Voice controller as a basic Human Machine Communication system

Wassila ABADI, Mohamed FEZARI, Rachid HAMDI
Badji Mokhtar Annaba University
Automatic and Signals Laboratory (LASA)
Faculty of Engineering, BP: 12, Annaba, 23000, Algeria
e-mail : wassila.abadi@gmail.com, mohamed fezari@uwe.ac.uk, hamdi_rach@yahoo.fr

Abstract— *Most of HCI (Human Computer Interfaces) applications are based on a graphical interface, mouse and keyboard; in this work we investigated the use of voice activation as input device to control the mouse pointer on the screen. The control of the windows icon mouse pointer (WIMP) by voice command is currently based on using vowel utterances, this category of letters is easy to learn and to be pronounced, especially for individuals who are physically disabled or have a partial voice disorder. In addition, vowels are quite easy to recognize by automatic speech recognition (ASR) systems. In this work we represent a system for the control of mouse cursor based on voice command, using the pronunciation of certain phonemes and short words. The Mel Frequency Cepstral Coefficients (MFCCs) and Predictive Coding (LPC) are selected as distinctive features. The non linear sequence alignment known as Dynamic Time Warping (DTW), Hidden Markov Models (HMMs) have been have been tested as classifiers for matching components (vowels and short words).*

Keywords- Human machine communication; vowels; windows icons mouse pointer; MFCC; DTW; HMMs.

I. INTRODUCTION

Recently, a lot of interest is put on improving all aspects of the interaction between human and computer. Here is some related works on human computer interaction, based on voice activation or control, which can be interested for individuals with motor impairments. Most of concepts of vocal commands are built on the pronunciation of vowels [1, 2, and 3], where the particularity of vowels used is the simple and the regular pronunciation of these phonemes. Many vocal characteristics are exploited in several works, but the most used are: energy [1, 2, 3 and 5], pitch and vowel quality [5, 6] speech rate (number of syllables per second) and volume level [3]. However, Mel Frequency Cepstral Coefficients (MFCCs) [7, 8 and 9] are used significantly of speech processing as bio-inspired features for automatic speech recognition of isolated words [11-12].

This paper is organized as follows:

Section 2, presentation of an overview on related works of mouse cursor control based on vocal commands.

Section 3, presentation of LPC and MFCC computation as features extraction techniques

Description of used classifiers: DTW then HMM are presented in section 4.

Finally in last paragraph, we describe the application, tests scenarios then we finish by presenting results and discussion.

II. RELATED WORKS:

We describe some related works with vocal command system in the literature review. Voice recognition allows you to provide input to an application with your voice. In the basic protocol, each vowel is associated to one direction for pointer motion [1].This technique is useful in situations where the user cannot use his other hands for controlling applications because of permanent physical disability or temporal task-induced disability. The limitation of this technique is that it requires an unnatural way of using the voice [1] [2]. Control by Continuous Voice: In this interface, the user's voice works as an on/off button. When the user is continuously producing vocal sound, the system responds as if the button is being pressed. When the user stops the sound, the system recognizes that the button is released. For example, one can say "Volume up, ahhhhhh", and the volume of a TV set continues to increases while the "ahhh" continues. The advantage of this technique compared with traditional approach of saying "Volume up twenty" or something is that the user can continuously observes the immediate feedback during the interaction. One can also use voiceless, breathed sound [2].

Alex Olwal et al. [3] have been experimenting with non verbal features in a prototype system in which the cursor speed and direction are controlled by speech commands. In one approach, speech commands provide the direction (right, left, up and down) and speech rate controls the cursor speed. Mapping speech rate to cursor speed is easy to understand and allows the user to execute slow. The cursor's speed can be changed while it is moving, by reissuing the command at a different pace. One limitation of using speech features is that they are normally used to convey emotion, rather than for interaction control.

The detection of gestures is based on discrete predesignated symbol sets, which are manually labeled during the training phase. The gesture-speech correlation is modeled by examining the co-occurring speech and gesture patterns. This correlation can be used to fuse gesture and speech modalities for edutainment applications (i.e. video games, 3-D animations) where natural gestures of talking avatars are animated from speech [3] [4].

J. Bilmes et al. [5] have been developed a portable modular library (the Vocal Joystick"VJ" engine) that can be incorporated into a variety of applications such as mouse and menu control, or robotic arm manipulation. Our design goal is to be modular, low-latency, and as computationally efficient as possible. The first of those, localized acoustic energy is used for voice activity detection, and it is normalized relatively to the current detected vowel, and is used by our mouse application to control the velocity of cursor movement. The second parameter, "pitch", is not used currently but it is left for the future use. The third parameter: "vowel quality", where the vowels are characterized by high energetic level [5, 6]. The classification of vowels is realized by extraction of two first formants frequencies, tongue height and tongue advancement [5, 6]. Thus, the VJ research has focused on real time extraction of continuous parameters since that is less like standard ASR technology [5]. The main advantage of VJ is the reaction of the system in real time.

