise	Train corpus (words)	Test corpus (words)
7	She, had, all, ask, rag, like, wash	3231	616
8	She, had, all, ask, rag, like, wash, carry	3693	704
9	She, had, all, ask, rag, like, wash, carry, that	4155	792
10	She, had, all, ask, rag, like, wash, carry, that, dark	4617	880

states. Several different sequences of states can correspond to a particular input sequence. If a_{ii} is large, the next state will more likely be the same state i, so state i will be associated with more input feature vectors representing a unit segment that is longer in time. If a_{ii} is small, the input feature vectors for that segment tend to be shorter in time. Because a transition is given by a probability, an HMM can represent input feature vectors with different lengths.

Most state-of-the-art HMM systems use Gaussian mixtures to represent the output probabilities. The speech parameter vector o_k is generated from the output probability function $b_j(o_k)$ which is computed using the following formula:

$$b_j(o_k) = \sum_{j=1}^{M} c_{ji} N\left(o_k, \mu_{ji}, \Sigma_{ji}\right) \quad (9)$$

where: $b_j(o_k)=b_j(k)$, is the output probability for state j; c_{ji} is a mixing weight; μ_{ji} is a mean vector, and Σ_{ji} is a covariance matrix for i^{th} Gaussian in state j.
Each individual Gaussian component is given by (Rabiner, 1989):

$$N(o_k, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{\frac{-1}{2}\left(o_k-\mu\right)^t \Sigma^{-1}\left(o_k-\mu\right)\right\} \quad (10)$$

where n is the dimensionality of o_k and $|\Sigma|$ is the determinant of Σ. More details about HMMs can be found in (Rabiner, 1989; Rabiner and Juang, 1993).

We use the compact notation $\lambda=(A, B, \Delta)$ to indicate the complete parameter set of the model; where:

$A=\{a_{ij}\}$, a_{ij} is the state transition probability distribution.
$B=\{b_j(k)\}$, $b_j(k)$ is the observation probability distribution.
$\pi = \{\pi_i\}$, π_i is the initial state distribution.

For a given test pattern (or a sequence of input feature vectors) $O=[o_1,o_2,\ldots,o_T]$, the likelihood $P(O|\lambda_m)$ of the pattern O being generated from a model $\lambda_m\Delta\Delta$ representing the m^{th} word is calculated. If a particular model $\lambda_m\Delta\Delta$ has larger likelihood than all other models, the test pattern is determined to be the m^{th} word. There are three main problems that have to be solved in order to use hidden Markov models for speech recognition:

Probability evaluation

Computation of P(O□Δλ), that is, the probability that the model λ gives rise to the observation sequence O.

State sequence

Finding the most likely state sequence q, This problem tries to look 'inside' the HMM to understand the way the observation sequence comes about. There is no unique solution to this problem since different state sequences can result in the same observation. Hence, it is only possible to find a most likely one.

Model training

Computing the parameters (A,B,λ) for the model λ. This step corresponds to the training of the recognizer because given a labeled observation sequence O, this algorithm finds the 'optimum' model parameters. The solution to this task is a re-estimation technique that finds the optimum solution iteratively.

The task of the training problem to find the parameter set, most likely to yield the observation sequence, is a standard maximum likelihood problem from estimation theory (Dempster et al., 1977). Iterative maximization procedures must be used. One such solution was determined by Baum and colleagues, and is generally referred to as the Baum or Baum-Welch re-estimation algorithm.

EXPERIMENTAL PROCEDURE

Recognition system description

The objective of the isolated word recognition system is to recognize isolated words extracted from TIMIT database (DARPA, 1990). A vocabulary of 7, 8, 9 and 10 words was used as indicated in Table 1. The train and core sets were constructed from the training and testing sets of the TIMIT database. The recognition system has been conceived and tested under the Cambridge University toolkit HTK (Hidden Markov Model Toolkit) platform

Table 2. Word accuracy (in %) for various number of words in vocabulary using different feature sets.

