International Journal of Soft Computing 3 (2): 106-111, 2008
ISSN: 1816-9503

Design of Speech Control System Based on HMM Approach Applied to a Robot Arm

Mohamed Fezari
Laboratory of Automatic and Signals, Department of Electronics, Faculty of Engineering,
University of Annaba Annaba, BP.12, Annaba, 23000, Algeria

Abstract: A voice command system for a robot arm is designed as a part of a research project. The methodology adopted is based on a spotted words recognition system based on a robust HMM (Hidden Markove Model) technique with cepstral coefficients as parameters used in automatic speech recognition system. 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 designed system user can control the movements of 5 Degree of Freedom (DOF) for a robot arm using a vocal phrase containing spotted words. Other applications are proposed.

Key words: Human-machine interaction, hidden markove model, voice control, stepper motors and robotics

INTRODUCTION

Human-robot voice interface has a key role in many application fields and various studies made in the last few years have given good results in both research and commercial applications (Beritelli *et al*., 1998; Rabiner and Juang, 1993; Roa *et al*., 1998; Gu and Rose, 2001). This study proposes a new approach to the problem of the recognition of spotted words within a phrase, using a statistical approach based on HMM (Rabiner and Juang, 1993; Djemili *et al*., 2004). The increase in complexity as compared to the use of only traditional approach is quite acceptable but not negligible, however the system achieves considerable improvement in the recognition phase, thus facilitating the final decision and reducing the number of errors in decision taken by the voice command guided system.

Speech recognition systems constitute the focus of a large research effort in Artificial Intelligence (AI), which has led to a large number of new theories and new techniques. However, it is only recently that the field of robot and AGV navigation has started to import some of the existing techniques developed in AI for dealing with uncertain information.

HMM is a robust technique developed all applied in pattern recognition. Very interesting results were obtained in isolated words speaker independent recognition system, especially in limited vocabulary. However, the rate of recognition is lower in continuous speaking system. The approach proposed here is to design a system that gets specific words within a large or small phrase, process the selected words (Spots) and then execute an order (Renals *et al*., 1994; Ribner and Ribner, 1989). As application, a set of 4 stepper motors were installed via a PC parallel port interface. The application uses a set of 12 commands in Arabic words, divided in 2 subsets one subset contains the names of main parts of a robot arm (base, arm, fore-arm, wrist (hand) and gripper), the second subset contains the actions that can be taken by one of the parts in subset one (left, right, up, down, stop, open and close). The application should be implemented in a DSP or a micro controller in the future in order to be autonomous. Hongyu *et al*. (2004) has to be robust to any background noise confronted by the system.

The aim of this study is, therefore, the recognition of spotted words from a limited vocabulary in the presence of background noise. The application is speaker-dependent. Therefore, it needs a training phase. It should, however, be pointed out that this limit does not depend on the overall approach but only on the method with which the reference patterns were chosen. So by leaving the approach unaltered and choosing the reference patterns appropriately, this application can be made speaker-independent (Ferrer *et al*., 2000).

As application, a vocal command for a set of stepper motors is chosen. There have been many research projects dealing with robot control and teltoperation of arm minipulators, among these projects, there are some projects that build intelligent systems (Kwee, 1997; Buhler *et al*., 1994). Since, we have seen human-like robots in science fiction movies such as in IROBOT movie, making intelligent robots or intelligent systems became an obsession within the research group. Voice

Corresponding Author: Mohamed Fezari, Laboratory of Automatic and Signals, Department of Electronics, Faculty of Engineering, University of Annaba, Annaba BP.12, Annaba, 23000, Algeria

command needs the recognition of spotted words from a limited vocabulary used in Automatic Guided Vehicle (AGV) system (Nishimoto *et al*., 1993) and in arm manipulator control (Larson, 1999).