In [10], Thiang et al., described the implementation of speech recognition system on a mobile robot for controlling movement of the robot. The methods used for speech recognition system are Linear Predictive Coding (LPC) and Artificial Neural Network (ANN). LPC method is used for extracting feature of a voice signal and ANN is used as the recognition method. Backpropagation method is used to train the ANN. Experimental results show that the highest recognition rate that can be achieved by this system is 91.4%. This result is obtained by using 25 samples per word, 1 hidden layer, 5 neurons for each hidden layer, and learning rate 0.1.

III. FEATURE EXTRACTION

One way to get better results in automatic speech recognition is to select better and easy to compute features in order to implement the application on embedded system in future, so the features would be robust and easy to compute. The LPC and MFCC with energy were selected based on literature reviews [11, 12].

A. MFCC Feature extraction[7]

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. The overall process of the MFCC can be presented in the following steps:

1. After the pre-emphasis filter, the speech signal is first divided into fixed-size windows distributed uniformly along the signal.

2. The FFT (Fast Fourier Transform) of the frame is calculated. Then the energy is calculated by squaring the value of the FFT. The energy is then passed through each filter Mel. S_k: is the energy of the signal at the output of the filter K, we have now m_p (number of filters) S_k parameters.

3. The logarithm of S_k is calculated.

4. Finally, the coefficients are calculated using the DCT (Discrete Cosine Transform).

$$c_i = \sqrt{\frac{2}{m_p}} \left\{ \sum_{k=1}^{m_p} \log(S_k) \cos\left[i \left(k - \frac{1}{2} \right) \frac{\pi}{m_p} \right] \right\} \quad (1)$$

$$pour\ i = 1 \dots \dots .. N$$

N: is the number of MFCC coefficients.

B. LPC parameters extraction

Linear predictive analysis of speech has become the predominant technique for estimating the basic parameters of speech. Linear predictive analysis provides both an accurate estimate of the speech parameters and also an efficient computational model of speech.

The basic idea behind linear predictive analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and linear predicted values a unique set of parameters or predictor coefficients can be determined.

LPC computation basic steps can be presented as follow [7]:

a) Pre-emphasis: The digitized speech signal, $s(n)$, is put through a low order digital system, to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing.

b) Frame Blocking: The output of pre-emphasis step $\tilde{s}(n)$ is blocked into frames of N samples, with adjacent frames being separated by M samples. If $x_l(n)$ is the l^{th} frame of speech, and there are L frames within entire speech signal.

c) Windowing: After frame blocking, the next step is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. If we define the window as $w(n)$, $0 \le n \le N-1$, then the result of windowing is the signal:

$$\tilde{x}_l(n) = x_l(n) w(n) \quad (2)$$

d) Autocorrelation Analysis: The next step is to auto correlate each frame of windowed signal in order to give:

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n) \tilde{x}_l(n+m) \quad (3)$$

$$m = 0, 1, \dots p$$

e) LPC Analysis: which converts each frame of $p + 1$ autocorrelations into LPC parameter set by using Durbin's method.

f) LPC Parameter Conversion to Cepstral Coefficients: LPC cepstral coefficients, is a very important LPC parameter set, which can be derived directly from the LPC coefficient set. The recursion used is:

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k} \quad (4)$$

$$1 \le m \le p$$

And:

$$c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k} \quad (5)$$

$$m > p$$

The LPC cepstral coefficients are the features that are extracted from voice signal and these coefficients are used as the input data for the classifier (Euclidian Distance or DTW). In this system, voice signal is sampled using sampling frequency of 8 kHz and the signal is sampled within 1.5 seconds, therefore, the sampling process results 1200 data. Because we choose LPC parameter N = 200, m = 100, and LPC order = 10 then there are 119 vector data of LPC cepstral coefficients.

IV. CLASSIFIERS

In pattern recognition in general, automatic speech recognition, speaker Identification, image or shape recognition we need some how an algorithm to classify.

A. DTW(Dynamic Time Warping)

DTW algorithm is based on Dynamic Programming techniques .This algorithm is for measuring similarity between two time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series.