	7 words	8 words	9 words	10 words
MFCC	97.24	97.73	96.21	96.02
PLP	98.21	97.3	95.83	95.11
MFCC_E	98.86	98.15	97.1	97.39
PLP_E	98.7	97.87	97.22	96.93
MFCC_0	98.05	97.59	96.72	96.7
MFCC_D	99.19	99.29	98.86	98.41
PLP_D	99.19	99.29	98.74	98.51
MFCC_D_A	99.51	99.57	98.99	98.52
PLP_D_A	99.35	99.43	98.86	98.41
MFCC_E_D_A	99.51	99.29	98.48	98.3
PLP_E_D_A	99.51	99.29	98.36	98.3
MFCC_0_D_A	99.19	99.15	98.61	98.52

(Young et al., 2002). HTK is composed of a set of tools enabling definition, initialization, re-estimation, and editing of sets of continuous mixture Gaussian HMMs.

HTK provides an analysis tool called HCopy which converts a specific train and test speech wave files into appropriate sequence of feature vectors. Typically, a speech recognition feature vector consists of 12 static coefficients (C_1, C_2,…, C_{12}) to which might be added one of the following component: a log energy (_E), first derivative (_D), log energy and first derivative (_D_E), first derivative and second derivative (_D_A), log energy, first derivative and second derivative (_E_D_A). If the mel cepstral coefficient of order 0 (C_0) has to be considered, then the qualifier (_0) is used.

All ASR systems, using HTK, operate in two phases. First, there is a training phase during which the system learns the reference patterns representing the different speech sounds (phrases, words, syllables, or phones) that constitute the vocabulary of the application. Each reference is learned from spoken examples and stored either in the form of templates obtained by some averaging procedure or models that characterize the statistical properties of the speech pattern. Second, there is a recognition phase during which an unknown speech signal is identified using the stored reference patterns. The training and testing phases are accomplished as follows:

Training phase

A prototype model is first created by specifying the HMM design considerations, that is, the number of states, the transition matrix and the parameters for the observation pdfs. Then a model initialization is performed using the tool HCompV. Finally, all the HMMs are updated simultaneously using the embedded re-estimation tool HERest. This tool has to be executed several times, because it only performs one iteration at a time.

Testing phase

The re-estimated models are tested by confronting them with so far unknown test data using the tool HVite. HVite takes as input, a network describing the allowable words, a dictionary defining how each word is pronounced and a set of HMMs. It operates by converting word network to unit network and then attaching the appropriate HMM definition to each unit instance. The best matching word is found according to the Viterbi-based speech recognition algorithm (Rabiner, 1989) and the output is stored in a master label file. Finally, the HR results were compared to the correct transcription with the output of HVite and used to computes error statistics.

Experiment

The left to right HMM word prototype models was used. The observation probability distribution is a Gaussian mixture density with diagonal covariance matrix. 12 static coefficients vectors were computed using 25 ms Hamming window, shifted with 10 ms steps and a pre-emphasis factor of 0.97. Two kinds of acoustic features were adopted for all our experiments (MFCC and PLP).

The HMM word models were 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. They were later refined using the Baum-Welch (forward-backward) algorithm (Rabiner and Juang 1993; Magdi and Gader, 2000), in order to find the optimal parameters for the recognition phase. The Viterbi decoder was used for recognition. To achieve a very good recognition performance, we tried throughout our experiments to modify the number of states in each HMM, the number of Gaussian Mixture components and the frame shift duration of each system.

RESULTS AND DISCUSSION

Comparison between different kinds of acoustic analysis

For our first experiment, we tried to study the effect of the energy and dynamic features when appended to the static MFCC and PLP vector components. For that, we gradually augment the vocabulary size from 7 to 10 words. The probability distribution functions (pdf) associated with each emitting state is one Gaussian mixture. The period of frame analysis is maintained equal to 10 ms. All our experimental results are gathered in Table 3. Figures 4 and 5 shows the recognition accuracy (RA) of the recognition systems for the MFCC and PLP feature kinds, respectively. When we compare the obtained experimental results of the two kinds of the

Table 3. Recognition accuracy (in %) as a function of a number of words in vocabulary for different HMM's number of states.

