International Journal of Control, Energy and Electrical Engineering (CEEE) Copyright – IPCO-2014
Vol.1, pp.57-61

Phoneme Recognition using Hidden Markov Models:
Evaluation with signal parameterization techniques

Ines BEN FREDJ and Kaïs OUNI

Research Unit Signals and Mechatronic Systems SMS,
Higher School of Technology and Computer Science,
Carthage University, Tunis, Tunisia.
ines_benfredj@yahoo.fr
kais.ouni@esti.rnu.tn

***Abstract* — HMM applications show that they are an effective and powerful tool for modelling especially stochastic signals. For this reason, we use HMM for Timit phoneme recognition. The main goal is to study the performance of an HMM phoneme recognizer to fix on an optimal signal parameters. So, we apply different techniques of speech parameterization such as MFCC, LPCC and PLP. Then, we compare the recognition rates obtained to check optimal features. We varied coefficient number of each sample from 12 to 39 for all features. Experimental results show that 39 PLP is the most appropriate parameters for our recognizer.**

***Keywords*— HMM, HTK, LPCC, MFCC, PLP, TIMIT**

I. INTRODUCTION

Speech recognition has become a topic of the most interesting research in signal processing. This is due particularly to the performance of processing and calculating that enable today's computers. Indeed, the availability of many commercial products in this area is the result of research, and especially the fact that computers have become faster and more accessible [1].

On another hand, actual processes usually produce outputs as signals that can be either discrete as quantized vectors from a set of values or continuous as a sample speech [2].

A fundamental problem is to characterize these signals in terms of a signal model. This fact because that a model can provide a notional explanation of a real signal and consequently control the system outputs [3] [4].

For this purpose, there are two types of modelling: deterministic and stochastic.

The deterministic model operates generally some properties of the signal to the model, such as the waveform. Also, stochastic modelling tries to determine the statistical characteristics of the signal. In this case, a probability distribution functions are introduced such as the Poisson and the Gaussian functions [5].

For speech processing applications, both models have provided relevant results.

In this study, we will focus on one type of stochastic modelling, namely the hidden Markov model (HMM).

Mainly, the HMM, introduced late 60s, early 70s, became the perfect solution to the speech recognition problems [6] [7].

In fact, these models are rich in mathematical structure and can be used in a wide range of applications.

So, these models give great results when they are correctly applied.

Many researchers have focused on different ways in applying HMM for speech recognition. In each case, the nature of the data and the parameters selected for the HMM make the difference of the results obtained [8] [9] [10].

For this purpose, the present work was prepared. We are interested to phoneme recognition of Timit database using HMM. For evaluation, we used different speech parameterization techniques such as MFCC, LPCC and PLP. This evaluation aims essentially to fix and choose the most appropriate features for the HMM recognizer.

The paper is organized as follows: In the next section, we present the concept of the HMM followed by an overview of the speech parameterization techniques used.

After that, we describe in section 3 Timit database and the Hidden Markov Model Toolkit (HTK). Then, we explain training and recognition stages. We expose and comment later experimental results and we finish by conclusions and some future works.

II. HIDDEN MARKOV MODELS: HMM

Hidden Markov Models are mostly used in speech recognition. HMM are probabilistic models useful for modelling stochastic sequence with underlying finite state structure. Indeed, these models are an intense mathematical structure which explains the remarkable results that they give.

An HMM is characterized by the number of states, the functions of observation and the transition probability between states.

In fact, the main goal is to determine the probability of a sequence of observations $O = o_1, o_2, ...o_N$ where N is the length of the sequence. An HMM with "n" states $S = s_1, s_2, ..., s_n$ can be presented by a set of parameters λ = { π , A, B} where:

- π represent the initial distribution probability that describes the probability division of the observation symbol in the initial moment noting $\sum_{i=1}^{n} \pi_i = 1$ and $\pi_i >= 0$.
- A is the transition probability matrix { $a_{i,j}$ | i,j=1,2,...,n} where $a_{i,j}$ is the probability of transition from state "i" to state "j" noting $\sum_{j=1}^{n} a_{i,j} = 1$ and $a_{i,j} >= 0$.
- B is the observation matrix { $b_{i,k}$ | i=1,2,...,n , k=1,2,...,m} where $b_{i,k}$ is the probability of observation symbol with index "k" emitted by the current state "i", "m" is the number of observation symbols, $\sum_{k=1}^{m} b_{i,k} = 1$, $b_{i,k} >= 0$ and "n" as noted is the number of states.

