Speech Interfacing For Control of PC Windows

Abhinav Sharma[1], Anshu Sharma[2]
[1]Depat of Electronics and Communication Engineering
[2]Dept of PDP
[1, 2] Graphic Era Hill University, Dehradun, Uttarakhand

Abstract- *In this research work Speech interfacing for computer control has been developed in MATLAB platform that enables physically disabled people also to gain access to computer. This speech interfacing enables users to execute common computer commands by spoken utterances and not by using standard input devices such as keyboards and mouse. The approach used, in this work for computer control is direct execution of recognized vocal command rather than the mouse movement control or the text matching between that produced by speech recognition system and the one from the application menu . The framework is based on PC Windows (windows 7), and is carried out in English language. In this work, recognition of speech is carried out by Gaussian Mixture Model (GMM). Mel Frequency Cepstral Coefficient (MFCC) technique is used for transferring input speech into spectral components. Average speech command recognition performance is observed to be 74.38%.*

Keywords- Speech Recognition, Speech Interfacing, Windows, GMM (Gaussian Mixture Model), MFCC(Mel Frequency Cepstral Coefficient)

I. INTRODUCTION

In this modern era of technology, continuous discoveries and inventions have facilitated us to improve our ease, comfort and efficiency in almost all fields of our daily lives. Computers have been at the center of all these inventions and achievements. In the last decade computers have shown a major impact on just about all aspects of our lives. All of its kinds from full size desktop, to Laptops and tablet PCs, to handheld Palm pilots and Blackberries, have facilitated its services for users to any location. So it is essential to provide the access of computers to most of the people including physically disabled people with hands. This can be achieved by interfacing the human speech for the computer control. This research paper will describe another approach of speech interfacing for hands-free computer control by speech recognition techniques.

Speech Recognition is a technology that allows the computer to identify and understand words spoken by a person using a microphone or telephone. The ultimate goal of the technology is to be able to produce a system that can recognize with 100% accuracy all words that are spoken by any person.

Even after years of research in this area, the best speech recognition software applications still cannot recognize speech with 100% accuracy. Some applications are able to recognize over 90% of words when spoken under specific constraints regarding content and previous training to recognize the speaker's speech characteristics.

Speech interfacing to computers has different advantages in real life: speech does not require use of physical devices such as keyboards or pointing devices. Computing devices can become more compact as keyboard and mouse pointing devices take a less prominent role. Individuals with physical challenges may also benefit from the use of speech based applications. Such system does not require the use of vision. Speech can be performed effectively in low light environments or by persons with low or no vision. Applications for this technology may include uses where the individual needs to keep their eyes on equipment or the environment if navigating a vehicle. People have been speaking for thousands of years and individuals start speaking at a young age. Speech systems can process speech from individuals that may be at some distance away from the computer. Wireless and handheld devices are becoming increasingly smaller with corresponding smaller displays and input keys. A voice interface is not constrained by the physical size of the device.

Speech recognition involves the extraction of features from the vocal commands, training of speech models and then testing. There are several features used for recognition of different speech commands from speech such as: (i) excitation source features: using these features, shape of glottal pulse, strength of excitation and characteristics of open and closed phase of glottis are captured. Features extracted from the vibration pattern of vocal folds are known as excitation source features. (ii) prosodic features are used to capture long term features like melody, rythm, intonation through duration, energy and pitch features. (iii) spectral features are used to capture the features from sizes and shapes of the vocal tract, formant information and so on. In this study excitation features are used.

Different pattern classifiers such as: AANN (Auto Associative Neural Network), GMM (Gaussian Mixture Models), HMM (Hidden Markov Model), SVM (Support Vector Machines) are being used for various speech tasks. AANN is used to capture non-linear, higher order relations among the samples. GMM is used to capture the distribution pattern of feature vectors. HMM is used when capturing sequential information is important. SVM is helpful during the classification of highly distinguishable feature vectors those are less in number. So, it is an important decision to choose suitable database, features and classifier while developing vocal command recognition models.

