Journal of Engineering and Applied Sciences 2 (11): 1587-1591, 2007
ISSN: 1816-949X

A Matlab Implementation of a Speech Recognition System Using HMM Models

A. Cherif, F. Chabane, M. Talbi and F.L. Agha
Laboratory of Signal Processing, Faculty of Science, 1060 Tunis, Tunisia

Abstract: In this study, we present a speech recognition interface designed for vocal control. The implementation has been realized under the Matlab environment with scripts in C. The program uses the statistical HMM (Hidden Markov Models) for speech modeling, the K-means, Baum-welch algorithms for training and codebook conception and finally the Viterbi decoding algorithm for the recognition process. The recognized word decision is based on the maximal likehood value. The speech database is constituted of 1000 words mono-speaker associated with a denoising module before be applied to the developed interface.

Key words: Speech recognition system, HMM, mono-speaker, Matlab environment, code back conception

INTRODUCTION

Speech processing, coding, synthesis and recognition are considered among the promising applications in telecommunications and data transmission. The difficulty of the automatic recognition depends on the implemented processing method and the acquisition enviroment (Calliope, 1989).

Now, it exists several softwares intended for speech recognition with variable vocabularies, but their performances are still inferior to the desired recognition ratio especially in the case of multi speaker or in the presence of noise perturbation. That's why we will develop a Matlab application with a real time speech acquisition interface which can be easily implemented on a DSP.

SPEECH RECOGNITION SYSTEM

The speech recognition system for isolated words that will be developed is based on HMM speech modelling Let's consider a vocabulary of the number (R) of words to recognize. The recognition method illustrated by Fig. 1 can be divided into the following steps (Boite *et al*., 2000):

- Describing a network which the topology reflects sentences, words of the vocabulary.
- Realizing a training of the database: Words with HMM model : $\lambda = (\pi, A, B)$.
- Computing the maximum likehood.
- Recognition decision.

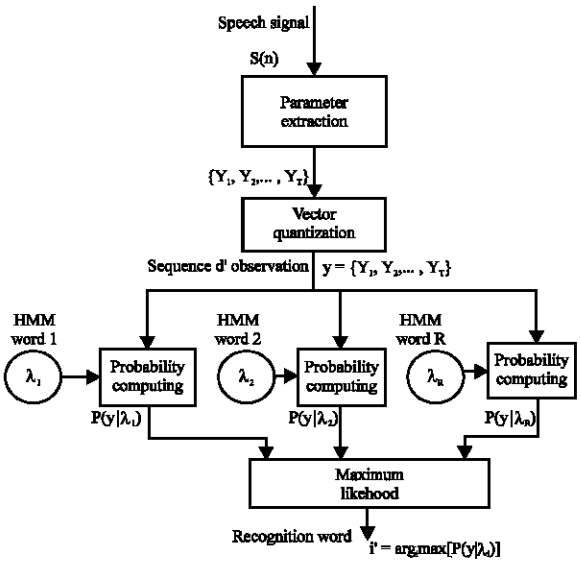

Fig. 1: A speech HMM recognition principle

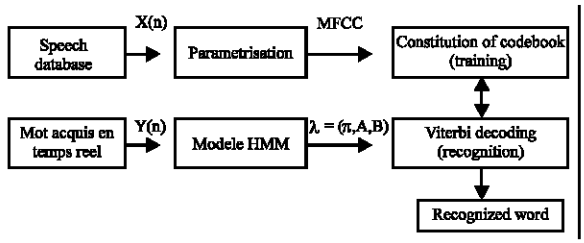

Fig. 2: The speech recognition system

To realize this system, we have developed a vocal control interface constituted of a new toolbox under Matlab. This toolbox uses next occurrences of the words

Corresponding Author: A. Cherif, Laboratory of Signal Processing, Faculty of Science, 1060 Tunis, Tunisia