As well HMM are widely used in speech recognition such as their powerful adaptation to the variability of the observation.

III. Speech Parameterization Techniques

A. MFCC

The analysis MFCC consists of the evaluation of Cepstral Coefficients from a frequency distribution according to the Mel scale [3].

The algorithm of MFCC is as follows:

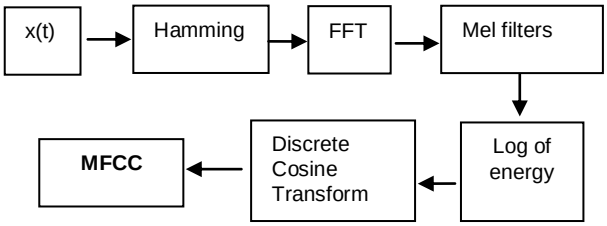

Fig. 1 MFCC algorithm

We take the Fourier transform of a signal windowed by the hamming window. We map the powers of the spectrum obtained above onto the Mel scale. We take the logs of the powers at each of the Mel frequencies.

We get the discrete cosine transform of the list of Mel log powers to obtain the MFCC coefficients.

B. LPCC

The LPCC can be calculated from the LPC signal analysis by a recursive procedure [11]. In other words, they are converted to LPC cepstrum coefficients.

LPCC algorithm is described in figure 4.

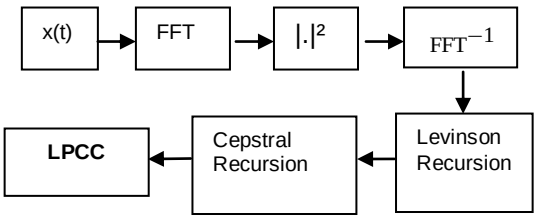

Fig. 2 LPCC algorithm

First, Fourier transform of the signal is applied. Then, calculating the inverse Fourier transform of its module squared. Finally, we pass to Levinson and cepstral recursion for getting LPCC coefficients.

C. PLP

PLP was studied by Hermansky in 1990 [12]. This technique is based on concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum: the critical-band spectral resolution, the equal-loudness curve and the intensity-loudness power law.

The power spectrum is obtained with a Bark filter bank with a subsequent equal loudness pre-emphasis and a compression based on cube-root.

The auditory spectrum is then approximated by an auto-regressive all-pole model.

PLP algorithm is presented as follows:

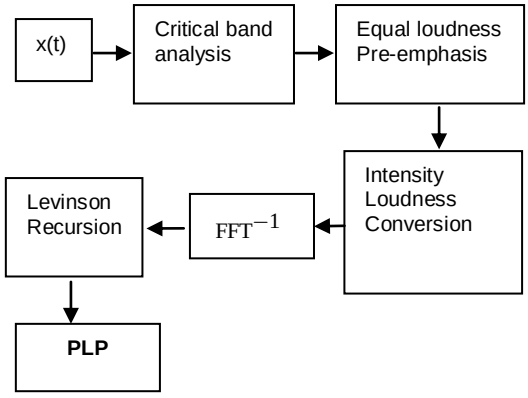

Fig. 3 PLP algorithm

IV. Materiel

A. Database

TIMIT database is used to train and evaluate speaker-independent phoneme recognizers. It consists of 630 speakers from 8 major dialect regions of the United States; each saying 10 sentences which gives 6300 sentences.

Table I describes the structure of Timit corpus [13].

TABLE I
TIMIT CORPUS

Dialect	Designation	Speakers number	
		Male	Female
DR1	New England	31	18

DR2	Northern	71	31
DR3	North Midland	79	23
DR4	South Midland	69	31
DR5	Southern	62	36
DR6	NewYork City	30	16
DR7	Western	74	26
DR8	Army Brat (moved round)	22	11

All dialects of TIMIT speech corpus sampled in 16 kHz were used.