II. LITERATURE REVIEW

Using the combination of voice recognition and speech synthesis, R. Evans, Wayne A. Tjoland and Lloyd G. Allred [2] in 2000 developed a hands-free computer interface which allows the technician to probe and tune electronic circuit cards without having to remove his hands from the circuit card or to focus his eyes on the computer screen. In this system, the operator was allowed to enter data and to control the software flow by voice command or from the keyboard or mouse. Grammar set, or legal set of commands could be specified dynamically. The computer voice helps the operator to give his attention to other activities like probing a circuit card and taking readings. To insure reliable entry, system echoes the entered voice when operator is taking readings. Operator can hear the resultant reading using electronic tuning. This enables operator to focus on the circuit card instead of constantly turning his head to see the computer screen. Operator training was not required. The Voice Control application was devised for use with the Microsoft Speech API version **4.0.** Any 32-bit Windows software, which has window messaging capability, can access this voice control system. Standard programming languages such as Borland or Microsoft C/C+ + and Visual Basic (VB) and commercial packages such as Lotus Notes and Microsoft Word support this system. The system uses a limited set of grammar, or legitimate commands, providing a restricted context, thus enabling very high reliability. To prevent inadvertent errors, verification from the user is required on crucial commands.

M. Abdeen, H. Mohammad, M. C.E. Yagoub in 2008 developed a language-independent framework for a hands-free control of desktop computer. It works for PC windows. It is tested on both English and Arabic languages. The approach used for this framework is based on matching of text between that produced by API functions and the speech recognition techniques. After successful matching both the texts, the system runs the vocal command. Speech recognition phases in general and for Arabic in particular consist of feature extraction, endpoint detection, segmentation, where the spoken word is segmented into its phonemes and finally Hidden Markov Model is used to train and recognize speech utterances from given observations . “Forward” algorithm and “Viterbi” algorithm are used for calculating the probability of observation sequences and state sequences. They also presented a comparison between their work and the XP and Vista. Some difficulties with speech recognition were encountered especially when users speak relatively above average. Reason for this is inability to recognize start and end of each word due to error in finding silences.

R. Maskeliunas, K. Ratkevicius, V. Rudzionis in 2011 proposed a Voice-based human-machine interaction model for automated information services. This model allows recognition of isolated commands together with some keywords. Different foreign language speech engines have different capabilities to recognize Lithuanian voice commands. The main aim of these experiments was to establish the limits of possibilities to improve the recognition accuracy of Lithuanian voice commands selecting more proper foreign language engine. English (Microsoft Speech Recognizer 9.0 for Microsoft Speech Server (English-US)) and Spanish (Microsoft Speech Recognizer 9.0 for Microsoft Speech Server (Spanish-US)) engines were selected for the comparison by using ten vocal commands. Spanish speech engine enabled to achieve significantly higher recognition accuracy than English engine: overall recognition accuracy increased from 77% for the English engine to the 97% for the Spanish engine.

Sandeep Kaur, in 2012 developed Mouse Movement control using Speech and Non-Speech Characteristics of Human Voice. In this research work, she presented a system called as Vocal Mouse (VM) which allows users to continuously control the mouse pointer using words as well as sounds by varying vocal parameters such as vowel quality, loudness and pitch. The difficulty of poor command recognition with the traditional methods, while performing continuous tasks was removed by this system as it allows users to work on both continuous and discrete motion control. Speech recognition of mouse movement commands involved extraction of low-level acoustic features in real time using LPC (Linear Predictive Coding). Pattern recognition is performed using a new proposed technique called “minimum feature distance technique”. This proposed technique is based on calculating distances between the spoken word and each stored word in the library during training process. Features from pattern recognition module are processed to produce output in the form of cursor’s 2-D movement. VM can be used by novice users without extensive training and it presents a viable alternative to existing speech based cursor control methods.

III. FEATURE EXTRACTION:

This analysis technique is very useful as it provides methodology for separating the excitation from the vocal tract shape [2]. In the linear acoustic model of speech production, the composite speech spectrum, consist of excitation signal filtered by a time-varying linear filter representing the vocal tract shape as shown in fig.1.