Number of words	4	5	6	7	8
7	98.7	99.51	99.68	99.68	99.68
8	98.58	99.57	99.72	99.86	99.72
9	96.59	98.99	99.24	99.37	99.37
10	96.25	98.52	98.75	98.98	99.2

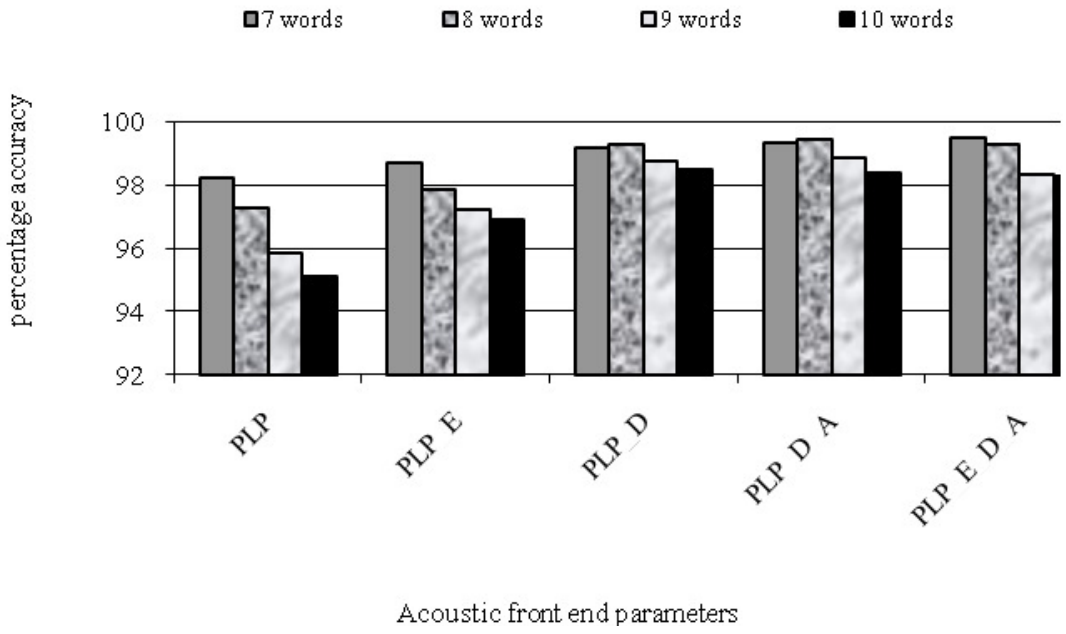

Figure 4. Performance (word accuracy in %) for PLP features appended by dynamic features for various vocabulary sizes.

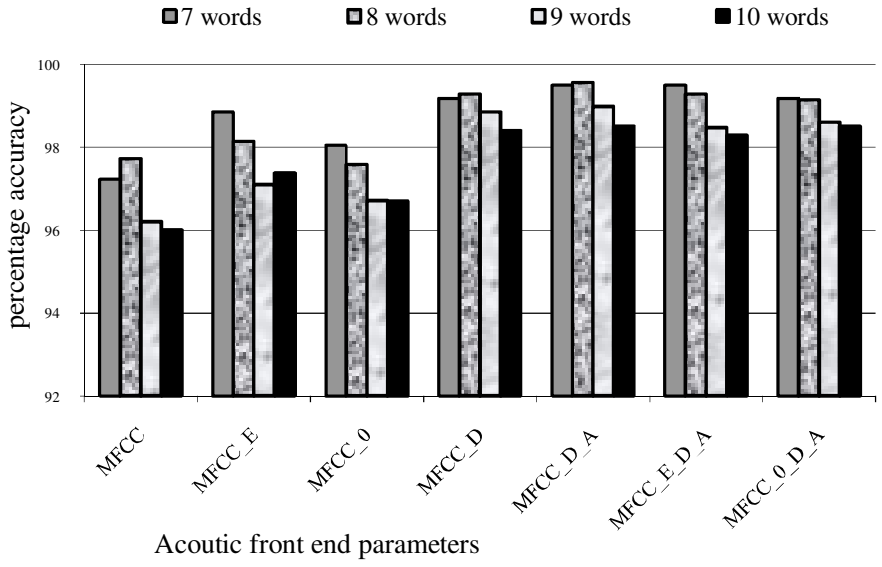

Figure 5. Performance (word accuracy in %) for MFCC features appended by dynamic features for various vocabulary sizes.

