Intelligent Robotic Arm for the Deaf and Dumb

Mr. Sanap Vikram A., Mr. Hullule Smadhan M., Mr. Balsaraf Gokul A., Ms. Bhandari N. K.
PREC, Loni (BE Electronics)

Abstract - This paper presents a speech recognition based 'Intelligent Robotic Arm for Deaf and Dumb' where speech is converted to text and passes the command as isolated word s to hard ware which represents the signs and also to the software which animates the signs. Sign language is used by the deaf and dumb people. It is a combination of shapes and movements of the different parts of the human body. These parts are face and hands. The area of performance of the movements may be from well above the head to the belt level. Signs are used in a sign language to communicate words and sentences to deaf and dumb. The sign language chosen for this project is the American Sign Language. It is the most well documented and most widely used language in the world.

1. INTRODUCTION

Speech is one of the natural forms of communication. Recent development has made it possible to use this in the security system. In speech identification, the task is to use a speech sample to select the identity of the person that produced the Speech from among a population of speeches. In speech verification, the task is to use a speech sample to test whether a speech has in fact done so [1]. In this paper for our project we are going to use MFCC (Mel-Frequency Cepstrum Coefficients) for Speech recognition

2. PRINCIPLES OF VOICE RECOGNITION

Speech recognition methods are divided into text-dependent and text independent methods. In a text Independent system, speech models capture Characteristics of somebody's voice which show up irrespective of what one is saying [1]. And
In a text-dependent system, the recognition of the speech's identity is based on person's voice one or more specific phrases, like passwords, card numbers, PIN codes etc.

Every technology of voice or speech recognition, identification and verification has some sort of advantages and disadvantages and sometimes require different operation and techniques. To choose the technology to use is application-specific. All speech recognition systems contain two main modules these are:
(1) Feature extraction
(2) Feature matching [2, 3].

3. FEATURE EXTRACTION

The intention of this module i.e. is 0f Feature extraction is to convert the speaker waveform to some type of parametric representation (at somewhat lower information rate). The speaker or voice signal is a slowly time varying signal (called a quasi-stationary). When examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are most constant. However, over long slots of time (of the order 0.2s or more) the signal characteristics change to reflect the different voice sounds being spoken. Hence, short-time spectral analysis is the most popular technique or way to characterize the voice signal. A wide range of chances are exist for parametrically showing or representing the voice signal for the speech or voice recognition technique, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC is possibly the well known and very popular technique; this feature has been used in our project for voice or speech recognition. Mel Frequency Cepstrum Coefficient (MFCC) is based on the known changes or variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, logarithmically spaced filters and linearly spaced filters. To underline the phonetically important characteristics of voice or speech; signal is expressed in the Mel frequency scale. This scale has linear frequency slots below 1 KHz and a logarithmic spacing or slots above 1 KHz i.e. 1000 Hz. Ordinary voice signal waveform may Change from time to time corresponding to the physical state of speeches' vocal cord. Instead of the voice waveforms themselves, MFFCs are less susceptible to the given changes [1, 4].

(a) The MFCC processor

A block diagram of the structure of an MFCC processor is given in Figure 1. The input is recorded at a sampling rate of 2.2050 KHz. This sampling frequency is selected to reduce the effects of aliasing in the analog to digital conversion process. Fig.1 shows the block diagram of an MFCC processor. Continuous Speech Mel cepstrum

