Scientific Research and Essays Vol. 6(8), pp. 1808-1816, 18 April, 2011
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE10.761
ISSN 1992-2248 ©2011 Academic Journals

Full Length Research Paper

Spotted words recognition system based on Kalman filter and HMM (hidden Markove model) models to control the movement of the manipulator arm

Ibrahiem M. M. El Emary[1]*, Hamza Atoui[2], Mohamed Fezari[2] and Mouldi Bedda[2]

[1]Information Technology Deanship, King Abdulaziz University, Kingdom of Saudi Arabia, Saudi Arabia.
[2]Laboratory of Automatic and Signals, Department of Electronics, Faculty of Engineering, University of Annaba, Annaba, BP.12, Annaba, 23000, Algeria.

Accepted 20 October, 2010

The main objective of this paper is to implement a real-time speech recognition module to use it in controlling the movements of a five degree of freedom (FOD) manipulator arm using a Kalman filter as a selector in noisy environment and HHM model to recognise 12 Arabic spotted words. The adopted methodology is based on detecting and spotting vocabulary words within a phrase generated by user. The system recognises the spotted words using Kalman filter to select these spotted words then a robust HMM (hidden Markove model) technique with cepstral coefficients is used to improve the recognition rate. To implement the approach on a real-time application, a "personal computer parallel port interface" was designed to control the movement of a set of stepper motors. The user can control the movements of five degree of freedom (DOF) for a robot arm using a vocal phrase containing spotted words.

Key words: Hidden Markove model, kinematics, Kalman filter, recognition, manipulator, robot, degree of freedom, artificial intelligence.

INTRODUCTION AND LITERATURE REVIEW

Human-machine voice interface has a key role in many application fields. Robotics has achieved its greatest success to date in the world of industrial manufacturing. Robot arms, or manipulators, comprise a 2 billion dollar industry. Bolted at its shoulder to a specific position in the assembly line, the robot arm can move with great speed and accuracy to perform repetitive tasks such as spot welding and painting. In the electronics industry, manipulators place surface-mounted components with superhuman precision, making the portable telephone and laptop computer possible (Beritelli et al., 1998; Renals et al., 1994). Moreover, speech recognition constitutes the focus of a large research effort in artificial intelligence (AI), which has led to a large number of new theories and new techniques. Automatic speech recognition performs poorly in noise, especially with crosstalk from other speakers. Humans are very tolerant of noisy environments, but automated speech recognition degrades rapidly as noise increases. Signal corruption from background speech in multiple-speaker environments is particularly troublesome. Biologically inspired neural networks show promise for noise-tolerant spoken-language interfaces in such situations. The problem of single-microphone speech enhancement was extensively studied. Specifically, the use of Kalman filter for estimating both the signal and the parameters is presented in Wan and van der Merwe (2000) and Gannot et al. (1998). By assuming AR model to the speech signal and giving dynamic model to the AR parameters, both dual and joint schemes can be formulated. Each of the two steps comprising the dual scheme is linear, while the joint scheme is consisting of a single nonlinear step. However, it is only recently that the field of robot and AGV navigation have started to import some of the

*Corresponding author. E-mail: omary57@hotmail.com.

existing techniques developed in AI for dealing with uncertain information. Yet, for all of their successes, these commercial robots suffer from a fundamental disadvantage represented by lack of human voice control.

A fixed manipulator has a limited range of commands provided by a manipulator and mainly it uses a keyboard, joystick or a mouse and sometimes a pre-registered program. Serial-link manipulators represent the simplest mechanisms for modelling in Mechanism_Model (http://www.giveng.com/Ergo%20Ams%201ndustrial).
The range of mechanisms on rover platforms that can be modelled as serial-link manipulators include passively set or actively controlled pan and tilt units for cameras, two degrees-of-freedom (DOF), three or four DOF masts arms mounted with navigation camera heads and four, five or six DOF instrument arms. These mechanisms are all mounted to mobile rover platforms (http://www.giveng.com/Ergo%20Ams%201ndustrial).
Instrument arms have the additional kinematic feature of typically holding many instruments on a turret at its end. In research and flight projects at JPL, serial-link manipulators are also deployed from non-mobile lander platforms. In addition to a scoop as its end effectors, the four DOF Mars Phoenix Mars manipulator has a camera attached to its lower arm. It is also desirable that Mechanism_Model be able to also handle parallel-link (or closed-chain) kinematics structures (http://www.giveng.com/Ergo%20Ams%201ndustrial). In Mohamed et al. (2007) a robust adaptive controller for mobile manipulator system in the presence of parametric uncertainties and external disturbances was proposed. The proposed control strategy was designed to drive simultaneously in task space desired end-effectors and platform trajectories without violating the nonholonomic constraints. The unknown parameters and the external disturbances are estimated by using update law in adaptive control scheme (Mohamed et al., 2007). In http://www.insa.fr/lagadic/pdf/2002_ijrr_marcha, a complete framework to control in position and in velocity the effectors of a manipulator using a vision-based control approach was proposed. On the practical side, the primary interest is to use the visual information provided by the camera to efficiently control a manipulator that was not instrumented by proprioceptive sensors (no odometry).