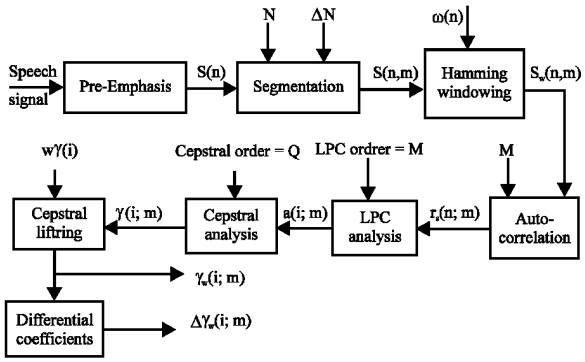

Fig. 3: Parameter extraction of a speech signal

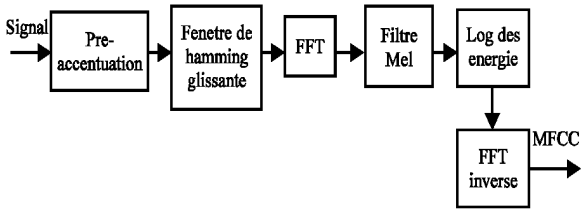

Fig. 4: MFCC parameters of a speech signal

which are the most used for a voice control (start, stop, yes, no, go, help, erase, rubout, repeat, escape,...) extracted from the database.

This interface is based on two modulus : Training and recognition (Fig. 2).

Parameters extraction: The signal parametrisation has for objective to compute the MFCC coefficients of every 20 ms frame. These parametrs characterize every word before be added to the codebook.

The training and recognition is based on the identification of the parameters of a 3 stage HMM (Dutoit, 2002). The acoustic vector parameters of each frame is constituted 16 coefficients (13 coefficients MFCC+1 pitch+1 energie+1 derivative of the energy) Fig. 3.

The cepstral analysis and MFCC coefficients are based on the algorithm of Fig. 4.

Vector quantification: The vector quantification is an operation which allows to represent a vector with N components. It must be organized to minimize quantization errors. Its implementation is conducted by the K-means algorithm in order to surmount the initialisation of the codebook parameters (Rabiner and Juang, 1993).

HMM training: In this phase, each word of the vocabulary will be connected to a hidden Markov model HMM given by

$$\lambda = (\pi, A, B) \quad (1)$$

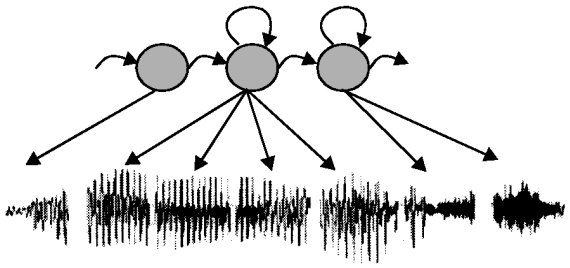

Fig. 5: A three state MMC of a speech signal

with:
Q = {q1, q2, ... qT} is the optimum sequence of the state which has given the sequence of observations y = {y1, y2 ..., yT} then the model is defined by the following parameters (Burno, 1995):

- Number of states S (three in Fig. 5) .
- The number K of observations y(k) of the codebook
- The Matrix A : A=[a_{ij}] of probability transitions:

$$a_{ij} = p(q_t = S_j | q_{t-1} = S_i) \quad (2)$$

- The Matrix B: B = {b_j (k)} of probabilities for each observation:

$$b_j(k) = p(y_t = v_k | q_t = S_j) \quad (3)$$

with: $1 \le j \le S$, $1 \le k \le K$

- The initial probability (p) of each states S.

$$\pi_i = p(q_0 = S_i) \quad (4)$$

with: $1 \le i \le S$

As each word will be represented by the vector y = {y1,y2,...yT} among the K vectors of the code book, the training is resolved by estimating the parameters (A,B,π) of the model l and maximising the probability p(y|λ) (Sakoe and Shiba, 1978; Levinson *et al*., 1988).

RESULTS AND DISCUSSION

We have programmed under MATLAB a GUI interface (Fig. 6) including the algorithms described previously. We have studied the effect of all the parameters on the recognition ratio, such as the codebook number K, the number of HMM states S, the MFCC coefficients. For example, we presented in Fig. 6 and 7 the maximal likehood and recognition ratio of the word "ENTER" .