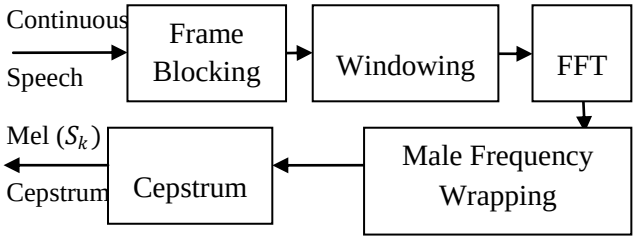

Fig. 1 Block diagram of the MFCC processor

(b) Mel-frequency wrapping

The voice signal consists of tones with various frequencies. For each tone with an actual Frequency (f), measured in Hz, a belonging pitch is measured on the 'Mel' scale. The Mel-frequency scale is linear frequency difference below 1 KHz and a logarithmic difference above 1 KHz. As a check point, the pitch of a 1 kHz tone, 40dB above the perceptual listening threshold, is defined as thousand Mels. Hence here we can use the formula to count the Mels for a given frequency f in Hz as follows [5]:

Mel (f) = 2595*log10 (1+f/700) …….. (1)

One way to simulating the subjective spectrum is to use a filter bank, in this approach one filter for each desired Mel-frequency component is used. The filter bank has a triangular band pass frequency response, and the spacing and the bandwidth is determined by a stationary Mel-frequency interval.

(C) Cepstrum

At the last, the log Mel spectrum should to be converted back to time domain. This result is called the Mel frequency cepstrum coefficients (MFCCs). The cepstral of the speech spectrum gives a better representation of the local spectral properties of the signal for the given frame analysis. As the Mel spectrum coefficients are real numbers, they may be changed to the time domain using the DCT i.e. Discrete Cosine Transform. The MFCCs may be calculated using following equation [3, 5]:

$$\tilde{C}_n = \sum_{k=1}^{K} (\log \dot{S}_k) \left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \ldots\ldots (2)$$

Where n=1, 2….K

The number of Mel cepstrum coefficients K, is typically selected as 20. The first component $\tilde{C}_0$ is achieved from the Discrete Cosine Transform i.e. DCT since it shows the mean value of the input signal which carries little speech specific information. By using the procedure described above, for every speech frame of nearly 30 ms with overlap, a set of Mel-frequency cepstrum coefficients is computed. This set of coefficients is called an *acoustic vector*. These acoustic vectors can be used to represent and recognize the voice characteristic of the speech [4]. Therefore each input word spoken is converted into a sequence of acoustic vectors. The next section clarifies how these vectors can be used to describe and recognize the voice characteristic of a speech.

4. FEATURE MATCHING

The state-of-the-art feature matching techniques used in speech recognition include, Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). The VQ approach has been used here for its ease of implementation and high accuracy.

Vector quantization

Vector quantization (VQ) is a data compression method based on principle of block coding [6]. It is a fixed-to-fixed length algorithm. Vector Quantization may be thought as an aproximator. Fig. 2 shows an example of a 2-D Vector Quantization.

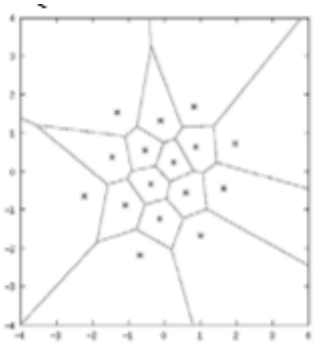

Fig. 2 2-D Vector Quantization

In vector Quantization, every set of two numbers falling in a particular region is guessed by a star associated with that region. In Figure 2, the stars are called code vectors and the regions defined by the borders are called encoding regions. The set of all code vectors is called the codebook. The set of all encoding regions is called the partition of the space [6].

5. SIGN LANGUGE

Sign language is the language used and can be understand by deaf and dumb people. It is a combination of shapes and movements of different parts of the body as the deaf and dumb people can't hear and speak the language used by normal people. These parts include face and hands. Facial expressions also count toward the gesture, at the same time. A posture on the other hand, is a static shape of the hand to indicate a sign. A sign language usually provides signs for whole words. The most popularly known sign language is ASSL i.e. American Standard Sign Language which is accepted world widely.