DESIGNED APPLICATION DESCRIPTION

The application is based on the voice command for a set of 4 stepper motors. 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 12 words divided into 2 subsets: Object 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 open 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 4 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. In this code, 4 higher bits are used to codify the object names and 4 lower bits are sued to codify the actions. The action is shown in real-time on parallel port interface card that includes a set 4 stepper motors to show what command is taken and a radio frequency emitter.

THE SPEECH RECOGNITION AGENT

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 12 vocabulary words repeated 20 times by 20 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.

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 (Renals *et al*., 1994), or speech recognition (Renals *et al*., 1994; Ferrer *et al*., 2000). 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 (Rabiner and Rabiner, 1989; Ferrer *et al*., 2000). 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 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 Sj is given by $B = \{b_j(O_t)\}$. The probability distribution of initial state is $\Pi = \{\pi_i\}$.

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

$$a_{ij} = P(q_{t+1} = S_j / q_t = S_i) \quad (2)$$

$$\pi_i = P(q_0 = S_i) \quad (3)$$

Given an observation O and a HMM model $\lambda = (A, B, \Pi)$, the probability of the observed sequence by

the forward-backward procedure P(O/λ) can be computed (Kwee, 1997). Consequently, the forward variable is defined as the probability of the partial observation sequence $O_1O_2,....O_t$ (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 follow:

$$P(O/\lambda) = \sum_{i=1}^{N} \alpha_t(i) * \beta_t(i) = \sum_{i=1}^{N} \alpha_T(i) \quad (4)$$

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

$$\pi_i = P(q_0 = S_i) \quad (5)$$

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 (Fig. 1):

$$a_{ij} = 0 \text{ if } j < i$$
$$a_{ij} = 0 \text{ if } j < i + \Delta \quad (6)$$

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

Pre-emphasis step: In general, the digitized speech waveform has a high dynamic range. In order to reduce this range pre-emphasis is applied. By pre-emphasis Beeritelli, 1998, we imply the application of a high pass filter, which is usually a first-order FIR of the form $H(z) = 1 \; a \times z^{-1}$. The pre-emphasize is implemented as a fixed-coefficient filter or as an adaptive one, where the coefficient *a* is adjusted with time according to the autocorrelation values of the speech. The pre-emphasize-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: To minimize the discontinuity of a signal at the beginning and end of each frame, we window each frame frames. 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 (Fig. 2a):

$$w(n) = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N-1 \quad (7)$$

Feature extraction: In this step, speech signal is converted into stream of feature vectors coefficients