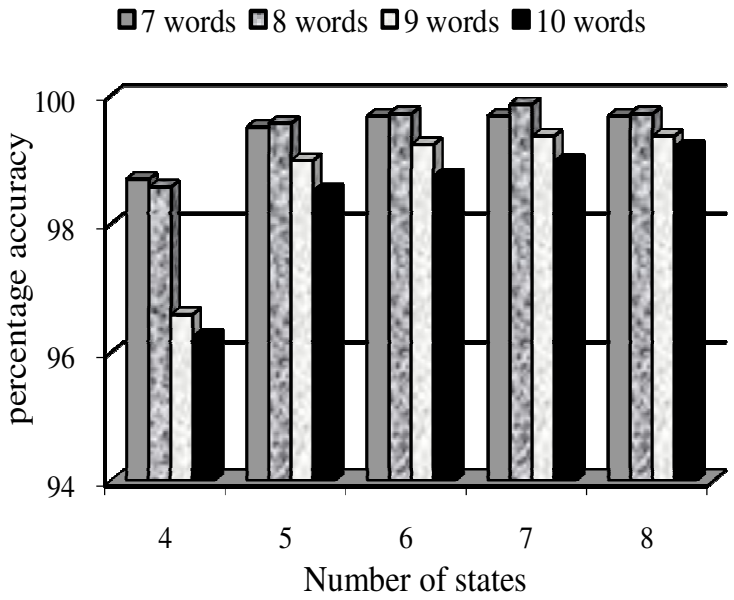

Figure 6. Performance (word accuracy in %) for MFCC_D_A features for different HMM's number of state.

Table 4. Recognition accuracy (in %) as a function of a number of words in vocabulary for different Gaussian mixtures.

Number of words	1	2	4	8	16
7	99.51	99.68	99.80	99.84	99.84
8	99.57	99.74	99.78	99.86	99.86
9	98.99	99.37	99.62	99.87	99.86
10	98.52	99.20	99.66	99.77	99.66

tested analysis, we remark that the dynamic coefficients which model the sequential properties of the speech signal and increases the performance of the recognition system. In fact, when apprehending the first and second derivatives features to the static components, a significant amount of temporal information of feature vectors is known. This yields to an improvement in recognition accuracy. Best results are hence obtained with the parameterization based on MFCC appended with dynamic coefficients (MFCC_D_A). Results given by PLP analysis are close to those obtained by the MFCC analysis since the procedure for extracting the corresponding feature vectors is motivated by the workings of the human auditory system. For the PLP parameterization, good results are obtained by PLP appended with the first order regression coefficient (PLP_D).

We will consider only (MFCC_D_A) for the remaining experiments in which we plan to explore the effects of the number of states of HMM as well as the number of Gaussian mixture and period of frame on the performance of the isolated word recognition system.

Effect of the number of states

For the second experiment, All HMMs are one mixture Gaussians. The frame shift is fixed at 10 ms. We varied the number of states in HMMs from 4 to 8. Table 3 summarizes all the obtained experimental results.

Figure 6 shows the word accuracy rate of the recognition system as a function of the number of states, HMM. When analysing the obtained experimental results, we noticed that, for better recognition rate, the number of states increases along with the number of the recognised words. In fact, for 8 and 9 words, the best accuracies are obtained for HMM models with seven states.

Effect of the number of Gaussian mixture

For the third experiment, we fixed the number of HMM states to five. The frame period is maintained equal to 10 ms. We varied the number of Gaussian Mixture (GM) from 2, 4, 8 and 16. Table 4 summarizes all the experimental results.