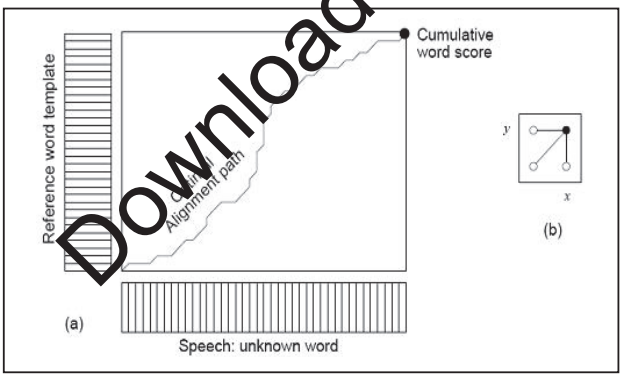

Figure 1. Example of a figure caption. *(figure caption)*

If D(x,y) is the Euclidean distance between frame x of the speech sample and frame y of the reference template, and if C(x,y) is the cumulative score along an optimal alignment path that leads to (x,y), then

$$C(x,y)=MIN(C(x-1,y),C(x-1,y-1),C(x,y-1))+ D(x,y) \quad (6)$$

B. HMMs Basics [13]

Over the past years, Hidden Markov Models have been widely applied in several models like pattern, or speech recognition. 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,A,B) for the HMM. This method is based on the maximum likelihood criterion. To compute the most probable state sequence, the Viterbi algorithm is the most suitable.

An HMM model is basically a stochastic finite state automaton, which generates an observation string, that is, the sequence of observation vectors O = O1, .. Ot, ... ,OT . Thus, a HMM model consists of a number of N states S={Si} and of the observation string produced as a result of emitting a vector Ot for each successive transitions from one state Si to a state Sj. Ot is d dimensional and in the discrete case takes its values in a library of M symbols.

The state transition probability distribution between state Si to Sj is A={aij}, and the observation probability distribution of emitting any vector Ot at state Sj is given by B={bj(Ot)}. The probability distribution of initial state is Π={ πi}.

$$a_{ij} = P(q_{t+1} = \frac{S_j}{q_t} = S_i) \quad (7)$$

$$B = \{b_j(O_t)\} \quad (8)$$

$$\pi_i = P(q_0 = S_i) \quad (9)$$

Given an observation O and a HMM model λ=(A,B,Π), the probability of the observed sequence by the forward-backward procedure P(O/λ) can be computed. Consequently, the forward variable is defined as the probability of the partial observation sequence O1, O2, ...,Ot (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:

$$p(o/\lambda) = \sum_{i=1}^{N} \alpha_t(i) * \beta_t(i) = \sum_{i=1}^{N} \alpha_T(i) \quad (10)$$

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

$$\pi_i = P(q_0 = S_i) \quad (11)$$

V. DESCRIPTION OF APPLICATION

The application is designed to control the mouse cursor by using the pronunciation of certain phonemes and words, which we chose as vocabulary: " aaa", " ooh", " iii"," eeu'", " Clic", " stop".

The choice of these phonemes and words is based on the following criteria:

- Easy to learn.
- Easy to pronounce.
- Easy to recognize by the system of automatic recognition of speech.

A. DataBase Description

The database consists of 5 women, 7 men, and 3 children, each speaker had: 5 trials for each phoneme or word. Collection of the database is performed in a quiet room without noise.

B. The parametrization

According to the tests, we found that the parameters more robust to noise than other parameters are the LPC coefficients and Mel Frequency Cepstral Coefficients (MFCCs).
The input signal is segmented by a window of 20 ms overlapping 10ms, from each segment parameters were extracted by both methods LPC (the order of the prediction: 10) then MFCC (42 coefficients: Energy and derivative and second derivatives).

C. Classification

For this moment, we have used: Dynamic Time Warping (DTW) and Hidden Markov chains (HMM) for classification phase.

For Hidden Markov models, in our system, we utilize left-to-right HMM structures with 3 states and 3 mixtures are used to model MFCCs coefficients.

D. Application

Our application is used to control the cursor by voice, pronouncing a phoneme or vocabulary words above. The directions of the mouse are:

- Up:" ooh"
- Down:" aah"
- To the right:" iii"
- Left:" eeu"
- To double-click (open):" clic"
- To exit the application by voice command:" stop"

VI. Results and Discussions

For the testing phase 20% of recorded sounds are selected for each phoneme or word.

In order to see the effect of training and making the system speaker independent, different scenarios for the tests were done, here we choose the results of recognition of three users out of database.