American Standard Sign Language

- American standard sign language was discovered around 1980 in United States and many other countries, to provide education for the people who have problem in speaking and hearing the words (the deaf and dumb).
- It is the most well documented and most widely used language in the world. American Standard Sign

Language (ASSL) is a complex visual-spatial language that is used by the Deaf community in the United States and English-speaking parts of Canada

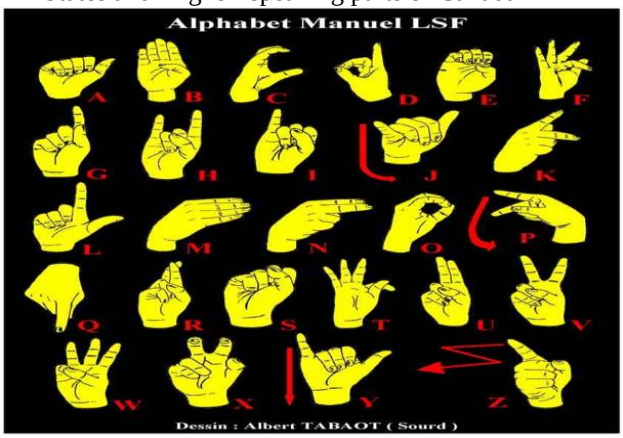

Fig. 3 American Standard Sign Language

6. BLOCK DIAGRAM

Block diagram of the overall system will be implemented. The system will be operating in close to real time and will take the speech input from the microphone and will convert it to synthesized speech or finger spelling. Speech recognition will

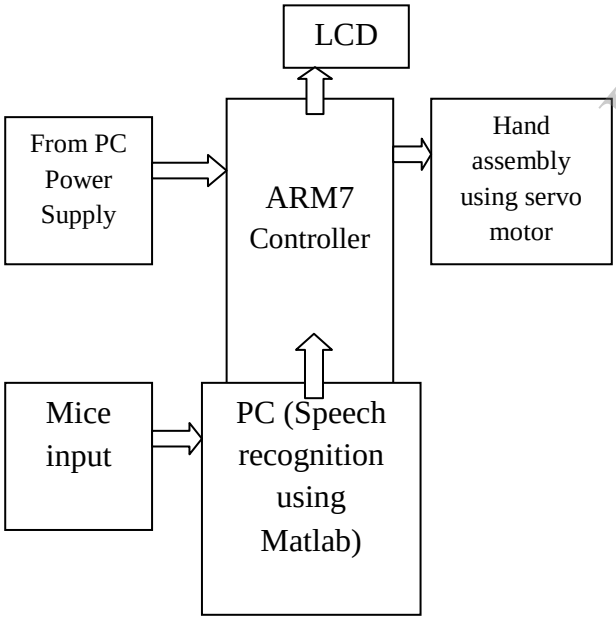

Fig.4 Block diagram of the overall system

be implemented for the considered languages. Language models will be used to solve ambiguities. Finger spelling synthesis will be implemented. Language model in mat lab is used to synthesis speech in to text. And given to arm controller which will have LCD to verify words spell and processed by processor.

ROBOTIC ARM

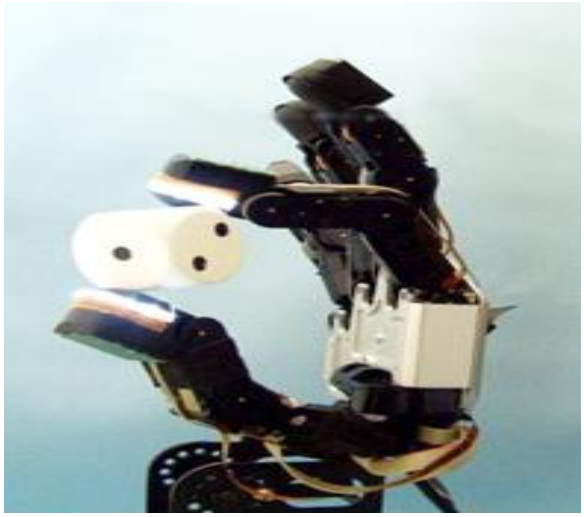

Fig. 5 Robotic Arm

Robotic arm is connected to arm to represent the vocal language in to sign language. The robotic arm uses servo motors for their movement. Speech recognition system based on the speech reading and the samples passed to the processing unit. The processing system consists of a speech recognition unit with symbol generator, which determines the speech signal and produces an equivalent coded symbol for their cognized speech signal.