In addition, the database was organized into six phoneme groups which represent vowels, semivowels, affricates, fricatives, stops and nasals classes as illustrated table II.

TABLE II
DISTIBUTION CLASSES OF TIMIT CORPUS

Class	Label
Affricates	/jh/ /ch/
Fricatives	/s/ /sh/ /z/ /zh/ /f/ /th/ /v/ /dh/
Nasals	/m/ /n/ /ng/ /em/ /en/ /eng/ /nx/
Semi-vowels	/l/ /r/ /w/ /y/ /hh/ /hv/ /el/
Stops	/b/ /d/ /g/ /p/ /t/ /k/ /dx/ /q/ /bcl/ /dcl/ /gcl/ /pcl/ /tcl/ /kcl/
Vowels	/iy/ /ih/ /eh/ /ey/ /ae/ /aa/ /aw/ /ay/ /ah/ /ao/ /oy/ /ow/ /uh/ /uw/ /ux/ /er/ /ax/ /ix/ /axr/ /ax-h/
Others	/pau/ /epi/ /h#/ /1/ /2/

Timit corpus was divided into two parts: the first part (about 70%) for the training stage and the second for the recognition stage.

We apply MFCC, LPCC and PLP to obtain a database of cepstral parameters. They were extracted from the speech signal with 256 sample frames and were Hamming windowed in segments of 25 ms length every 10 ms with a sampling frequency equal to 16000 KHz. Coefficients number varies from 12 to 39 including first and second derivatives and energy.

B. Hidden Markov Model Toolkit: HTK

HTK is a portable toolkit for building and manipulating hmms.

The first version of HTK was developed by the Cambridge University Engineering Department (CUED) in 1989 [14].

HTK is mainly used for speech recognition purpose.

HTK consists of a set of library modules and tools available in C source form. It is available on free download, beside with a good and complete documentation.

HTK offers sophisticated solutions for the vocal analysis, the training HMM and the test results.

V. TRAINING AND RECOGNITION

A. Training

The first step is to prepare a dictionary that contains a list of all the possible case of phoneme. Then, the wav files are labelling to mark the beginning and the end of each phoneme and to get a database of labels relative to each sentence.

After features extraction to get a database of MFCC, LPCC and PLP coefficients, we define a prototype HMM for each phoneme since we are interesting on phoneme recognition. A prototype is characterized by the number of states, the functions of observation and the transition probability between states. We have used a prototype of five states defined by the following transition probability matrix:

$$A=\begin{pmatrix} \mathbf{0} & \mathbf{0,6} & \mathbf{0,4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0,6} & \mathbf{0,4} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0,7} & \mathbf{0,3} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Fig. 4 Probability transition matrix of HMMs states

Each HMM is initialized and trained with the corresponding training set to get a model set which will be included for recognition step [15] (see fig 5).

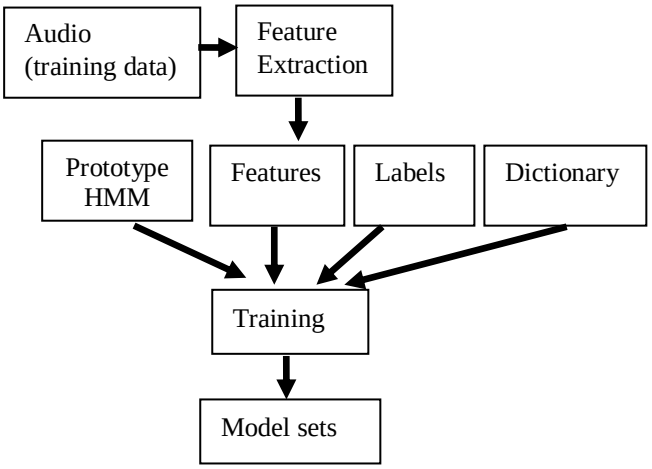

Fig. 5 Training schema

B. Recognition

Before using the model sets obtained by the training step, we have to define the task grammar. It describes all combinations which can form a possible phoneme. Our grammar is illustrated by a start silence, followed by a single phoneme, followed by an end silence. The task grammar has to be compiled to obtain the task network.

