Embedded Design of a Remote Voice Control and Security System

B. D. Subudhi, A.K. Patra, N. Bhattacharya, and P. Kuanar
Department of Electrical Engineering, National Institute of Technology, Rourkela-769008, India
bidyadharnitrkl@gmail.com

Abstract — **It has been observed that modern heavy industries are plagued with hazardous working environments like high temperature, high voltage etc. that pose a severe threat to the life of a human operator. Most common examples are steel plants and mines which report huge losses of life every year. Thus, there is a need to develop a system which any human personnel can use to control a process without exposing himself to any type of prevailing work hazards. This paper describes a Remote Voice Control and Security System (RVCS) to solve the above problem. It is a system that can control any process (e.g. switching ON/OFF a dc motor for experimental purposes) from anywhere in the world by vocalizing appropriate commands into a mobile phone. The hardware of the system comprises of a GSM handset connected to a TI TMS C6713 DSP Starter Kit via a stereo audio cable. An authorized operator at a remote location uses his GSM handset to connect to this system located at the process site. The system prompts the operator for access authentication (security feature) and once authenticated the user controls the plant operation by voicing the appropriate predefined commands (e.g. starting and stopping a motor). The DSK kit carries out real time speaker recognition for access authentication and speech verification [2-3] for decoding the vocal commands. The software of the system has been developed in ANSIC in Code Composer Studio Platform.**

***Index Terms*—Mel Frequency Cepstrum Coefficients (MFCC), Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), Speaker Verification, Speech Recognition, Code Composer Studio, Euclidean Distance, Vector Quantization.**

I. INTRODUCTION

The dawn of twenty-first century has witnessed an unprecedented penetration of mobile technology in day to day life. The use of mobile technology has so far been limited to the field of communication. Simultaneously we have also seen an increased loss and injury to human operators in heavy industries (like mines, steel, heavy electrical etc) due to exposure to hazardous working conditions. We propose to solve the above problem by combining mobile technology and embedded DSP to build RVCS, a Remote Voice Control and Security System. It is a wireless system for remote voice control of plants and processes from any corner in the world by vocalizing appropriate commands into his mobile phone. The system is armed with real time speaker authentication facility for enhanced security and real time speech recognition facility for decoding the vocal commands. Some of the related work in this area has been reported in [1-6]. Motivated by the above works, we have built an improved system that can integrate both security and control facilities into a single entity. This is especially useful in the case of large industries with distributed operations.

The project is a fusion of two modern technologies i.e. mobile telephony and Embedded Digital Signal Processing [4]. For our purpose we needed a cheap, readily available and reliable communication device so that vocal commands can be transported to the RVCS system without much hassle. GSM handset or mobile phone best fitted for this job. Secondly we needed an Embedded DSP Platform to build a standalone Real Time System for online Speaker Verification and Speech Recognition [4].TMSC6713 TI DSP Starter Kit platform was chosen as it is best suited for the above purpose [7]. It is available in the Real Time and Embedded Systems Laboratory, Department of Electrical Engineering, National Institute of Technology (NIT), Rourkela.

The RVCS project has accomplished the following objectives. The first one is to implement speaker recognition (for access authentication to enhance security) and speech verification (to decode the various vocal commands) in TI DSP TMS 320C6713 kit followed by interfacing a cell phone into the DSP Kit. The final one is full system integration of all the individual hardware and software modules. Provisions can be made for added facilities like voice driven menu system for ease of operation. RVCS system can be very useful in the following areas

- Heavy Industries where human operators work in hazardous working conditions. The RVCS system removes the necessity of physical presence of the operator at the process site thus protecting him from any peril to life.
- Automatic Robot or Unmanned Aerial Vehicle (UAV) navigation in areas like hostile mines, enemy territory (for defense purposes) etc.
- Home Automation: Control of consumer electronic appliances in our home like switching on/off washing machine, lights, TV etc.

The speaker and speech recognition algorithms used in the system have certain novelties

- The vocabulary database is self made i.e. the database is constructed during training by the authorized users. This makes it less complicated compared to standard bulky database like MIT Digit, LANIC etc. The implementation becomes hassle free in TI DSK due to its simplicity
- As the database is self made, **language independent recognition [2]** can be implemented i.e. the system can recognize commands in any language if it is trained.