The second interest of this work (http://www.insa.fr/lagadic/pdf/2002_ijrr_marcha) lies in the independence of the position reached by the arm compared to the various modeling errors and calibration errors, of the sensor as well as the arm if it is possible to express the task assigned with the manipulator directly in the space of measurement of the sensor (in fact the image plane of the camera). To achieve manipulator control, the values of the manipulator joints are estimated on-line and the specified displacement is achieved using a 3D visual servoing control law. The orientation of the camera is also controlled by 2D visual servoing so that the effector always remains in the camera field of view. Because of the camera motion, it is appropriate to measure this motion (measured by odometry) and to compute at each iteration of the control the position to reach or the adequate velocity to follow (http://www.insa.fr/lagadic/pdf/2002_ijrr_marcha). This paper is organized as follows: The designed application is described, after which the Kalman filter theory is explained and then the HMM models which cover the Kalman filter are discussed to enhance spoken words. Subsequently, hardware interface design is devoted to the experimental results. Finally, conclusion and future works of this study are presented.

DESCRIPTION OF THE DESIGNED APPLICATION

Our used application in this paper is based on the voice command for a set of five stepper motors to emulate the elements of the robot arm. It therefore involves the recognition of spotted words from a limited vocabulary used to recognise the elements and control the movement of a robot arm. The vocabulary is limited to twelve words divided into two subsets: element name subset necessary to select the part of the robot arm to move and command subset necessary to control the movement of the arm example: turn left, turn right and stop for the base (shoulder), open close and stop for the gripper. The number of words in the vocabulary was kept to a minimum both to make the application simpler and easier for the user. The user selects the robot arm part by its name then gives the movement order on a microphone, connected to sound card of the PC. The user can give the order in a natural language phrase as example: “hey, gripper opens please”. A speech recognition agent based on HMM technique detects the spotted words within the phrase, recognises the words, then the system will generate a byte where the four most significant bits represent a code for the part of the robot arm and the four less significant bits represent the action to be taken by the robot arm. Finally, the byte is sent to the parallel port of the PC and then it is transmitted to the robots via a radio frequency emitter.

The application is first simulated on PC. It includes three phases: the training phase, where a reference pattern file is created, the recognition phase where the decision to generate an accurate action is taken and the appropriate code generation, where the system generates a code of 8 bits on parallel port.

BRIEF KALMAN FILTER THEORY

Kalman filter is an adaptive least square error filter that provides an efficient computational recursive solution for estimating a signal in presence of Gaussian noises. It is

an algorithm which makes optimal use of imprecise data on a linear (or nearly linear) system with Gaussian errors to continuously update the best estimate of the system's current state. The use of Kalman filter for speech enhancement in the form that is presented here was first introduced by Paliwal and Basu (1987). This method is best suitable for reduction of white noise to comply with Kalman assumption. In deriving Kalman equations it is normally assumed that the process noise (the additive noise that is observed in the observation vector) is uncorrelated and has a normal distribution. This assumption leads to whiteness character of this noise. There are, however, different methods developed to fit the Kalman approach to colored noises (Gannot et al., 1998; Paliwal and Basu, 1987). It is assumed that speech signal is stationary during each frame, that is, the AR model of speech remains the same across the segment. To fit the one-dimensional speech signal to the state space model of Kalman filter, we introduce the state vector as:

$$x(k)=(x(k-p+1)\ x(k-p+2)\ x(k-p+3)\ \ldots\ x(k))^T \quad (1)$$

Where x (k) is the speech signal at time k.

Speech signal is contaminated by additive white noise n(k):

$$y(k)=x(k)+n(k) \quad (2)$$

The speech signal could be modelled with an AR process of order p:

$$x(k)=\Sigma a_i x(k-i) + u(k)\ i=1\ldots p \quad (3)$$

Where a_is are AR (LP) coefficients and u(k) is the prediction error which is assumed to have a normal distribution ~N(0,Q).