7. RESULTS

1. Input speech signal

The input speech signal from mice to PC is given.fig. 6 shows the waveform of input speech signal.This speech signal is then recognized using matlab and given for windowing.

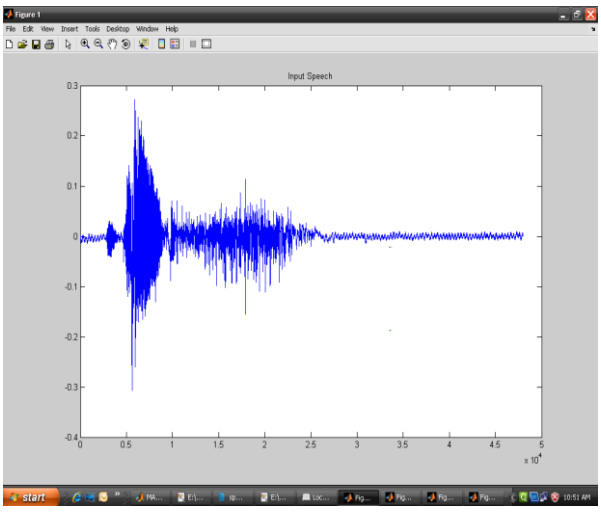

Fig. 6 Input speech signal

1. Before Windowing

Fig.7 shows the speech signal before windowing.

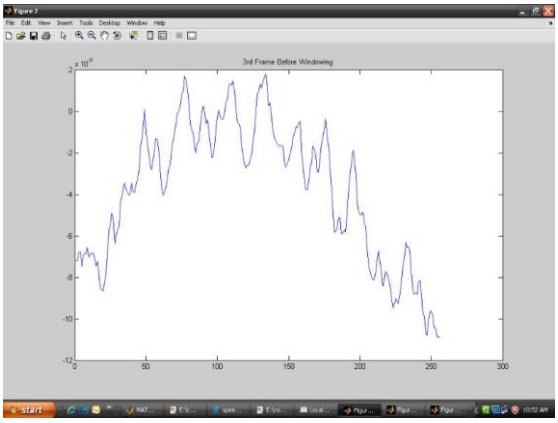

Fig. 8 Speech signal before widowing

2. After windowing of the signal

Windowing is used to minimize the discontinuities at both ends of the frame by approximating to zero and in remaining part of the wave bringing regularity by multiplying the wave by Hamming window function H (n).

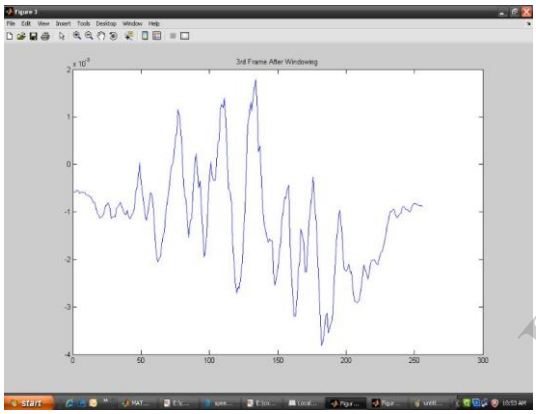

Fig. 9 Speech signal after windowing

H(n) is given as,

H(n)=0.54-0.46*cos(2*pi*n/(N-1));

Y[n]=x[n]*H[n]

3. Mel filter

Fig. 10 Shows filtering of the speech signal using Mel filter.Mel filter filters the speech signal on mel scale