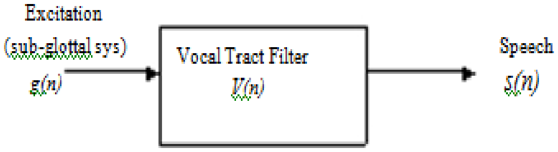

Fig. 1 Model of Vocal Tract filter

The speech signal is given as

$$s(n) = g(n) * v(n)$$

where *v(n):* vocal tract impulse response
g(n): excitation signal

The frequency domain representation

$$S(f) = G(f) . V(f)$$

Taking log on both sides

$$\log(S(f)) = \log(G(f)) + \log(V(f))$$

Hence in log domain the excitation and the vocal tract shape are superimposed, and can be separated. Cepstrum is computed by taking inverse discrete Fourier transform (IDFT) of logarithm of magnitude of discrete Fourier transform finite length input signal.

Details of calculating the features based on MFCCs

Delta and Acceleration Coefficients:

The first order regression coefficients (delta coefficients) are computed by the following regression equation:

$$d_i = \frac{\sum_{n=1}^{N} n(c_{n+i} - c_{n-i})}{2\sum_{n=1}^{N} n^2}$$

where di is the delta coefficient at frame *i* computed in terms of the corresponding basic coeffecients *cn+1 to cn-i.*

The same equation is used to compute the acceleration coefficients by replacing the basic coefficients with the delta
coefficients.

Cepstral Mean Normalization:

This option is aimed at reducing the effect of multiplicative noise on the feature vectors. Mathematically it is:

$$c_i = c_i - \frac{1}{N}\sum_{k=1}^{N} c_{ik}$$

where *ci* is the *i*th feature element in the feature vector and *c*ik is the *i*th feature element at frame *k*. *N* is the number of total input frames of data.

Energy and Energy Normalization:

Energy is calculated as the log signal energy using the following equation:

$$E = \log \sum_{n=1}^{N} s_n^{\,2}$$

where *sn* is the *n*th input speech data sample and *N* is the total number of input samples per frame. The corresponding normalization is to subtract the noise floor from the input data. Note that the silence floor is usually set to 50 dB and the minimum energy value is -1.0e+10.

Liftering:

Liftering is applied according to the following equation.

$$c_n' = \left(1+\frac{N}{2}\sin\frac{\pi n}{N}\right)c_n$$

Pre-emphasis:

The first order difference equation:

$$s_n' = s_n - \alpha s_{n-1}$$

is applied a window of input samples. Here *a* is the pre-emphasis filter coefficient in the range [0,1].

IV. FEATURE CLASSIFICATION AND MATCHING

Fundamentals of Gaussian Mixture Model

In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data-set should identify the sub-population to which an individual observation belongs.

Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. However, while problems associated with "mixture distributions" relate to deriving the properties of the overall population from those of the sub-populations, "mixture models" are used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population, without sub-population-identity information.

Some ways of implementing mixture models involve steps that attribute postulated sub-population-identities to individual observations (or weights towards such sub-populations), in which case these can be regarded as types of unsupervised learning or clustering procedures. However not all inference procedures involve such steps.

Mixture models should not be confused with models for compositional data, i.e., data whose components are constrained to sum to a constant value (1, 100%, etc.).

GMM FOR SPEECH RECOGNITION

GMM are used specifically for capturing distribution of data points from the input features.

Given a set of inputs, GMM refines the weights of each distribution through expectation-maximization algorithm.Once a model is generated, conditional probabilities can be computed for test patterns (unknown data points). Gausses in the mixture model is known as number of components. They indicate the number of clusters in which data points are to be classified. The components within each GMM capture finer level details among the feature vectors of each speech command. In this work, GMM’s are designed with 64 components and iterated for 100 times to attain convergence of weights.

III. APPROACH

The approach used for speech based computer control is based on the recognition of vocal computer command in real time, involving feature extraction and testing of all trained GMM classifiers of different commands with the features extracted from the test command. The total log- likelihood of test vectors of one test utterance with respect to the trained GMMs corresponding to each command is computed. The test utterance is considered to belong to that command with respect to which the total log-likelihood becomes the largest.

Using the MATLAB-Windows interface, the recognized command is executed in the Microsoft Windows platform.