Substituting Equation 1 into Equation 3 we get:

$$x(k)=Ax(k-1)+Gu(k) \quad (4)$$

Where,

$$G = (0\ 0\ \ldots\ 0\ 1)^T$$

$$\mathbf{A}=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \alpha_p & \alpha_{p-1} & \alpha_{p-2} & \cdots & \alpha_1 \end{bmatrix}$$

G has a length of p (LP order) and the observation equation would be:

$$y(k)=Hx(k)+n(k) \quad (5)$$

$$H=G^T$$

n(k) has a Gaussian distribution ~N(0,R).

HMM models

The speech recognition agent is based on HMM. In this paragraph, a brief definition of HMM is presented and speech processing main blocks are explained. However, a pre-requisite phase is necessary to process a data base composed of twelve vocabulary words repeated twenty times by twenty persons. So, before starting in the creation of parameters, 20*20*12 “wav” files are recoded in a repertory. The training phase will, each utterance (saved wav file) is converted to a 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 optimisation of the likelihood of the training vectors corresponding to each word in the vocabulary. This optimisation is carried by the Baum-Welch algorithm (Ferrer et al., 2000).

HMM basics

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 (Rabiner, 1989), or speech recognition (Rabiner, 1989; Buhler et al., 1994). 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 (Ferrer et al., 2000; Nishimoto et al., 1993). This method is based on the maximum likelihood criterion. To compute the most probable state sequence, the Viterbi algorithm is the most suitable. A HMM model is basically a stochastic finite state automaton, which generates an observation string, that is, the sequence of observation vectors, $O = O_1,..O_t ,\ldots ,O_T$. 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 O_t for each successive transitions from one state S_i to a state S_j. O_t is d dimension and in the discrete

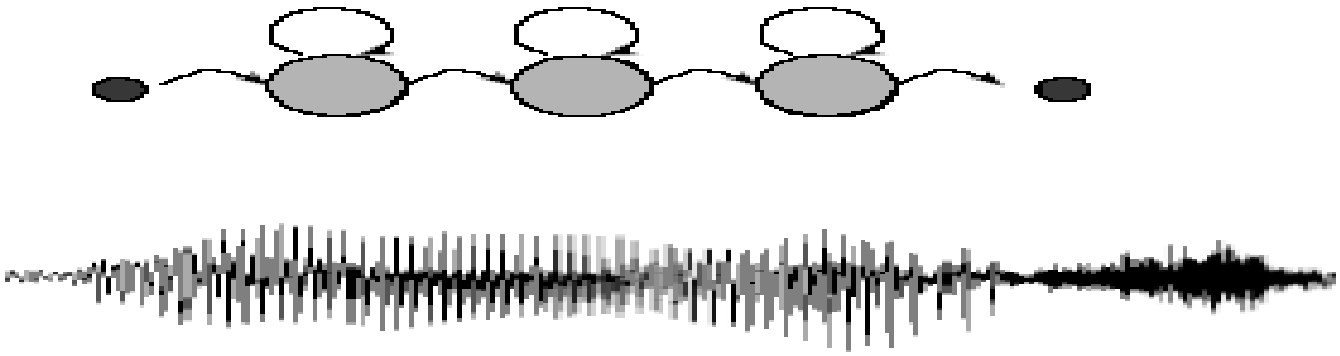


Figure 1. Presentation of left-right (Bakis) HMM.

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 O_t at state S_j is given by B = ($b_j(O_t)$). The probability distribution of the initial state is Π = (π_i).

$$a_{ij} = P(q_{t+1} = S_j / q_t = S_i) \quad (6)$$

B = ($b_j(O_t)$) (7)

$$\pi_i = P(q_0 = S_i) \quad (8)$$

Given an observation O and a HMM model λ = (A, B and ∏), the probability of the observed sequence by the forward-backward procedure P(O/λ) can be computed (Nishimoto et al., 1993). Consequently, the forward variable is defined as the probability of the partial observation sequence $O_1 O_2, ...O_t$ (until time t) and the state S at time t, with the model λ as α(i) and the backward variable is defined as the probability of the partial observation sequence from t+1 to the end, given state S at time t and the model λ as β(i). The probability of the observation sequence is computed as follows:

$$P(O/\lambda) = \sum_{i=1}^{N} \alpha_t(i) * \beta_t(i) = \sum_{i=1}^{N} \alpha_T(i) \quad (9)$$

and the probability of being in state I at time t, given the observation sequence *O* and the model λ is computed as follows:

$$\pi_i = P(q_0 = S_i) \quad (10)$$

The flowchart of a connected HMM is an HMM with all the states linked altogether (every state can be reached from any state). The Bakis HMM is left to right transition HMM with a matrix transition defined as shown in Figure 1.