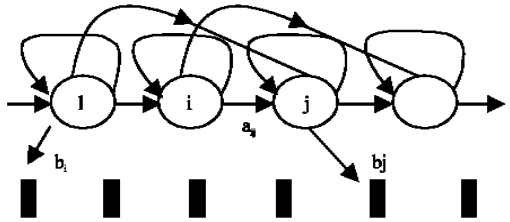

Fig. 1: Presentation of left-right (Bakis) HMM

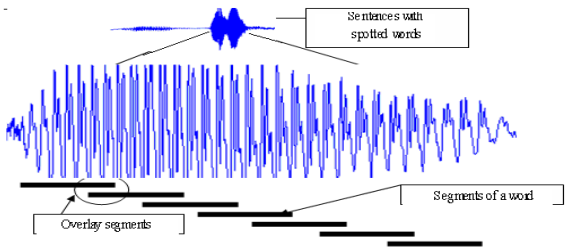

Fig. 2a: Windowing

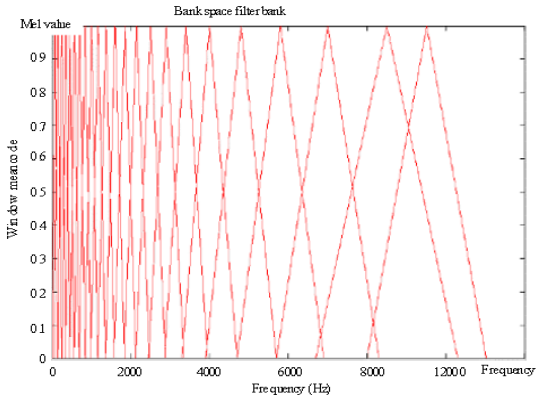

Fig. 2b: Mel-spaced filter Bank

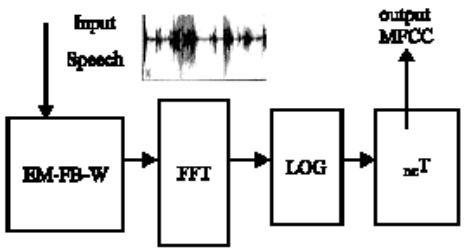

Fig. 3: MFCC design block diagram

which contain only that information about given utterance that is important for its correct recognition. An important property of feature extraction is the suppression of information irrelevant for correct classification, such as information about speaker (e.g., fundamental frequency) and information about transmission channel (e.g., 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 (Rabiner and Rabiner, 1989).

MFCC analysis: The Mel-Filter Cepstral Coefficients are extracted from the speech signal. The speech signal is pre-emphasized, framed and then windowed, usually with a Hamming window. Mel-spaced filter banks are then utilized to get the Melspectrum. Figure 2b shows the Mel-spaced filter banks that are used to get the Mel-spectrum.

The natural Logarithm is then taken to transform into the Cepstral domain and the Discrete Cosine Transform is finally computed to get the MFCCs. as shown in the block diagram of Fig. 3.

Where the acronyms signify:

EM-FB-W : Pre-Emphasis, Frame Blocking and Windowing.
FFT : Fast Fourier Transform.
LOG : Natural Logarithm.
DCT : Discrete Cosine Transform.

$$C_k = \sum_{i=1}^{N} \log(E_i) * \cos\left[\frac{\pi k(i-1/2)}{N}\right] \quad (8)$$

PARALLEL INTERFACE CIRCUIT

The speech recognition agent based on HMM will detect words and process each word. 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. It is obvious that within these words, some are object names and other are command names. The code to be transmitted is composed of 8 bits, 4 most significant bits are used to code the object name and the four least significant bits are used to code the command to be executed by the selected object. The block diagram of the parallel interface is shown in Fig. 4. Example: Robot diraa fawk please.

Table 1: The meaning of the vocabulary voice commands, affected code and motor controlled

Assas (1)	Base (M0)
Diraa (2)	Upper limb motor (M1)
Saad (3)	Limb motor (M2)
Meassam (4)	Wrist (hand) motor (M3)
Mikbath (5)	Gripper motor (M4)
Yamine (1)	Left turn (M0)
Yassar (2)	Right turn (M0)
Fawk (3)	Up movement M1, M2 and M3
Iahta (4)	Down movement M1, M2 and M3
Iftah (5)	Open grip, action on M4
Ighlak (6)	Close grip, action on M4
Kif (7)	Stop the movement, stops M0, M1, M2, M3r or M4

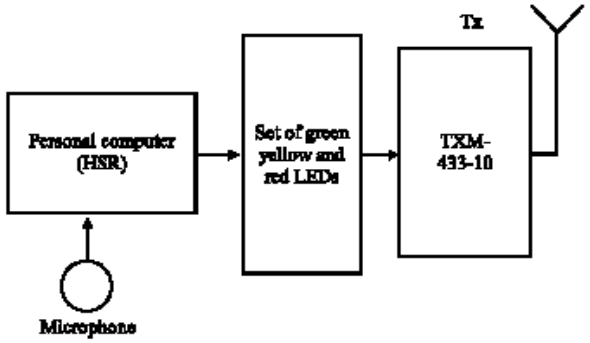

Fig. 4: Parallel interface circuit

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 frequency 433 Mhz and transmission rate 10 Kbs). However, a simulation card were designed to control the set of four stepper motor directly by Matlab software. It is based on a buffer 74LS245, a 74LS138 3/8 decoder and 4 power circuits for the motors.