IV. DATABASE COLLECTIONS

A headphone-mic, a Laptop computer, the Wavesurfer software were used for single channel recording of ten selected system commands for which the computer is to be interfaced, in a closed-room noise-free environment (samples were called as clean samples). 20 speech samples of each command were taken for enhancing the probability of recognition. Secondly, 4 artificial noises, namely crowd talking, fan- sound, traffic-sound and White Gaussian Noise were added to all 200 clean samples for reliability of the system. For digitization 48000 Hz of sampling frequency and 16-bit quantization were used.

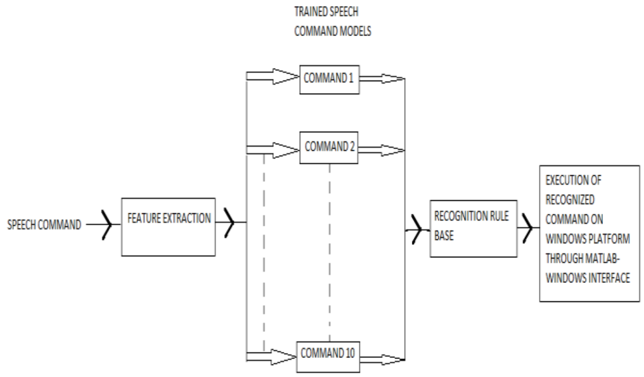

Fig.2 System Proposed

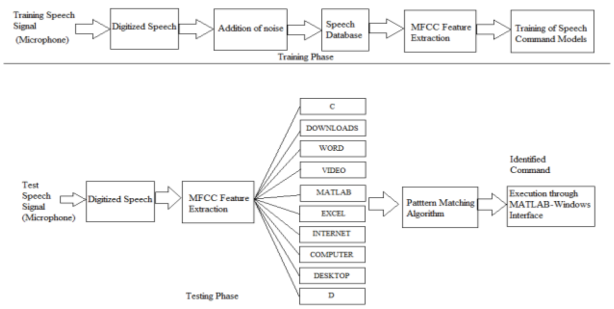

Fig.3 Flow Diagram

V. EXPERIMENTAL RESULTS IN DYNAMIC ENVIRONMENT

Table-1 Description of selected commands

No.	Commands	Description
1	C	Open C drive
2	COMPUTER	Open My Computer
3	D	Open D drive
4	DESKTOP	Open Desktop
5	DOWNLOADS	Open My Downloads
6	E	Open E drive
7	E-MAILS	Open e-mails
8	EXCEL	Open Microsoft Office Excel 2007 program
9	INTERNET	Open Default internet browser window
10	MATLAB	Open MATLAB
11	SONG	Play pre-selected song
12	VIDEO	Play pre-selected video
13	WORD	Open Microsoft Office Word 2007 program

Table-2 Command Recognition Performance

Command	Command recognition performance (%)												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	100	0	0	0	0	0	0	0	0	0	0	0	0
2	0	100	0	0	0	0	0	0	0	0	0	0	0
3	47	0	47	7	0	0	0	0	0	0	0	0	0
4	0	0	0	100	0	0	0	0	0	0	0	0	0
5	0	0	0	0	100	0	0	0	0	0	0	0	0
6	0	0	0	0	0	100	0	0	0	0	0	0	0
7	80	0	0	0	0	20	0	0	0	0	0	0	0
8	0	0	0	40	0	0	0	60	0	0	0	0	0
9	0	0	0	20	0	0	0	0	80	0	0	0	0
10	0	0	0	0	0	0	0	13	0	87	0	0	0
11	0	0	0	13	74	0	0	0	0	0	13	0	0
12	0	20	0	0	0	0	0	0	0	0	0	80	0
13	0	0	0	0	0	0	0	0	0	0	0	0	100

Table-3 Recognition Robustness Performance

Speech Commands	Closed Test Utterances	Open Test Utterances
Command 1	100	100
Command 2	100	100
Command 3	100	47
Command 4	100	100
Command 5	100	100
Command 6	100	100
Command 7	100	0
Command 8	100	60
Command 9	100	80
Command 10	100	87
Command 11	100	13
Command 12	100	80
Command 13	100	100