At this stage, our speech recognition task completely defined by its network, its dictionary, and its HMM Model set, is ready for use.

Evaluation and recognition should be done on the test data which should be labelled as for the training data.

An input speech signal is first transformed into a series of acoustical vectors, in the same way as what was done with the

training. The input features are then process by a Viterbi algorithm, which matches them against the Markov models recognizer.

The output is stored in a file which contains the transcription of the input.

The performance measures will just result from the comparison between the reference transcription and the recognition hypothesis of each data.

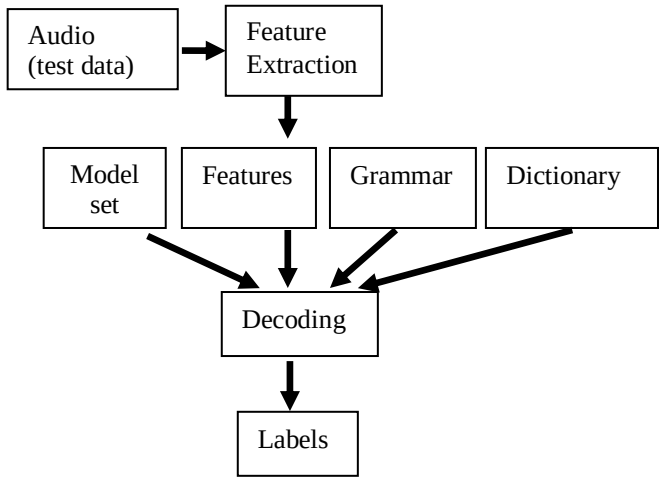

Fig. 6 Recognition schema

VI. EXPERIMENTAL RESULTS

Recognition was applied for all dialects of Timit corpus. Coefficients are varied from 12 to 39 including first and second derivatives and energy.

Results are described as follow.

TABLE III
RECOGNITION RATES USING MFCC (%)

Features / Dialects	MFCC			
	12	**24**	**36**	**39**
DR1	47.22	58.17	61.54	65.42
DR2	50.24	60.79	63.26	67.15
DR3	49.95	60.50	64.09	67.71
DR4	47.91	57.45	60.12	63.85
DR5	47.27	57.75	61.35	64.28
DR6	45.97	56.34	60.91	63.72
DR7	49.45	59.56	62.63	66.24
DR8	47.57	56.04	59.77	63.24
Mean rate	**48.50**	**58.66**	**61.89**	**65.41**

TABLE IV
RECOGNITION RATES USING LPCC (%)

Features / Dialects	LPCC			
	12	**24**	**36**	**39**
DR1	33.55	35.63	37.67	41.71
DR2	33.46	37.00	38.47	44.26
DR3	34.16	36.11	37.74	43.12
DR4	33.02	36.21	38.67	42.25
DR5	31.89	34.96	37.15	41.28
DR6	32.55	35.60	37.74	41.45
DR7	33.76	37.26	39.14	43.65
DR8	33.91	37.66	38.38	42.86
Mean rate	**33.24**	**36.27**	**38.16**	**42.68**

TABLE V
RECOGNITION RATES USING PLP (%)

Features / Dialects	PLP			
	12	**24**	**36**	**39**
DR1	47.63	58.38	62.33	65.85
DR2	49.64	61.03	63.63	67.69
DR3	49.88	61.08	63.48	68.07
DR4	46.40	57.39	60.51	63.88
DR5	46.94	57.58	60.70	64.34
DR6	46.81	57.28	59.95	63.41
DR7	48.88	59.98	62.46	66.83
DR8	46.88	56.38	59.23	62.93
Mean rate	**48.01**	**58.90**	**61.75**	**65.63**

We notice that MFCC and PLP gave very similar results using different number of coefficients. As well, LPCC coefficients have yielded modest results. Also, we see that increasing number of features affects positively recognition rates.

For most dialects, reliable features were firstly 39 PLP, then 39 MFCC. This result was obtained by introducing first and second derivatives and energy. It was established that signal dynamic parameters showed an advantageous ability to improve the recognition task by introducing the transitory characteristics of the speech signal.

This conclusion confirms that the recognizer can run well using dynamic features and energy; this is also assured for all features.