II. DESIGN ARCHITECTURE OF RVCS SYSTEM

A. Hardware Architecture

Fig 1 given below shows a schematic diagram of the RVCS system.

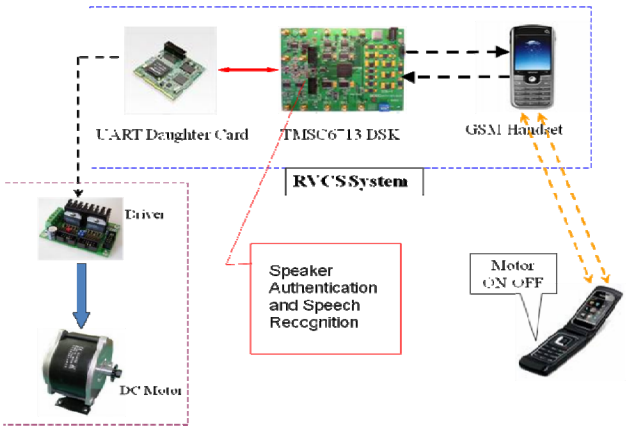

Fig 1.Block Diagram of RVCS

The hardware of RVCS comprises of a TI DSP (TMS320C6713) starter kit, GSM handset, 12v DC motor (process to be controlled), DC motor driver and UART daughter card(or General Purpose Input/output) for giving governing signals to the driver).

The TI DSK is responsible for implementing real time text dependent speaker authentication and language independent speech verification algorithms. It is in fact the brain of the system as it performs all the acquisition, decoding and control operations.

The system GSM handset is connected to the DSK via a standard stereo audio cable i.e. the audio output port of the handset is connected to the audio input (mic in /line in) of the DSK kit so that the audio signal received by the handset is directly transmitted to the kit in line (without any speaker) The handset is continuously monitored for incoming calls in auto-answer mode. An authorized operator uses his personal GSM handset to connect to the handset located at the process site. On successful connection the operator is prompted for voice authentication. This allows access to authorized personnel only for whose voice the system has already been trained, thereby enhancing the security of the system. No unauthorized user (for whom the system has not been trained) can access the system.

Once the authentication is achieved the operator takes necessary action by voicing the appropriate commands into his/her handset. The system then processes the voice commands and executes them. This process continues till the operator is satisfied with the results.

B. Interfacing DSK with microphone

The microphone used here is an Electret microphone as it is sensitive, very durable, and extremely compact in size and has low power requirements. It works well for PC application like sound cards which includes a preamp inside it. But to interface with the DSK kit it needs an amplifier. The diagram below shows the pin diagram of a 3.5 mm jack.

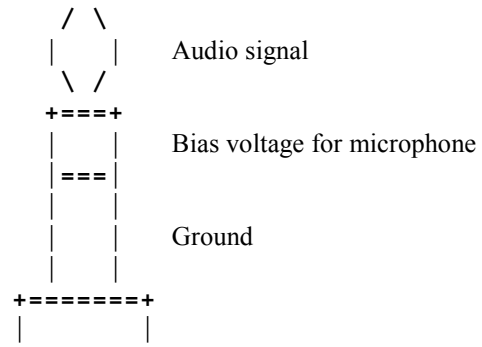

Fig 2.Schematic Diagram of Electret microphone

Input Type: Unbalanced Low Impedance
Input Sensitivity: Approx. -20dBV (100mV or 0.1Volt)
Input Impedance: 600 to 1500. (Ohms)
Input Connector: 3.5mm Mini plug (Stereo Jack)
Input Wiring: Audio on Tip, Ground on Sleeve, 5Volts DC Bias on Ring.

This circuit is suitable for interfacing two wire electret microphone capsules to soundcards (Sound Blaster soundcards) which supply bias voltage for powering electret microphones.

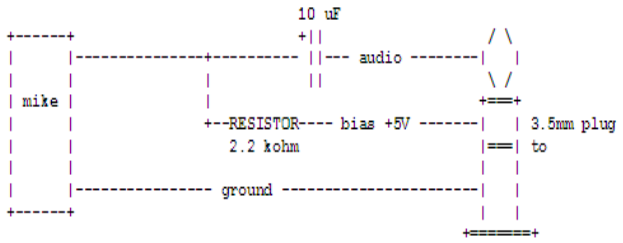

Fig 3.Experimental Picture of Electret microphone

C. DC Motor control with DSK Kit

The TMS320C6713 is actually meant for signal processing not for motor control. So it has no direct general purpose input and output (GPIO) pins on its own. But this project requires to do both signal processing as well as motor control in 6713 DSK. So some tampering is done with DSK kit. This can be avoided if a UART daughter card is available.