Some phonemes and words were correctly classified with some confusion, where a phoneme (or word) test classified as another phoneme (or word), the misclassification presented in the tables below (I, II):

TABLE I. Confusion Table (MFCC/HMM)

Pronounced	Classified as:					
	aaa	***ooh***	***eeu***	***iii***	***clic***	***stop***
aaa	-	-	-	x	x	x
ooh	x	-	-	-	x	x
eeu	x	x	-	x	-	-
iii	-	x	-	-	x	-
clic	x	-	-	x	-	x
stop	-	-	-	-	-	

x: means that pronounced phoneme classified as an other

TABLE II. Confusion Table (LPC/DTW)

Pronounced	Classified as:					
	aaa	***ooh***	***eeu***	***iii***	***clic***	***stop***
aaa	-			-	x	-
ooh	x		x	-	-	-
eeu	x	x	-	x	-	-
iii	x	-	x	-	x	-
clic	x	-	-	x	-	-
stop	-	-	x	-	x	-

x: means that pronounced phoneme classified as an other

TABLE III. Classification of LPCs using DTW

phoneme	Recognition Ratio (%)
aaa	92.86
ooh	58.33
eeu	0
iii	30.77
clic	54.55
stop	55.56

TABLE IV. Classification of MFCCs using HMM

phoneme	Recognition Ratio (%)
aaa	78.57
ooh	75.00
eeu	64.28
iii	84.62
clic	54.55
stop	100

According to the results presented above (Table: III, IV), the recognition rates using MFCCs parameterization classification with DTW or HMM classification is better than: LPCs and MFCCs with DTW. So we can say that the MFCCs / HMM system is partially independent of the speaker.

In addition, we must consider the preprocessing for noise in future work, as well as the database training models need from the category of children.

CONCLUSION

According to the results, we note that the classification using HMM is better than the DTW, and the decision based on MFCC coefficients is more certain than the coefficients LPCs.

From experimental results, it can be concluded that MFCC features and HMM as classifier can recognize the speech signal well. Where the highest recognition rate that can be achieved in the last scenario. This result is achieved by using MFCCs and HMM.

In addition, the variety of signals of database, the recording conditions and the environment, have an impact in classification phase.

REFERENCES

[1] C. de Mauro, M. Gori, M. Maggini, and E. Martinelli, "A voice device with an application-adapted protocol for Microsoft windows," In Proc. IEEE Int. Conf. on Multimedia Comp. and Systems, vol. 2, pp. 1015–1016, Firenze, Italy, 1999.

[2] T. Igarashi and J. F. Hughes, "Voice as sound: Using non-verbal voice input for interactive control," In ACM UIST 2001, November.

[3] Alex Olwal and Steven Feiner, "Interaction techniques using prosodic features of speech and audio localization," In IUI '05: Proc. 10th Int. Conf. on Intelligent User Interfaces. New York: NY, USA, 2005. ACM Press, pp. 284– 286.

[4] M.E. Sargin,O. Aran,A. Karpov,F. Ofli1,Y. Yasinnik,S. Wilson,E. Erzin,Y. Yemez and A.M. Tekalp, ''Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis,'' ICME 2006 : IEEE International Conference on Multimedia and Expo, July 2006, pp: 893–896.

[5] J. Bilmes, X. Li, J. Malkin, K. Kilanski, R. Wright, K. Kirchhoff, A. Subramanya, S. Harada, J. Landay, P. Dowden, and H. Chizeck, "The vocal joystick: A voice-based human-computer interface for individuals with motor impairments," in Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing, Vancouver, October 2005.

[6] S. Harada, J. Landay, J. Malkin, X. Li, J. Bilmes, ''The Vocal Joystick: Evaluation of Voice-based Cursor Control Techniques", *ASSETS'06,* October 2006.

[7] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, ''Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," Journal of Computing, Volume 2, Issue 3, March 2010, pp : 138-143.

[8] Mahdi Shaneh and Azizollah Taheri, "Voice Command Recognition System Based on MFCC and VQ Algorithms," World Academy of Science, Engineering and Technology 57 2009, pp: 534-538.

[9] A Bala, A Kumar, N Birla - Anjali Bala et al., ''Voice command recognition system based on MFCC and DTW International Journal of Engineering Science and Technology, Vol. 2 (12), 2010, pp :7335-7342.

[10] Thiang, S. Wijoyo, "Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile robot", International Conference on Information and Electronics Engineering IPCSIT vol.6, 2011, 179-183.

[11] C. Snani, " conception d'un system dereconnaissance de mots isolés à base de l'approche dynamique plastique en temps réel : Application commmande vocale d'un fauteuil roulant", Mémoire de magister ,Institut d'electronique univ. Badji mokhtar Annaba,2004.

[12] C. HAMDI, M. boughazi and M fezari, "improvement of Arabic digits recognition rate based in the parameters choice", in proceedings of international conf. CISA Annaba, june 2008.

[13] M. Fezari and A. Al-dhioud, "An Approach For: Improving Voice Command processor Based On Better Features and Classifiers Selection," pp. 1–5. The 13th International Arab Conference on Information Technology ACIT'2012 Dec.10-13 ,2012.