ROBOT ARM INTERFACE

As in Fig. 5a and b, the structures of the mechanical hardware and the computer board of the robot arm in this

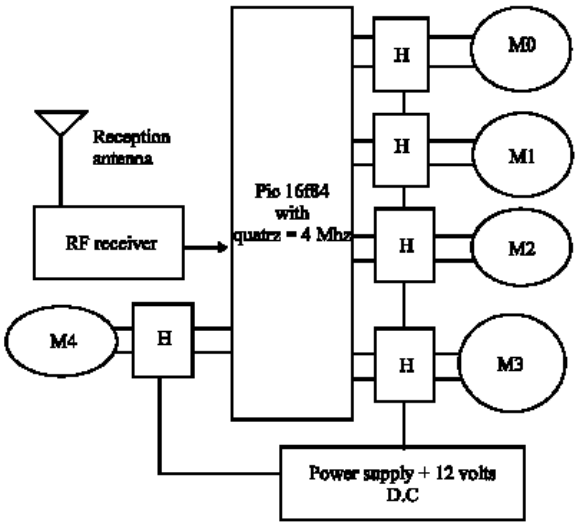

Fig. 5a: Robot Arm block diagram

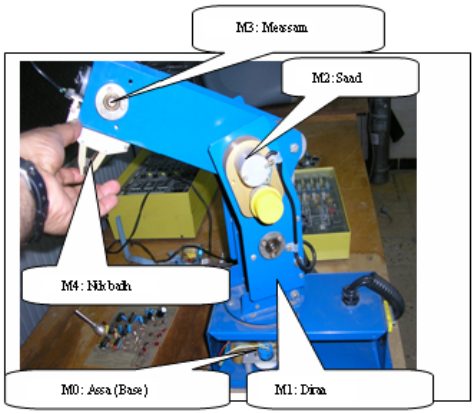

Fig. 5b: Overview of the Robot arm Parallel interface

study is similar to MANUS (Kwee, 1997; Buhler *et al*., 1994; Heck, 1997). However, since, the robot arm in this study needs to perform simpler tasks than those in Mishimoto *et al*. (1993) do, the computer board of the robot arm consists of a PIC16F84, with 1K-instruction EEPROM (Electrically Programmable Read Only Memory) (http://www.microchip.com., 2002) 4 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 (modulation frequency 433MHz and transmission rate is 10 Kbs) (http//www.radiometrix.com). Each motor in the robot arm performs the corresponding task to a received command (example: yamin, kif or Fawk) as in Table 1. Commands and their corresponding tasks in autonomous robots may be changed in order to enhance or change the application.

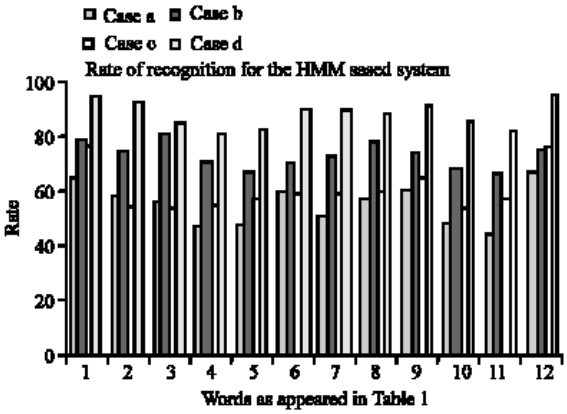

Fig. 6: The effect of PN or NSN, in and out the laboratory

In the recognition phase, the application gets the word to be processed, treats the word, then takes a decision by setting the corresponding bit on the parallel port data register and hence the corresponding LED is on. The code is also transmitted in serial mode to the TXM-433-10.