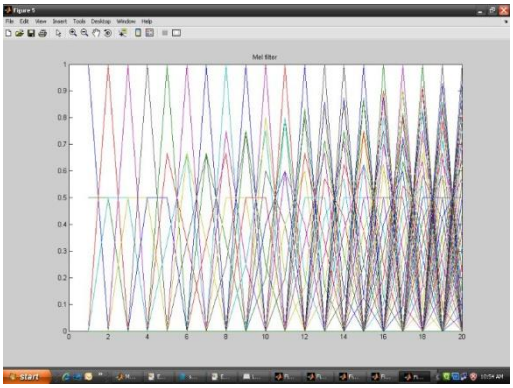

Fig. 10 Mel filter

4. After DCT (Discrete Cosine Transform)

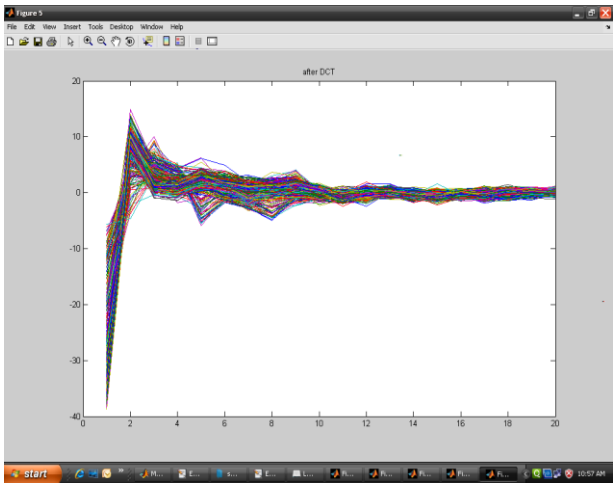

Fig. 11 speech signal after DCT

8. CONCLUSION

The MFCC technique has been applied for speech identification. VQ is used to minimize the data of the extracted feature. The study reveals that as number of centroids increases, identification rate of the system increases. It has been found that combination of Mel frequency and Hamming window gives the best performance. It also suggests that in order to obtain satisfactory result, the number of centroids has to be raised as the number of speakers increases. The observation shows that the linear scale can also have a reasonable identification rate if a relatively higher number of centroids are used. However, the recognition rate using a linear scale would be much lower if the number of speakers increases. Mel scale is also less vulnerable to the changes of speech's vocal cord in course of time.

The present study is still ongoing, which may include following further works. HMM may be used to improve the efficiency and precision of the segmentation to deal with crosstalk, laughter and uncharacteristic speech sounds. A more effective normalization algorithm can be adopted on extracted parametric representations of the acoustic signal, which would improve the identification rate further. Finally, a combination of features (MFCC, LPC, LPCC, Formant etc) may be used to implement a robust parametric representation for speech identification.

REFERENCES

1. Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology,Dhaka-1000, “Speaker Identification Using Mel Frequency Cepstral Coefficients”
2. Zhong-Xuan, Yuan & Bo-Ling, Xu & Chong-Zhi, Yu. (1999). “Binary Quantization of Feature Vectors for Robust Text-Independent Speech Identification” in IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 1, January 1999. IEEE, New York, NY, U.S.A.

3. F. Soong, E. Rosenberg, B. Juang, and L. Rabiner, "A Vector Quantization Approach to Speech Recognition", AT&T Technical Journal, vol. 66, March/April 1987, pp. 14-26
4. Comp.speech Frequently Asked Questions WWW site, http://svr-www.eng.cam.ac.uk/comp.speech/
5. Jr., J. D., Hansen, J., and Proakis, J. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York, 2000.
6. R. M. Gray, ``Vector Quantization," IEEE ASSP Magazine, pp. 4--29, April 1984.
7. Y. Linde, A. Buzo & R. Gray, “An algorithm for vector quantizer design”, IEEE Transactions on