Speech processing phase

Once the phrase is acquired via a microphone and the PC sound card, the samples are stored in a wav file. Then the speech processing phase is activated. During this phase the signal (samples) goes through different steps: pre-emphasis, frame-blocking, windowing, feature extraction and MFCC analysis.

Pre-emphasis step

The pre-emphasis block 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.

Frame blocking

Since the vocal tract moves mechanically slowly, speech can be assumed to be a random process with slowly varying properties (Beritelli et al., 1998). Hence, the speech is divided into overlapping frames of 20 ms every 10 ms. The speech signal is assumed to be stationary over each frame and this property will prove useful in the following steps:

Windowing: 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 (Figure 2):

$$W(n) = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right) 0 \le n \le N-1 \quad (11)$$

Feature extraction: An important property of feature extraction is the suppression of information irrelevant for correct classification, such as information about speaker

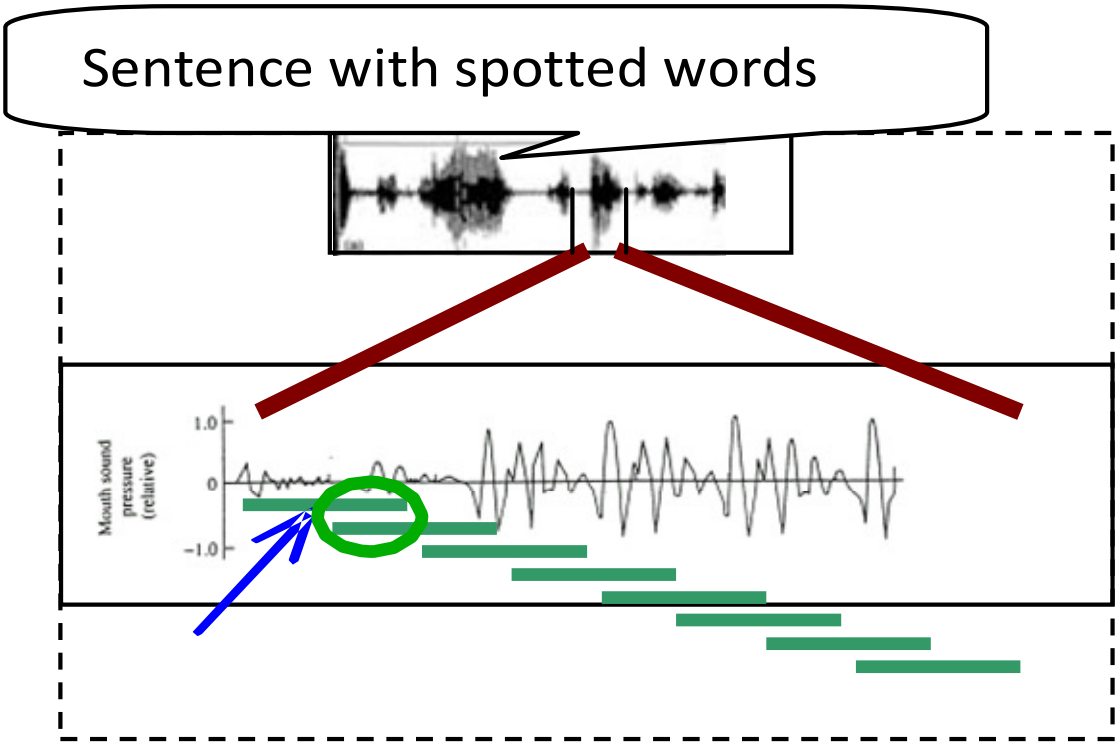


Figure 2. A Windowing.

(for example, fundamental frequency) and information about transmission channel (for example, 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 (Ferrer et al., 2000).

KALMAN FILTER FOR ENHANCING THE SPOKEN WORD

There are several methods for extraction of LP model parameters from noisy data (Wan and van der Merwe, 2000). In this part of the work however, these parameters are assumed to be given so that we can assess the potential of Kalman Filter for speech enhancement without worrying about the extraction of these parameters and the effect of this error on the system. Other methods try to calculate the LP model parameters first and then use them for de-noising the speech signal or iteratively estimate and correct these values and enhance the speech (EM algorithm). A pre-cleaning block may also be used to extract an estimate of these values (like simple spectral subtraction methods). The initial value for x is the noisy data providing the posteriory estimate error covariance matrix with diagonal value of R. The LP coefficients are calculated for segments that might or might not be overlapping. In the latter case special care should be taken to guarantee the continuity of the filter parameters (for example, make sure you store filter parameters midway the segment where you are going to start your next segment filtering, so that you can use these values when going to next segment). It was mentioned in (Gannot et al., 1998) that the use of x(k-p+1) calculated at time k would result in better performance relative to the value that was filtered for the first time (for example, x(k-p+1) calculated at time k-p+1) since more information is incorporated for in calculating this value. The designed agent is presented in Figure 5.