RESULTS

The developed system has been tested within the laboratory of L.A.S.A, there were 2 different conditions to be tested: the distance of the microphone from the speaker and the rate of recognition in periodic noise and nonstationary noise (NSN) environment . The system, first, had been tested in the laboratory and outside in order to detect the environment effect on the recognition rate, then. After testing the recognition of each word 25 times in the following conditions: outside the Laboratory (LASA) with NSN, outside the LASA with periodic noise (PN), inside the LASA with NSN and inside the LASA with PN. The results are shown in Fig. 6 where the numbers in abscess axe corresponds to the order of voice command word as they appear in Table 1.

CONCLUSION AND FUTURE WORK

A voice command system for robot arm is proposed and is implemented based on a HMM model for spotted words. 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 exremums approach. It is speaker independent. However, by computing parameters based on speakers pronunciation the system can be speaker dependant.

The increase in computational complexity as compared with a traditional approach is, however, negligible. Spotted words detection is based on speech detection then processing of the detected. Once 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. Finally, we notice that by simply changing the set of command words, we can use this system to control other objects by voice command such as an electric wheelchair movements or a set of autonomous robots (Kim *et al*., 1998; Fezari *et al*., 2005).

REFERENCES

Beritelli, F., S. Casale and A. Cavallaro, 1998. A robust voice activity detector for wireless communications using soft computing. IEEE J. Selected Areas in Communications (JSAC), special Issue on Signal Processing for Wireless Communications, 16(9).

Buhler, C., H. Heck, J. Nedza and D. Schulte, 1994. MANUS wheelchair-Mountable Manipulator-Further Devepolements and Tests. Manus usergroup Magazine, 2: 9-22.

Data sheet PIC16F876 from Microchip inc. User's Manual, 2002. http://www.microchip.com.

Djemili, R., M. Bedda and H. Bourouba, 2004. Recognition of spoken arabic digits using neural predictive hidden markov models. Int. Arab J. Inform. Technol. IAJIT, 1: 226-233.

Ferrer, M.A., I. Alonso and C. Travieso, 2000. Influence of initialization and Stop Criteria on HMM based recognizers. Elec. Lett. IEEE., 36: 1165-1166.

Fezari, M., M. Bousbia-Salah and M. Bedda, 2005. Hybrid technique to enhance voice command system for a wheelchair. In Proc. ACIT, Jordan, pp: 102-109.

Gu, L. and K. Rose, 2001. Perceptual harmonic cepstral coefficients for speech recognition in noisy environment. In Proceedings ICASSP.

Heck Helmut, 1997. User Requirements for a personal Assistive Robot. In proceeding of the 1st MobiNet Symposium on Mobile Robotics Technology for Health Care Services, Athens, pp: 121-124.

Kim, W.J. *et al*., 1998. Development of A voice remote control system. Proceedings of the 1998, Korea Automatic Control Conference, Pusan, Korea, pp: 1401-1404.

Kwee Hok, 1997. Intelligent control of Manus Wheelchair. In proceedings Conference on Reabilitation Robotics, ICORR, pp: 91-94.

Larson, M., 1999. Speech Control for Robotic arm within rehabilitation. MSc. thesis, Lund Unversity.

Nishimoto, T. *et al*., 1993. Improving human interface in drawing tool using speech, mouse and Key-board. Proc. 4th IEEE. Int. Workshop on Robot and Human Commun. Tokyo, Japan, pp: 107-112.

Rabiner, L. and B.H. Juang, 1993. Fundamentals of speech recognition, Prentice Hall International.

Rabiner, L. and R. Rabiner, 1989. Tutorial on hidden markov models and selected applications in speech recognition readings in speech recognition. Chapter A, pp: 267-295.

Radiometrix components, 2006. Txm-433 and SILRX-433 Manual, HF Electronics Company. http://www.radiometrix.com.

Rao, R.S., K. Rose and A. Gersho, 1998. Deterministically annealed design of speech recognizers and its performance on isolated letters. Proc. IEEE. ICASSP, pp: 461-464.

Renals, S., N. Morgan, H. Bourlard, M. Cohen and H. Franco, 1994. Connectionist probability estimators in HMM speech recognition. IEEE. Trans. Speech Audio Processing, 2: 161-174.