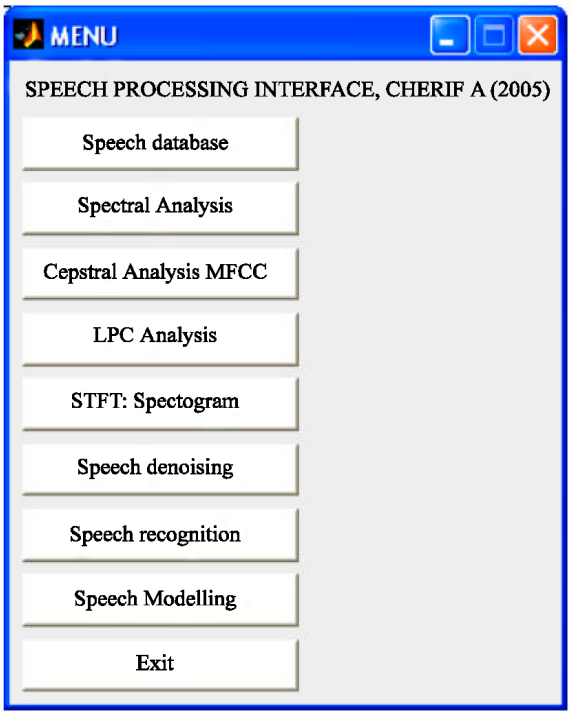

Fig. 6: The ASR main menu

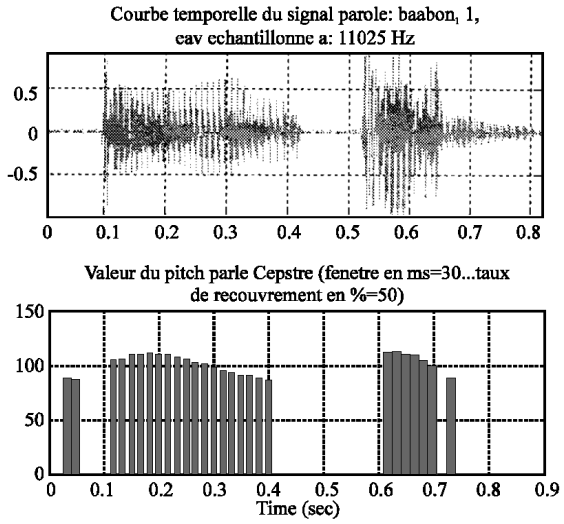

Fig. 7: The pitch evolution with cepstral analysis

The main menu interface: The ASR interface has ten sub-menus for speech analysis and representation, speech denoising, parameters extraction, HMM modeling, training and recognition.

Speech acquisition and analysis: Figure 7- 9 illustrate the speech analysis of a male sound "baabon.wav' with a sample frequency 11025 Hz. For example, Fig. 7 gives the pitch values by using the