HARDWARE INTERFACE DESIGN

The speech recognition agent based on Kalman filter and HMM will detect words, and process each word (Djemili et al., 2004). Depending on the probability of recognition of the object name and the command word a code will be transmitted to the parallel port of the PC. The vocabulary to be recognized by the system and their meanings are listed as in Table 1, a parallel port interface was designed to show the real-time commands. It is based on the following TTL IC (integrated circuits): a 74LS245 buffer, a microcontroller PIC16F84 and a radio frequency transmitter from RADIOMETRIX TX433-10 (modulation

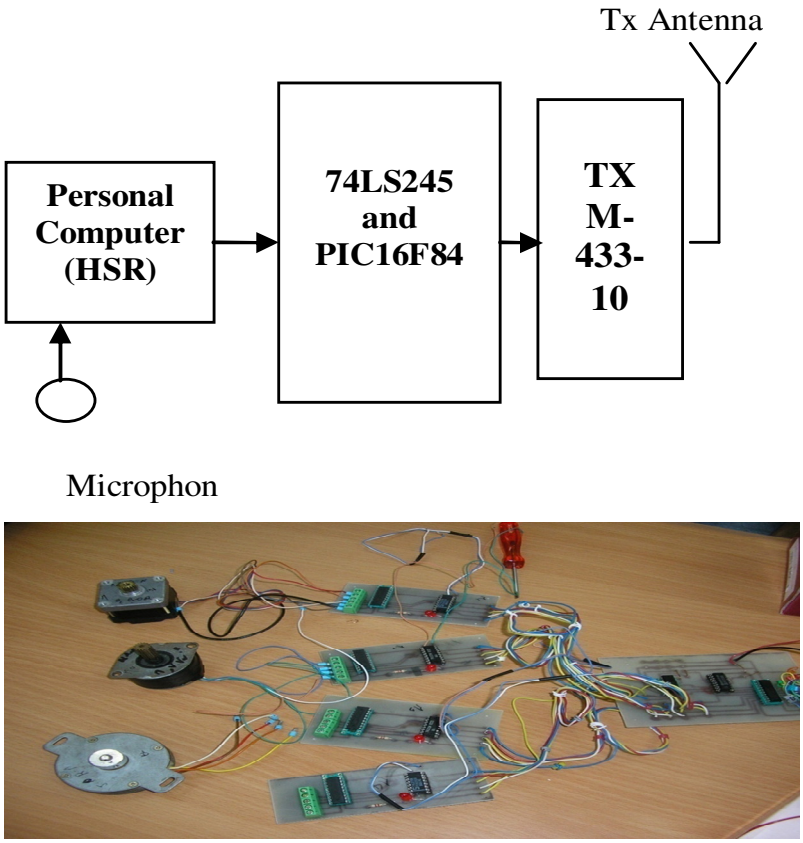


Figure 3. Parallel interface circuit and stepper motors.

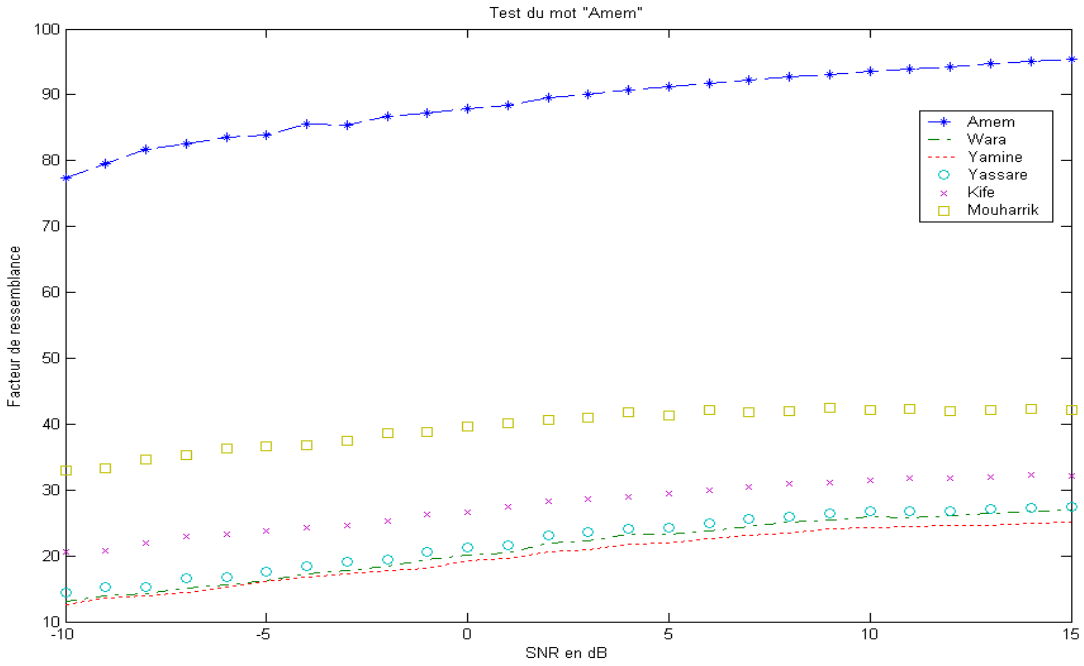
Table 1. The meaning of the vocabulary voice commands.