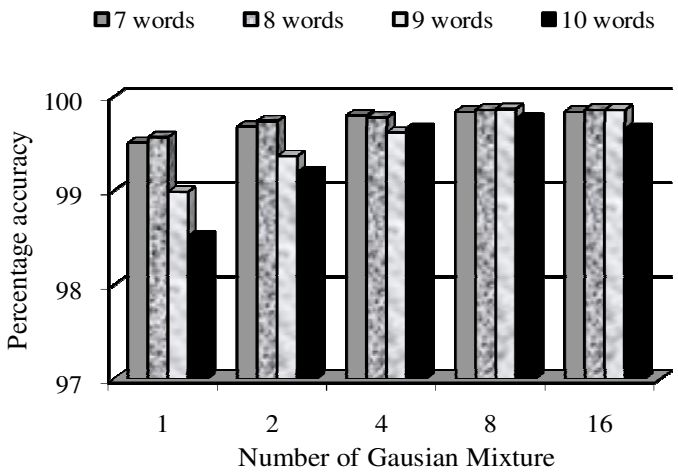

Figure 7. Performance (word accuracy in %) for MFCC_D_A features for different Gaussian mixtures.

Table 5. Recognition accuracy (in %) as a function of a number of words in vocabulary for different frame's shift duration.

Number of words	5 (ms)	10 (ms)	15 (ms)	20 (ms)	25 (ms)
7	98.86	99.51	99.51	99.19	99.19
8	99.15	99.57	99.57	99.29	99.20
9	97.47	98.99	99.12	98.86	99.12
10	96.59	98.52	98.86	98.86	98.75

Figure 7 shows the word recognition accuracy as a function of the number of Gaussian mixtures. From the depicted experimental results, we observed that the increase of the number of GMs produces better recognition accuracy for HMM models with 8 GMs.

Effect of the frame shift duration

For the forth experiment, the HMM emission probability distributions associated to each state are modelled with one Gaussian mixture. The number of states for HMMs is fixed equal to 5. We varied the frame shift of the Hamming analysis window from 5 to 25 ms with 5 ms step. The Table 5 summarizes all our experimental results.

Figure 8 shows the word recognition accuracy as a function of the frame shift. From the obtained results, the best word recognition accuracy is reached with frame shift of 10 ms.

Conclusion

We proposed in this research, several optimization strategies for HMM classifiers using Baum-Welch These optimization strategies were validated with several experiments using different vocabulary sizes extracted from TIMIT database. For our first experiment, we used a single Gaussian mixture distribution for each of the three emitting states of the left to right HMM. Here we chose to fix the frame shift duration to 10 ms. We confirmed similarity in performance of MFCC and PLP features which were both based on psychophysical studies on human auditory perception. The dynamic coefficients delta and acceleration enhanced the recognition accuracy. In fact, an average relative improvement in performance of 1.7% over the MFCC baseline system is obtained with MFCC appended by first and second derivatives (MFCC_D_A). In the second experiment, we modified the number of states in the left to right HMM topology with only one Gaussian mixture associated to each state. For the selected MFCC_D_A feature, the best recognition performance is achieved with seven HMM states. In the third experiment, we used five states left to right models and we modified the number of mixtures in each state. For one 10 ms fixed period of frame, the best word accuracy was reached with eight HMM states. Finally, for the forth experiment, we explored the effect of the frame shift duration analysis for the MFCC_D_A features when left to right five HMM states classifier with one Gaussian mixture in each state used. In such

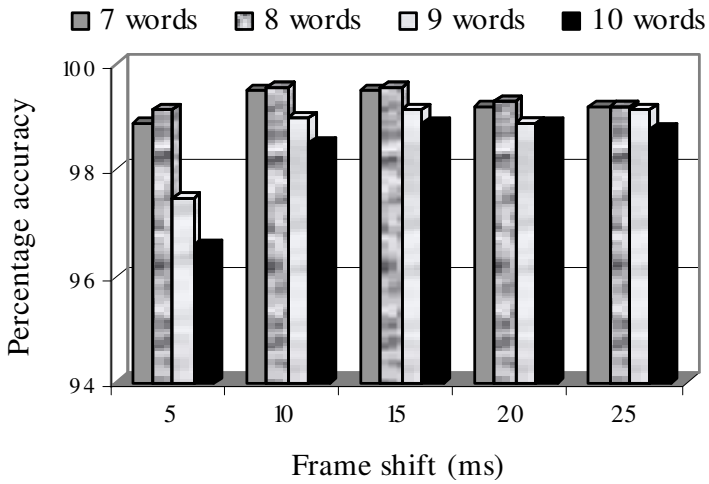

Figure 8. Performance (word accuracy in %) for MFCC_D_A features for different frame's shift duration.

conditions, we showed that 10 ms frame shift duration was the optimal choice as far as the word accuracy criterion was concerned.