The 6713 DSK kit is provided with Host peripheral Interface (HPI) which is multiplexed with GPIO pins. The details of GPIO are given below.

D. General-purpose input/output (GPIO) in HPI

To use the 0-15 software-configurable GPIO pins, the GPxEN bits in the GP Enable (GPEN) Register and the GPxDIR bits in the GP Direction (GPDIR) Register must be properly configured.

GPxEN = 1 GP[x] pin is enabled
GPxDIR = 0/1 GP[x] pin is an input/output

where “x” represents one of the GPIO pins.

31	24	23	16
Reserved			
R-0			

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
GP15 EN	GP14 EN	GP13 EN	GP12 EN	GP11 EN	GP10 EN	GP9 EN	GP8 EN	GP7 EN	GP6 EN	GP5 EN	GP4 EN	GP3 EN	GP2 EN	GP1 EN	GP0 EN
R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-1	R/W-1	R/W-1	R/W-1	R/W-0	R/W-0	R/W-0	R/W-0

Legend: R/W = Readable/Writeable; -*n* = value after reset, -x = undefined value after reset

Table 1.GPEN Register Configuration

Table 1 shows the GPIO enable bits in the GPEN register for the C6713/13B device. To use any of the GPx pins as general-purpose input/output functions, the corresponding GPxEN bit must be set to “1” (enabled).

31	24	23	16
Reserved			
R-0			

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
GP15 DIR	GP14 DIR	GP13 DIR	GP12 DIR	GP11 DIR	GP10 DIR	GP9 DIR	GP8 DIR	GP7 DIR	GP6 DIR	GP5 DIR	GP4 DIR	GP3 DIR	GP2 DIR	GP1 DIR	GP0 DIR
R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0	R/W-0

Legend: R/W = Readable/Writeable; -*n* = value after reset, -x = undefined value after reset

Table 2.GPIO Register Configuration

Table 2 shows the GPIO direction bits in the GPDIR register. This register determines if a given GPIO pin is an input or an output providing the corresponding GPxEN bit is enabled (set to “1”) in the GPEN register. By default, all the GPIO pins are configured as input pins. Any of the available grounds can be connected with the motor while GPIO(2) is used as the output pin.

E. Experimental Setup

This project has been implemented in hardware and completed successfully in Real Time Embedded Systems Laboratory, NIT Rourkela. A TMS320C6713 kit, a dc motor, a microphone and 2 mobile phones are the integral components of the setup. For better understanding and clarification we have provided a few snapshots of our experimental setup.

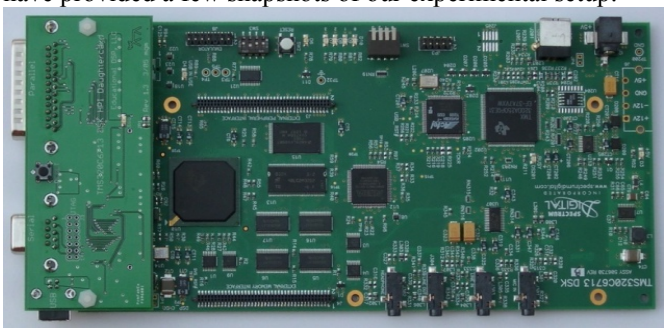

Fig 4.TMS320C6713 DSP Starter Kit, Vendor:-Spectrum Digital Inc.