Korse	Base (M0)
Diraa	Upper limb motor (M1)
Saad	Limb motor (M2)
Meassam	Wrist (hand) motor (M3)
Mikbath	Gripper motor (M4)
Yamine	Left turn (M0)
Yassar	Right turn (M0)
Fawk	Up movement M1, M2 and M3
Tahta	Down movement M1, M2 and M3
Iftah	Open Grip, action on M4
Ighlak	Close grip, action on M4
Kif	Stops movement M0,M1, M2, M3r or M4

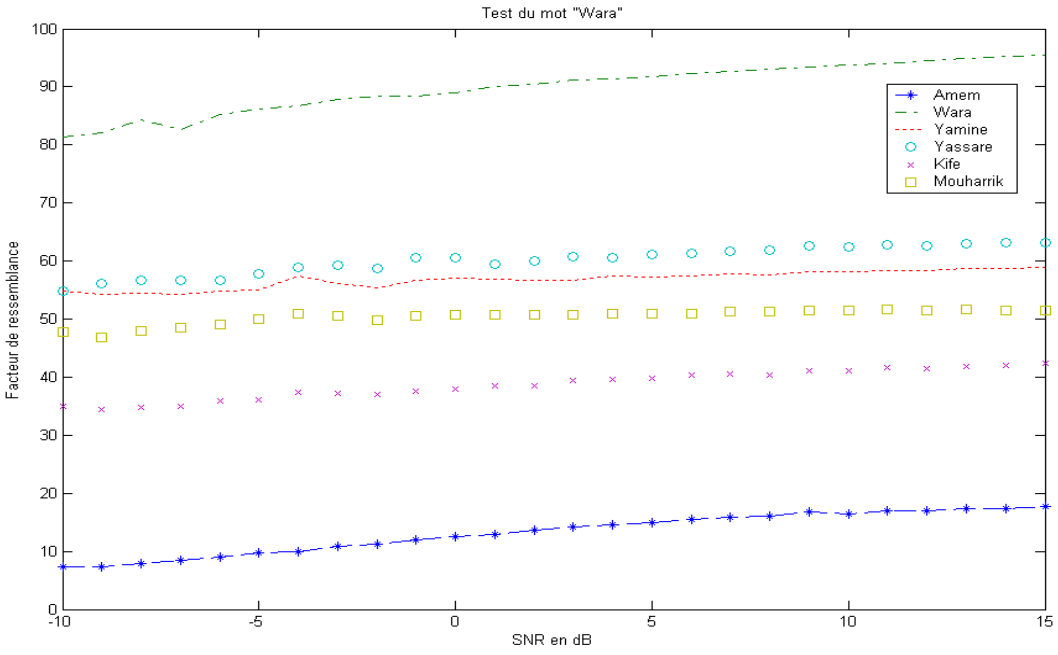
frequency 433 Mhz and transmission rate 10 Kbs) as shown in Figure 3. As shown in Figure 3, the structures of the mechanical hardware and the computer board of the robot arm in this paper are quit similar to MANUS. However, since the robot arm in this paper needs to perform simpler tasks than those in Buhler et al. (1994) do, the computer board of the robot arm consists of a PIC16F84, four power circuits to drive the stepper motors and one H bridges driver using BD134 and BD133 transistors for DC motor to control the gripper, a RF receiver module from RADIOMETRIX which is the SILRX-433-10 and a four bit micro-switch to fix the address of each robot (Kwee, 1997). Each autonomous robot performs the corresponding task to a received command as in Table 1. Commands and their corresponding tasks in autonomous robots may be changed in order to enhance or change the application. In the recognition phase, the application gets the word to be processed, treats the word, and then takes a decision by setting the corresponding bit on the parallel port data register.