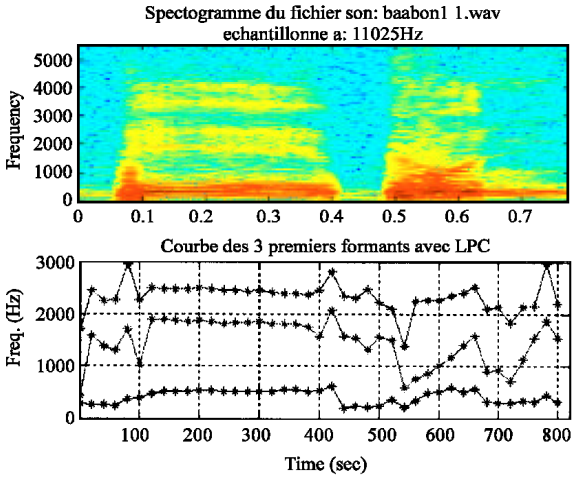

Fig. 8: Spectrogram and the 3 formants extraction

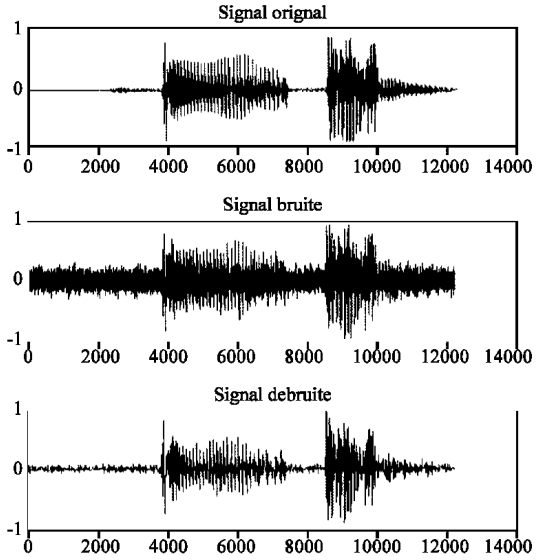

Fig. 9: Speech denoising: Original, noisy and enhanced speech

cepstral method. Yet, Fig. 8 gives the first 3 formants of the same speech signal.

Recognition results: The recognition procedure is based into two stages:
The training and the recognition.

As illustrated in Fig. 10-12, for each stage, there are four steps:

- Speech processing and parametrisation.
- Speech HMM model of the acquired word.
- Maximal likehood ratio computing.
- Recognition decoding and decision.

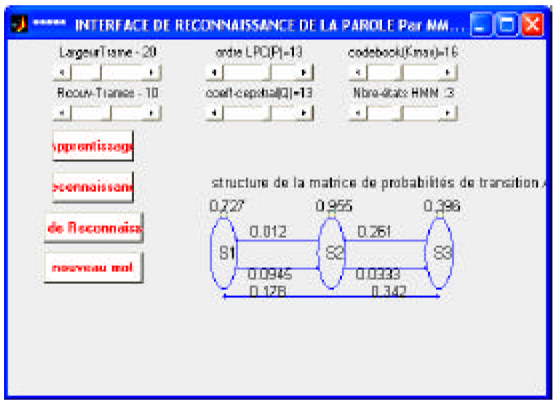

Fig. 10: HMM model of the word ENTER (S = 3)

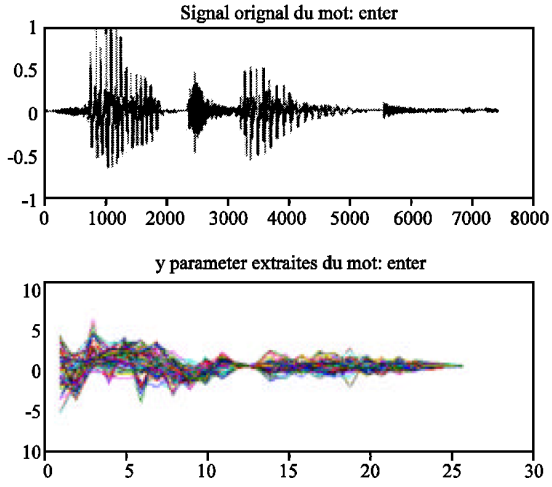

Fig. 11: Parameters extraction of the word 'Enter'

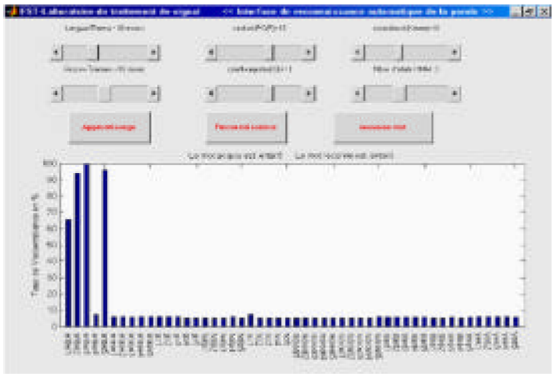

Fig. 12: Maximal likehood ratio of the word: ENTER

After a real time acquisition of the word, a speech processing stage, and a HMM modelling, it is compared with all the codbook models thanks to the Maximal likehood ratio MLR. The maximal value corresponds to the recognected word (Fig. 13: word : Ariana : MLR = 100%).