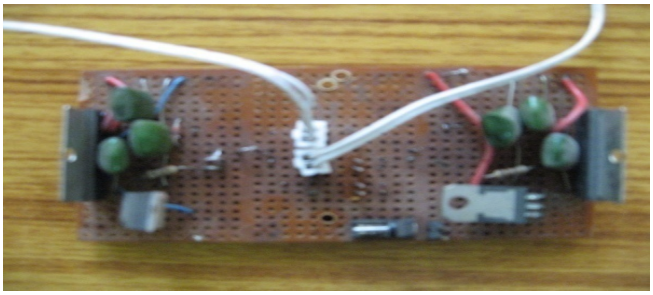

Fig 5.BA6209 DC Motor Driver Circuit
I/P: - 5V, O/P:-12V, Vendor:-Self Made

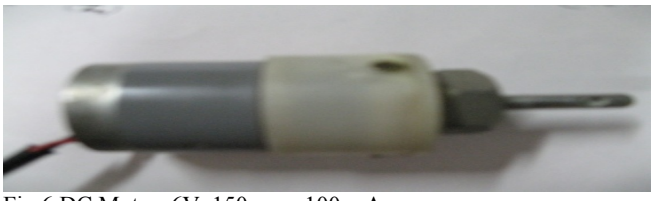

Fig 6.DC Motor, 6V, 150 rpm, 100 mA
Vendor:-Gala Electronics Lamington Road, Mumbai

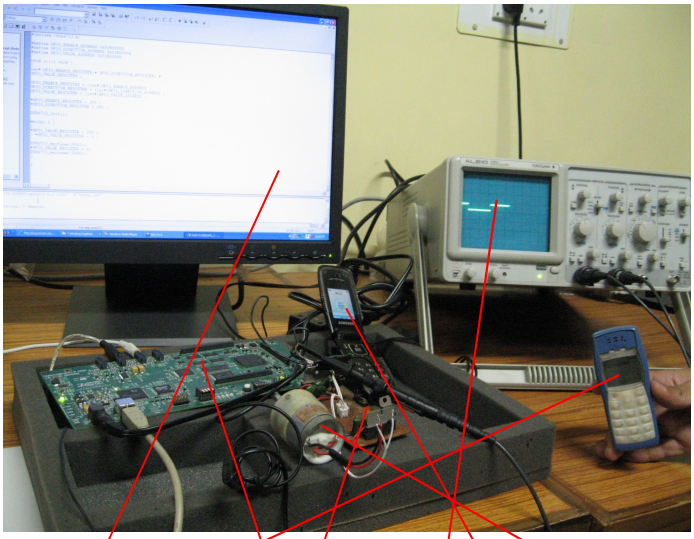

Fig 7.Complete Experimental Setup of RVCS System in Real time Embedded Systems, NIT Rourkela

F. Software Interface

This software for the system has been written in ANSIC in Code Composer Studio Platform provided by Texas Instrument.

G. Code Composer Studio

The CCStudio IDE extends the basic code generation tools with a set of debugging and real-time analysis capabilities. The CCStudio IDE supports all phases of the development cycle shown here:

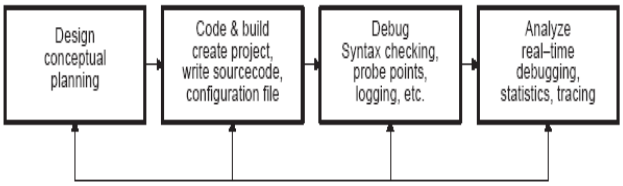

Fig 8.Code Composer Studio Platform

CCStudio Setup allows configuring the CCStudio IDE software to work with different hardware or simulator targets. In this case hardware interface is TMS320C6713 as mentioned above. Code composer studio has its inbuilt libraries for every DSP Kit available at Texas Instruments.

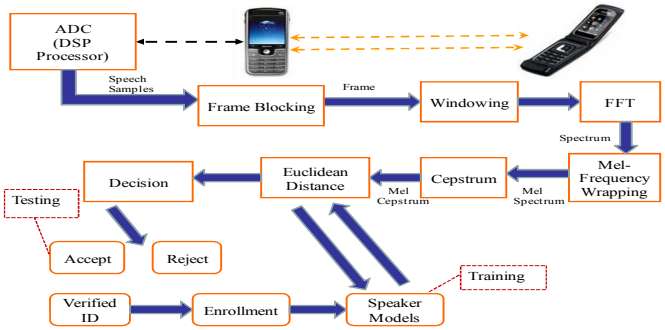

Fig 9.Process Flow of Automatic Speaker Verification

H. Algorithm for Automatic Speaker Verification

The following algorithm is to be followed for the implementation of automatic speaker verification [3-9]:

- Trained vectors are derived from the speech sample of the speaker at a different time.
- First the input analog speech signal is digitized at 8KHZ Sampling Frequency using the on board ADC (Analog to Digital Converter)
- The Speech sample is stored in a one-dimensional array. Speech signals are quasi-stationary. It means that the speech signal over a very short time frame can be considered to be stationary.
- The speech signal is split into frames. Each frame consists of 256 Samples of Speech signal and the subsequent frame will start from the 100th sample of the previous frame. Thus each frame will overlap with two other subsequent frames. This technique is called Framing. Speech sample in one frame is considered to be stationary.
- After Framing, to prevent the spectral leakage we apply windowing. Here Hamming Window [1-10] with 256 coefficients is used.
- The next step is to convert the time domain speech signal into frequency domain using Discrete Fourier Transform. Here Fast Fourier Transform is used [1-10].
- The resultant transformation will result in a signal being complex in nature. Speech is a real signal but its Fourier Transform will be a complex one (Signal having both real and imaginary).
- The power of the signal in Frequency domain is calculated by summing the square of Real and Imaginary part of the signal in Frequency Domain. The power signal will be a real one. Since second half of the samples in the frame will be symmetric to the first half (because the speech signal is a real one) we ignore the second half (second 128 samples in each frame).
- Triangular filters are designed using Mel Frequency Scale. These banks of filters will approximate our ears. The power signal is then applied to these banks of filters to determine the frequency content across each filter.
- In our implementation we choose total number of filters to be 20. These 20 filters are uniformly spaced in Mel Frequency scale between 0-4 kHz.
- After computing the Mel-Frequency Spectrum, log of Mel-Frequency Spectrum is computed. Discrete Cosine Transform of the resulting signal will result in the computation of the Mel-Frequency Cepstral Coefficient.
- Euclidean distance between the trained vectors and the Mel-Frequency Cepstral Coefficients are computed for each trained vectors. The trained vector that produces the smallest Euclidean distance will be identified as the speaker.

I. Algorithm for speech recognition

Since this project deals with isolated word and small vocabulary following algorithm is implemented.

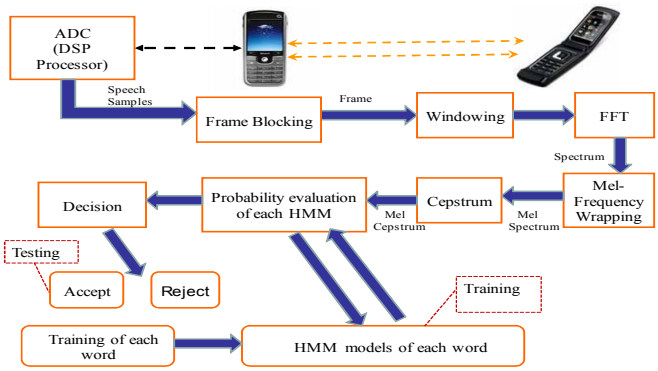

Fig 10.Process Flow of Automatic Speech Recognition

For Speech recognition [2-3] the following steps are implemented subsequent to the execution of the above steps mentioned in speaker recognition.

- Delta Values of MFCC Coefficients are calculated.
- Flat Start Approach is implemented to construct initial model of HMM (Hidden Markov Models).All the speech frames are divided into equal segments and they are assigned a state.
- The initial HMM Model has an architecture where each state can lead to its subsequent or the same state only. Thus the diagonal elements of the state transition matrix (A matrix) are assumed to be 0.95 and their subsequent elements in the same row are assumed to be 0.05. Rest elements are all zero.

- Similarly pi values (initial sate probability) of the HMM are assumed to be 1 for state 1 and rest are assumed to be 0.
- $\mathbf{b_j}$ of each state is calculated where $\mathbf{b_j}$ is the probability of symbol in that state.
- The HMM model is trained using Viterbi search algorithm.
- $\mathbf{a_{ij}}$ and pi values are re estimated up to convergence.

Repeat the above steps for the training of each word to obtain their respective models.

The above algorithm is implemented during both training and verification process. For verification the MFCC coefficients of the word (observation vector) spoken by the user are calculated. Now using Viterbi search algorithm the probability of the observation vector corresponding to the HMM model of each word is calculated.
The word model giving the highest probability is the recognized word.