However, we got some low recognition rates for some features and coefficients such as MFCC, LPCC and PLP using 12 coefficients and all the rates obtained with LPCC.

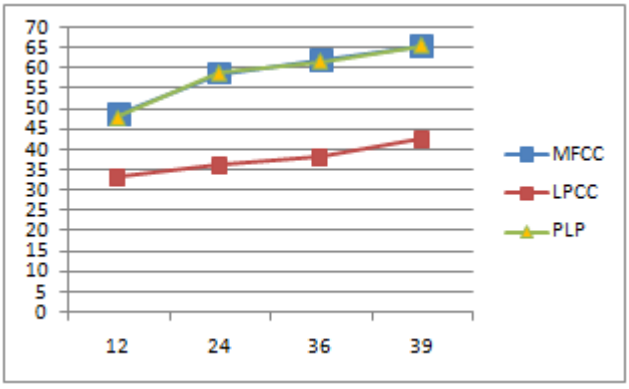

Fig. 7 Comparison of mean recognition rate using MFCC, LPCC and PLP

At last, even features selection, error rates require other solutions to further increase. On the other hand, the difference of the error rates between dialects open a remarkable research topic about the variability influence of speaker accent on speech recognition.

VII. CONCLUSIONS AND FUTURE WORKS

In this work, we presented an approach of phoneme recognition of Timit database using HTK toolkit.

We evaluated the recognizer with different techniques of features extraction such as MFCC, LPCC and PLP.

Number of features varied from 12 to 39 by introducing first and second derivatives and energy to implementing temporal variation.

Results showed the relevance of PLP and MFCC coefficients including signal dynamic coefficients.

Though, LPCC technique remains to be improved.

In future, we will focus to improve this preliminary approach by studying the HMM parameters and find out a reliable prototype.

REFERENCES

[1] B.H. Juang and L.R. Rabiner, "Automatic speech recognition - A brief history of the technology development," *Elsevier Encyclopedia of Language and Linguistics*, 2005.

[2] F.Jelinek, "*Statistical Methods for Speech Recognition*," MIT Press, 1997.

[3] L.Rabiner and B.H. Juang, "*Fundamentals of Speech Recognition*," Prentice Hall,1993.

[4] L.Rabiner, "A tutorial on Hidden Markov Model and Selected Applications in Speech recognition," in *Proc. IEEE*, vol. 77, 1989.

[5] V.Luba and A.Younes, "*Modèles de Markov cachés, Reconnaissance de la parole*," Faculté Polytechnique de Mons, 2004.

[6] L. R. Rabiner, "An introduction to Hidden Markov Models," *IEEE ASSP Magazine*, pp. 4-16, 1986.

[7] C.Gagné, "*Modèles de Markov cachés*, " Université Laval, 2010.

[8] A.G. Veeravalli, W.D. Pan, R. Adhami and P.G. Cox, "A tutoriel on using hidden markov models for phoneme recognition," in *Proc. Thirty-Seventh Southeastern Symposium on System Theory (SSST05)*, pp. 154 – 157, 2005.

[9] J.Picone, "Continuous Speech Recognition Using Hidden Markov Models," In *ASSP Magazine IEEE, vol.* 7, 1990.

[10] M. Gales and S.Young, "The Application of Hidden Markov Models in Speech Recognition," *Foundations and Trends in Signal Processing*, pp. 195-304, vol. 1, 2007.

[11] D. O'Shaughnessy, "Linear predictive coding," In *Potentials, IEEE,* vol.7, pp. 29-32, 1988.

[12] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," The Journal of the Acoustical Society of America, pp. 1738-1752, 1990.

[13] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett and N.L. Dahlgren, "Darpa Timit: Acoustic-phonetic Continuous Speech Corpus," *National Institute of Standards and Technology*, 1993.

[14] S.J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odel, D. Ollason, D.Povey, V. Valtchev and P. Woodland, "*The HTK Book (for HTK Version 3.2)* ," Cambridge University, 2002.

[15] I. Ben Fredj and K. Ouni, Optimization of features parameters for HMM phoneme recognition of TIMIT corpus. In Proc the *International Conference on Control, Engineering & Information Technology*, 2013.