EXPERIMENTAL RESULTS

The developed system has been tested within the laboratory of L.A.S.A. There are two different conditions to be tested. Test of the enhancement using Kalman filter are presented here as graphs presented in Figures 4a and b. The Kalman filter selects the clear word from the vocabulary which is close to the noised uttered word. The real-time acquired command words were added a Gaussian noise with ratio of SNR from -10 to 15. The recognition for the word “Amem” is higher than the other words as seen in Figure 4a and same for each word;



(a)



(b)

Figure 4. (a) Result of test of word “Amem”; (b) result of test of word “Yamine”.

however we present the result of just 2 words in Figures 4a and b. Results of the experimental tests are shown in Table 2.

DISCUSSION

A voice command system for robot arm is proposed and is implemented based on a hybrid model HMM/GMM for spotted words. The results of the tests shows that a better recognition rate can be achieved using hybrid techniques and especially if the phonemes of the selected word for voice command are quite different. The effect of the used microphone for tests is proved in the results. However, a good position of the microphone and additional filtering may enhance the recognition rate. The

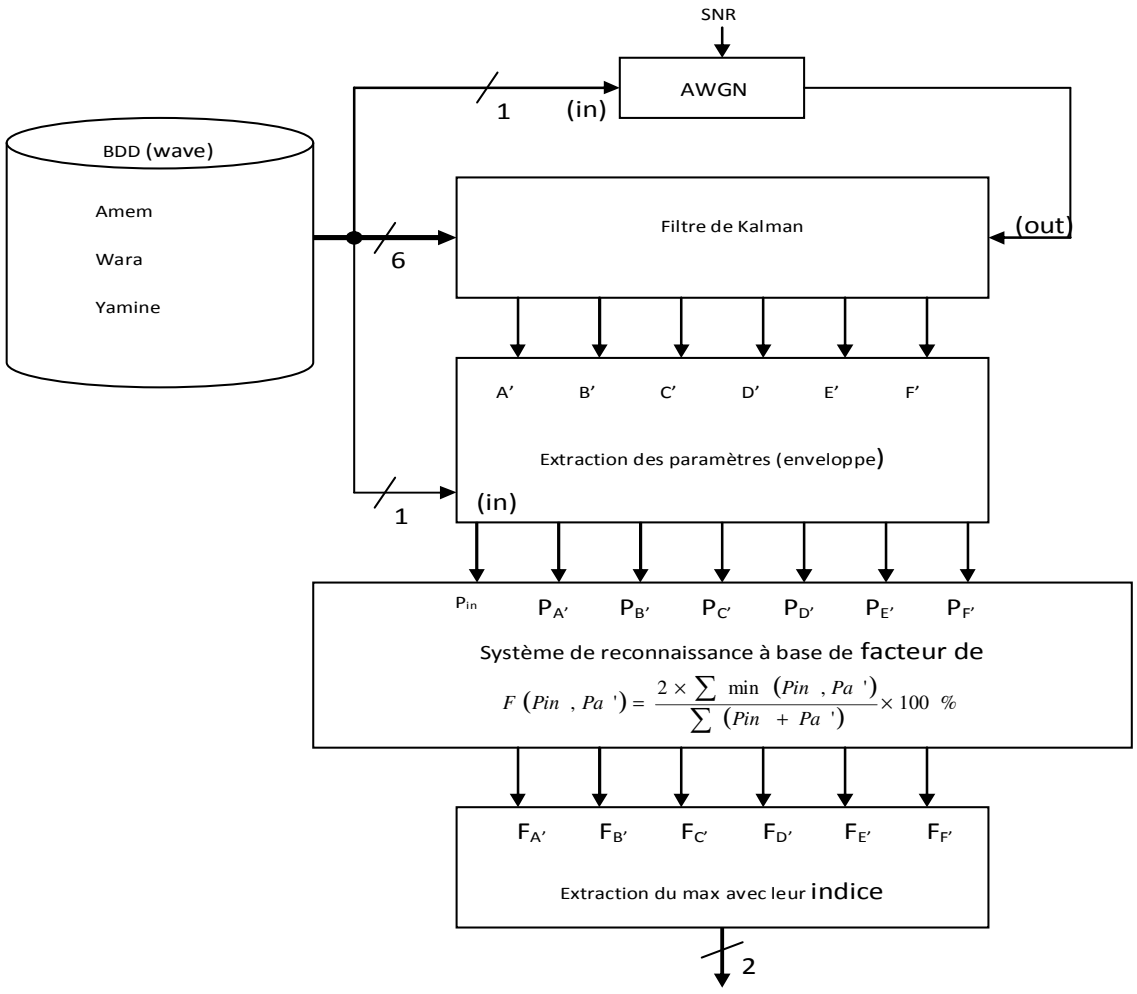


Figure 5. Bloc diagram using Kalamn filter.