IV. RESULTS AND DISCUSSIONS

B. Speaker Recognition:

The speaker recognition for this project was implemented using two algorithms. One program uses the Euclidean distance method and the other uses VQ method. The program which uses the Euclidean Distance method has lower accuracy than the one using the VQ method [7-8]. The accuracy for speaker recognition using Euclidean Distance method is approximately 60 percent which increases to about 80 percent once we use the VQ algorithm for implementation of the same. The following snapshot shows the result obtained for speaker recognition using the VQ algorithm. Better accuracy in the VQ algorithm is achieved at the expense of larger computation time and a more complex program. The Euclidean Distance method is simpler to implement but due to its lower accuracy is hardly ever used in real world applications. By the implementation of speaker recognition using both the algorithms we are thus able to bring out the inherent differences in the two algorithms.

B. Speech Recognition:

Speech Recognition Training:
The speech recognition algorithm employs Hidden Markov Model for modeling a speech sequence. For switching on the motor the voice command is the word "Start" and for switching it off it is "Stop". The model of both the words were trained and convergence for the probability of observation matrix (prob_obs_symb matrix) reached after 4 epochs while convergence for the state transition matrix (A matrix) reached after 6 epochs with satisfactory differential criteria of .02 within the respective values. At the end of the training the state transition matrix was averaged for each training sequence and a new matrix was obtained. This proved a better result while verification. The other alternative was to take the matrix with the minimum or maximum probability but they were at a disadvantage while verification process. Similarly the final probability of training was also averaged for better result. Training for the word Stop was carried out in similar manner. Convergence for the probability of observation matrix (prob_obs_symb) reached after 5 epochs while convergence for the state transition matrix (A matrix) reached after 6 epochs with satisfactory differential criteria of .02 within the respective values. The state transition matrix (A matrix) and probability values (prob_obs_symb matrix) were averaged. Fig.11 and Fig.12 show the waveforms of utterance during the start and stop respectively.

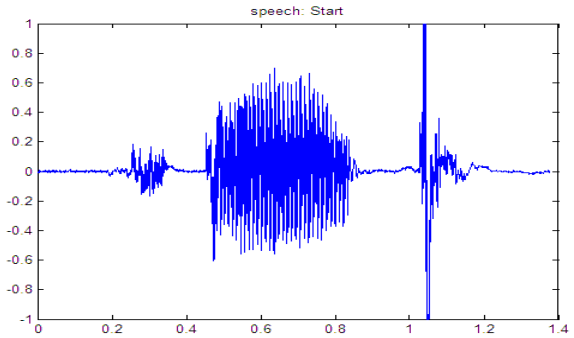

Fig. 11 Waveform of utterance "start"

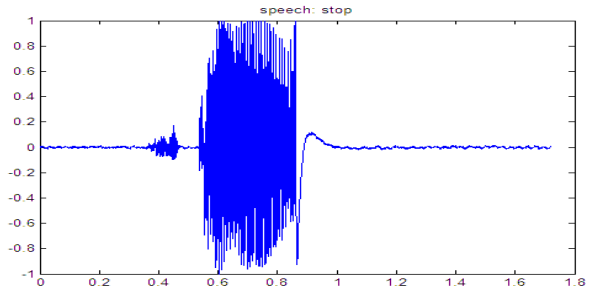

Fig 12.Waveform of utterance "stop"

C. Verification of an Unknown Speech Sequence

The robustness of the algorithm is tested by the verification accuracy of an unknown speech signal. For the purpose of verification viterbi algorithm is used to determine the best state sequence and the probability is calculated for the model of each of the word "start" and "stop". Option was there to calculate probability for all the state sequences but it was discouraged for additional complexity in calculation although accuracy was increased. After evaluation of probability the model that showed the highest probability was assumed to be the verified word. The accuracy was 85% but chances were there that an erroneous verification could occur which is highly dangerous. In order to prevent this Euclidean Distance between the observed probability of the unknown word and the trained probability of the model was calculated for each word. The one with the least difference was provisionally taken as the verified word. A difference criterion or threshold value was empirically determined between 0 and 25. A word is confirmed if the difference meets the above threshold else the unknown signal is discarded as a corrupt sequence and no confirmed verification is shown.

This algorithm reduces the accuracy of correct verification to 75% but drastically reduces the possibility of erroneous detection.