Table-4 Recognition Performance with different numbers of MFCCs

Speech Commands	No. of MFCCs				
	6	8	13	21	29
	Recognition Performance (in %)				
Command 1	0	0	80	100	20
Command 2	60	20	100	100	40
Command 3	0	0	20	47	0
Command 4	80	100	93	100	67
Command 5	100	100	100	100	100
Command 6	100	100	100	100	100
Command 7	80	20	0	0	13
Command 8	13	7	100	60	47
Command 9	87	80	87	80	80
Command 10	47	60	60	87	67
Command 11	0	33	0	13	0
Command 12	67	93	87	80	93
Command 13	27	80	100	100	100
Average	50.84	53.30	71.30	74.38	55.92

VI. CONCLUSION AND FUTURE WORK

The paper covers the importance of speech recognition technology to fasten the process of communication. In addition the research achieves and meets the objectives of developing it, and it is hoped that the research will benefit the end users as it is designated for that purpose.

The speech recognition technology has been widely used in this system. As a result, this study manages to show and emphasize the need and importance of such technology in our daily life. This system could be implemented in various business, organizations, financial institutions and many other academic institutions.

Computers are the utmost needs of today's s world. But it is not possible for physically challenged people with hand (or amputees) to operate on a computer machine. Moreover passwords protection of computer machines in today's world is not safe as passwords are easily hacked by the hackers, which makes the computer machines and different application systems controlled by it easily damageable. The above problems can be solved if the computer machine, operated by a physically handicapped person can be controlled by the speech commands of its user. The system will contribute a lot for the disabled people to make them able to get connected to the fast changing technology. This also contributes to make personal computers very much user friendly as the user is able to operate computer with the utterance of commands only. Biometric identification with speech is a good tool for prevention of mis-use of computer machine.

REFERENCES

[1] J. W. Picone, "Signal modelling technique in speech recognition," *Proc. Of the IEEE,* vol. 81, no.9, pp. 1215-1247, Sep. 1993.

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

[3] D.O. Shaughnessy, *Speech Communication: Human and Machine.* India:University Press ,2001.

[4] B. Gold and L. R. Rabiner,"Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. America,* vol.46, pt. 2, no. 2, pp 442-448, Aug. 1969.

[5] Dr.E.Chandra, A.Akila ,"An Overview of Speech Recognition and Speech Synthesis Algorithms", Int.J.Computer Technology & Applications,Vol 3 (4), 1426-1430 ISSN:2229-6093

[6] Ahmad A. M. Abushariah, Reddy Surya Gunawan, " Speech Recognition System using MATLAB ", LAB LAMBERT Aceademic Publishing GmbH & Co. KG, 2011

[7] Ahmad A. M. Abushariah, Teddy S. Gunawan,Othman O. Khalifa, "English Digits Speech Recognition System Based on Hidden Markov Models", International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia

[8] Susumu Harada, "Harnessing the Capacity of the Human Voice for Fluidly Controlling Computer Interface", University of Washington, 2010

[9] James R Evans, Wayne A Tjoland and Lloyd G Allred, "Achieving a Hands-Free Computer Interface using Voice Recognition and Speech Synthesis ",IEEE AES Systems Magazine, 2000

[10] Susumu Harada and James A Landay, "The Vocal Joystick: Evaluation of Voice-based Cursor Control Techniques", Portland, Oregon, USA: ASSETS, 2006

[11] M Abdeen, H Moshammad and M C E Yagoub, "An Architecture for Multi-Lingual Hands Free Desktop Control System for PC Windows", Niagara Falls, Canada : IEEE , 2008

[12] R Maskeliunas, K Ratkevicius and V Rudzionis, "Voice-based Human-Machine Interaction Modeling for Automated Information Services", ISSN 1392-1215 Electronics and Electrical Engineering, 2011

[13] R Norma Conn and Michael McTear, "Speech Technology: A Solution for People with Disabilities", Savoy Place, London WCPR OBL, UK: IEE, 2000

[14] Shashidhar G. Koolagudi, K.Srinivas Rao, "Recognition of Emotions from Speech using Excitation Source Features", International Journal of Speech Technology, June 2012, Volume 15, Issue 2, pp 265-289

[15] Sandeep Kaur , "Mouse Movement using Speech and Non-Speech Characteristics of Human Voice", International Journal of Engineering and Advanced Technology (IJEAT) ,ISSN: 2249 – 8958, Volume-1, Issue-5, June 2012