REFERENCES

Baum L (1972). ''An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes''. Inequalities 3:1-8

Ben Messaoud S, Frikha M, Lahyani M, Ben Hamida A (2005). ''A Study of the Effects of Acoustic Front Ends and Gaussian Mixtures on an Isolated Speech Recognition System based on HMM'', *IEEE conference ACIDCA05*, Tozeur, Tunisia.

Boite R, Bourlard H, Dutoit T, Hang J, Leich H (2000), ''Traitement de la parole'', *Presses* Polytechniques et universitaires Romandes, ISBN 2-88074-388-5. Lausanne, Suisse.

Calliope (1989)., ''La parole et son traitement automatique'', Collection technique et scientifique des télécommunications. Masson, CENT, ENST.

Davis SB, Mermelstein P (1980), ''Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences''. IEEE Trans. Acoust. Speech Signal Process., 28(4): 357– 366

Dempster AP, Laird NM, Rubin DB (1977). ''Maximum Likelihood from Incomplete Data via the EM Algorithm''. J. Roy. Stat. Soc., 39: 1–38.

Ephraim YE, Merhav N (2002). "Hidden Markov processes," IEEE Trans. Inf. Theory, 48(6): 1518–1569.

Frikha M (2008). ''Approche Markovienne Pour une Reconnaissance Robuste de Mots isolés dans un Environnement Acoustique Variable'', PhD thesis, National School of Engineering of Sfax, Tunisia.

Frikha M, Ben Massaoud S, Kammoun MA, Gargouri D, Lahyani M, Ben Hamida A (2005). "Optimizing Some HMM Model Parameters in an Isolated Speech Recognition System'', IEEE conference SSD05, Sousse, Tunisia.

Furui S (1986). "Speaker independent isolated word recognizer using dynamic features of speech spectrum''. IEEE Trans. Acoust. Speech Signal Process., 34(1): 52–59.

Hermansky H (1990). ''Perceptual Linear Predictive (PLP) Analysis of Speech''. J. Acoust. Soc. Am., pp. 1738-1752.

Jankowski Jr. CR, Vo HHD., Lippmann RP (1995)., "A Comparison of Signal Processing Front Ends for Automatic Word Recognition". IEEE Trans. Speech Audio Process., 3: 286–293.

Jelinek F (1976). ''Continuous speech recognition by statistical methods''. IEEE Proc., 64(4): 532–556.

Kammoun MA, Gargouri D, Frikha M, Ben Hamida A (2006). "Cepstrum vs. LPC: A Comparative Study for Speech Formant Frequencies Estimation", GESTS Int. Trans. Commun. Signal Proc., 9(1): 87-102.

Magdi MA, Gader P (2000), " Generalized Hidden Markov Models — Part I: Theoretical Frameworks". IEEE Trans. Fuzzy Syst., 8(1): 67-80.

Makhoul JI (1975), ''Linear Prediction: A Tutorial Review'', Proc. IEEE, 63(4): 561-580.

Morgan N, Wu SL, Bourlard H (1995). ''Digit Recognition with Stochastic Perceptual Models'', Proc. Eurospeech' 95, Madrid, Spain, pp. 771–774.

Picone J (1990). "Continuous Speech Recognition Using Hidden Markov Models". IEEE ASP Mag., pp. 26-40.

Picone J (1993). "Signal Modeling Techniques in Speech Recognition", *Proceedings of the IEEE*, 8(9): 1215-1247.

Rabiner LR (1989). ''A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition'', Proc. IEEE, 77(2): 257–285.

Rabiner LR, Juang B (1993), ''Fundamentals of Speech Recognition". Prentice Hall, Englewood Cliffs, New Jersey, USA.

Rabiner LR, Juang BH (1986). ''An Introduction to Hidden Markov Models'', IEEE ASSP Mag., pp. 4-16.

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) (1990). Training and Test Data and Speech Header Software NIST Speech Disc CD1-1.1.

Young S, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P (2002). ''The HTK Book'', *University of Cambridge*, United Kingdom.