V. CONCLUSIONS

The system so designed has an accuracy of around 60 % for speaker recognition using Euclidean distance algorithm and around 80% using VQ algorithm. Accuracy is compromised if conditions like duration of silence, ambient noise content, duress, emotional and physical health of the speaker vary during training and testing period. Thus we have to ensure that these conditions remain same during both the training and testing phases. The accuracy of speaker recognition could be improved by using a larger database of samples for training purposes. These samples may be taken under varying conditions and thus can present a complete representation of the trained speaker during training.

The speech recognition part has an accuracy of 75% of recognition. This means that 3 out of 4 sequences are detected correctly, but the chance of erroneous detection is 10% which is very satisfactory i.e. if a signal is corrupt it would be verified as a corrupt signal rather than erroneous detection to any of the word. Thus there is a tradeoff between accuracy of righteous recognition and accuracy of erroneous recognition. One has to be sacrificed for the improvement of the other. Another method which we have developed by observation for detection purposes is called “Fall in Distance” method. In this algorithm a signal is determined to be corrupt if both the evaluated probabilities fall within the model probability space of both the words i.e. if the evaluated probabilities for both the model lies between -105(model probability of start) and -107(model probability of stop) it is taken to be a corrupt signal. This method is expected to show a near zero chance of an erroneous recognition.

Finally we have achieved the following deliverables:- Complete control of simple 12v DC motor through the DSK kit, Speech verification and Speaker Verification in the DSK and using it to control the motor. Mobile phone was also interfaced to the DSK kit but showed some poor results for crude model like Nokia 1100.Better results were obtained for new age phones like Nokia N95.We have used a word dependent model for speaker recognition. Ideally one would like to implement word independent models which are more useful from the security point of view. Word dependent models have the drawback that they can be fooled by the use of recording systems. Moreover we have used a word based acoustic model. This model can be used only for limited vocabulary. We would have to move towards a phone based acoustic model. The problem of lack of good public domain acoustic model for India language needs to be addressed. There are a few public domain speech recognition systems that are available. But they have their own drawbacks. For example, there is speech recognition software called ocvolume [5]. The problem with ocvolume is that it does not use HMM, but vector quantization. It will not scale up to many words. Thus as future works it would be useful if someone could do some work in improving these libraries and software so that they can be integrated together and thus become more useful. Further enhancements can be incorporated into the RVCS system like in case of any unexpected event in the plant the system can be trained to call and inform the authorized personnel. Thus the reliability and dependability of the system improves further.

The proposed system has been implemented in the Real Time and Embedded Systems Laboratory, Department of Electrical Engineering, NIT Rourkela. Since it implements a stochastic approach to speaker recognition the accuracy is around 80%.This will definitely improve upon application of more rigorous and robust algorithms. The speech recognition algorithms have also been implemented in the DSP kit. The most effective algorithm is found to be Hidden Markov Models for training and verification purposes.

This project can be very helpful for industries where immense danger is present in controlling any system manually. With the help of the RVCS system the danger can be avoided by controlling the device from a remote location that too with voice which is hassle free and also enhances its security features.

VI. REFERENCES

[1] John G. Proakis and Dimitris K Manolakis, Digital Signal Processing, Prentice Hall India, New Delhi, 2003.
[2] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, New Jersey, 1993.
[3] Lawrence R. Rabiner and Ronald W.Schafer , Digital Processing of Speech signals, Tata McGraw Hill, New Delhi, 2002.
[4] Raulph Chassaing, Digital Signal Processing and Applications with C6713 and C6416 DSK Wiley Inter Science Pub, London, 2004.
[5] Heungsik Chun and I Kim, Realization of Speech Processing, IEEE, 2001
[6] Erdal Bekiroglu and Nihat Daldal, Remote control of an ultrasonic motor by a GSM mobile phone, Science Direct**,** 2004.
[7] T. Kambe, H. Matsuno and A.Yamada, C-Based Design of a real Time Speech Recognition System, ICSAS Conference, 2006.
[8] Y. Linde, A. Buzo and R. Gray, An algorithm for vector quantizer design, IEEE Transactions on Communications, Vol. 28, page no:84-95, 1980.
[9] Lawrence R. Rabiner and Ronald W. Schafer, Speaker Recognition in TMSC6713, Tata McGraw Hill, New Delhi, 2004.
[10] John G. Proakis and Dimitris K Manolakis, Computer based Digital Signal Processing in MATLAB, Prentice Hall India, New Delhi, 2003.