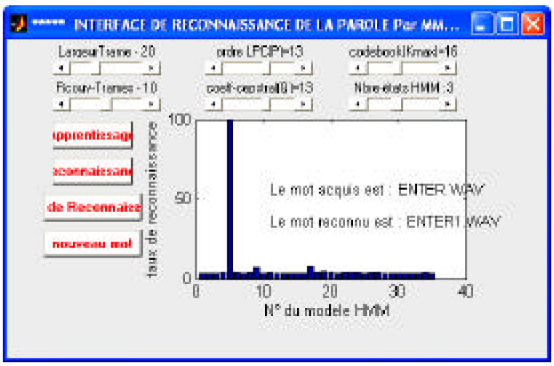

Fig. 13: Maximal likehood ratio of the word 'ARIANA' without noise (SNR= 10dB)

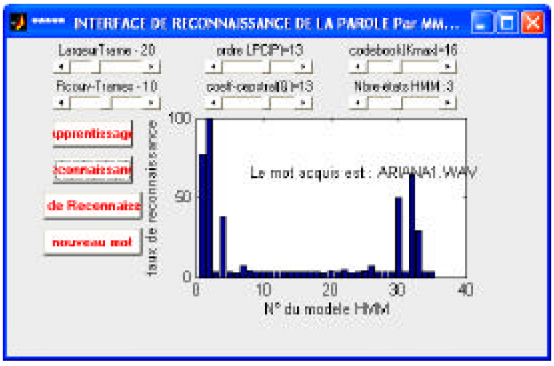

Fig. 14: Maximal likehood ratio of the word 'ARIANA' with noise (SNR= -5 dB)

We have tested our ASR program under a noisy environment (noisy speech such as in Fig. 9).

Figure 13 and 14 demonstrate that the Maximal likehood ratio MLR has a confusion value (60%) for a noisy speech signal. Hence, the recognition ratio is very affected by the speech quality (noisy signal) that's why we have used a denoising program before the speech processing and recognition.

This experience is repeated for all the 100 words of the codebook, we found that 94 words were exactly recognized, which corresponds to a recognition ratio of 94 %.

CONCLUSION

In this study, we succeed to develop a HMM speech recognition interface which was implemented under Matlab and tested in a real time functionning. The simulation results demonstrated that the recognition ratio is performed if the system is associated with a speech denoising module. To resolve this problem we integrated in our interface a denoising program which

uses the wavelets analysis. This deionizing toolbox was presented in a previous study. Besides, with the experimental and simulation results (training tests on 1000 words codebook) we obtained a recognition ratio of 94% in a clean environment (without noise).

REFERENCES

Boite, R., H. Bourlard, T. Dutoit, J. Hancq and H. Leich, 2000. Traitement de la Parole, Presses Polytechniques Universitaires Romandes, Lausanne.

Bruno, J., 1995. Un outil informatique de gestion de Modèles de Markov Cachés: Expérimentations en Reconnaissance Automatique de la Parole. Thèse de l'Institut de Recherche en Informatique de Toulouse.

Calliope, 1989. La parole et son traitement automatique. Masson.

Dutoit, T., 2002. Introduction au Traitement Automatique de la Parole Notes de cours, Faculté Polytechnique de Mons.

Junod, F. and B. Mathieu, 2002. A la découverte des Réseaux de neurones, RN Course, Yverdon, juin.

Levinson, S.E., L.R. Rabiner and M.M. Sondhi, 1988. An introduction to the application of the theory of probabilistic functions of Markov process to A.S. Recognition. The Bell System Technical Journal, Vol. 62.

Rabiner, L. and B.H. Juang, 1993. Fundamentals of Speech Recognition. Edition Prentice Hall.

Sakoe, H. and S. Shiba, 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE. Trans. Acoustics Speech and Signal Proc., 26: 143-165.