Table 2. Results of off-line tests.

Vocabulary	Rate (%)	Comment (confuse with)
Korse	96	
Diraa	83	3, 4 and 10
Saad	93	
Meassam	88	2 and 5
Mikbath	95	
Yamine	94	
Yassar	92	
Amam	88	9, 10 and 1
Wara	87	3 and 7
Iftah	*90*	
Ighlak	89	
Kif	80	2, 3 and 10

HMM based model gives better results than GMM independently, by combining GMM and HMM and using as features MFCC and differentials we increased the recognition rate. The application is speaker independent. However by computing parameters based on speakers' pronunciation the system can be speaker dependant. Off-line test for the HMM: where the system is tested based on pre-registered words from the vocabulary, these words were taken from the database. Within a program test, the words are taken randomly and tested on the system. Each word is tested 100 times. In this case the rate of recognition is acceptable. In the off-line test the words are recorded in same condition of that used in training phase (same place, same conditions and same material: microphone and PC).

CONCLUSION AND FUTURE WORK

An enhancement of the voice command system for manipulator is investigated and then is implemented based on Kalman filter to enhance the quality of the noised signal and then HMM model for spotted words

recognition. Since the designed electronic command for the robot arm consists of a microcontroller and other low-cost components namely RF transmitters, the hardware design can easily be carried out. The results of the tests shows that a better recognition rate can be achieved inside the laboratory and especially if the phonemes of the selected word for voice command are quite different. However, a good position of the microphone and additional filtering may enhance the recognition rate. Several interesting applications of the proposed system different from previous ones are possible, such as command of a set of autonomous robots or a set of home electronic goods. The HMM based model gives better results than DTW (dynamic time warping) or crossing zero and extremums approach. Spotted words detection is based on speech detection then processing of the detected words as isolated word recognition.

Once the filter computes and selects desired words and the parameters were computed, the idea can be implemented easily within a hybrid design using a DSP with a microcontroller since it does not need too much memory capacity.

REFERENCES

Beritelli F, Casale S, Cavallaro A (1998). “A Robust Voice Activity Detector for Wireless Communications Using Soft Computing”, IEEE J. Selected Areas Commun., (JSAC), special Issue on Signal Processing for Wireless Commun., 16: 9.

Buhler C, Heck H, Nedza J, Schulte D (1994). “MANUS wheelchair-Mountable Manipulator- Further Developments and Tests”, Manus Usergroup Mag., 2(1): 9-22.

Djemili R, Bedda M, Bourouba H (2004). “Recognition Of Spoken Arabic Digits Using Neural Predictive Hidden Markov Models” Int. Arab J. Inf. Technol. IAJIT., 1(2): 226-233.

Ferrer MA, Alonso I, Travieso C (2000). “Influence of initialization and Stop Criteria on HMM based recognizers”, Electronics Lett. IEE., 36: 1165-1166

Gannot S, Burshtein D, Weinstein E (1998). “Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms,” IEEE Trans. Speech Audio Proc., 6(4): 373-385.

http://www.giveng.com/Ergo%20Ams%201ndustrial

http://www.insa.fr/lagadic/pdf/2002_ijrr_marcha

Kwee H (1997). “Intelligent control of Manus Wheelchair”, in proceedings Conference on Reabilitation Robotics, ICORR’97, Bath, pp. 91-94.

Mohamed B, Tarek D, Mohamed J (2007). Robust Adaptive Control for Mobile Manipulator, Int. J. Autom. Comput., p. 3.

Nishimoto T, Tokuya N, Nobutashi S, Tetsunori K, Katsuhiko S (1993). Improving human interface in drawing tool using speech, mouse and Key-board, Proceedings of the 4th IEEE International Workshop on Robot and Human Communication, Tokyo, Japan, pp. 107-112.

Paliwal KK, Basu A (1987). A speech enhancement method based on Kalman filtering, in Proc. Int. Conf. Acoust, Speech, Signal Processing, pp. 177-180.

Rabiner L (1989). “Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition Readings in Speech Recognition”, chapter A, pp. 267-295.

Renals S, Morgan N, Bourlard H, Cohen M, Franco H (1994). “Connectionist probability estimators in HMM speech recognition”, IEEE Trans. Speech Audio Process., 2(1): 161-174.

Wan EA, van der Merwe R (2000). “The Unscented Kalman Filter for Nonlinear Estimation,” in Symposium 2000 on Adaptive Systems for Signal Processing, Communication and Control (AS-SPCC), Lake Louise, Alberta, Canada, IEEE, pp. 34.13.1 - 34.13.10.