



Intelligent System Enabling Speech to Sign Language Conversion

¹Manoj Sanker P.R, ²C.SameerDatta, ³Karthik A.R, ⁴Akshay Sarvesh, ⁵H L Shashidhar

Dept. of Telecommunication, PESIT,

Dept. of Electronics and communication, PESIT Bangalore

Email : ¹manoj.sanker.1992@gmail.com, ²csameerdatta.93@gamil.com, ³karthikreddy30@gmail.com, ⁴akshayaries.2511@gmail.com

Abstract- The main intent of this paper is to discuss an effective algorithm to convert speech signals using speech recognition algorithms to text output in the form of a multimedia display like image/video and its applications. The Speech recognition is done using the Mel Frequency coefficient calculations and the corresponding mapping is done using the Hidden Markov Model. We have focused on the application of the system in a class room environment the lecturer's speech signals are acquired and analyzed. This system can be adopted from a primary level in the education system of the deaf so that they are made better equipped for the challenges of the world. This would also tune them and get them adapted to a similar system at a higher level of such tech driven education. Thus the speech signal is converted into a American sign language(ASI) such that it's understandable for the deaf.

I. INTRODUCTION

Speech to text conversion of normal spoken syllables to understandable visual signal is the process of taking the spoken word as an input to an application or a program. This process is important to virtual reality because it provides a fairly natural and intuitive way of controlling the simulation while allowing the user's hands to remain free by not using keystrokes/board. This paper will delve into the process of speech recognition in the field of virtual reality, examine how speech recognition is accomplished, and list the academic disciplines that are central to the understanding and advancement of speech recognition technology. The above mentioned process is carried out by using software tools available to everyone commercially like Lab VIEW[1], hardware tools include mike, etc.

Speech recognition is "the technology by which sounds, words or phrases spoken by humans are converted into electrical signals, and these signals are transformed into classified patterns to which meaning has been assigned". While the abstraction could more generally be called "sound recognition", the focus is on human voice because we most often and most naturally use our voices to communicate to others. In the context of a virtual environment, the user would presumably gain the greatest

feeling of immersion, or being part of the simulation, if they could use their most common form of communication, the voice. The difficulty in using voice as an input to a program lies in the fundamental differences between human speech and the more traditional forms of computer input. While computer programs are commonly designed to produce a precise and well-defined response upon receiving the proper (and equally precise) input, the human voice and spoken words are anything but precise. Each human voice is different, and identical words can have different meanings if spoken with different inflections or in different contexts. Several approaches have been tried, with varying degrees of success, to overcome these difficulties.

Human speech contains quite a few discriminative features that can be used for identification. Speech contains significant energy from zero frequency up to around 5 kHz. The objective of Intelligent System Enabling Speech To Sign Language Conversion, is to extract, characterize and recognize the information about speaker's voice. The property of speech signal changes markedly as a function of time. To study the spectral properties of speech signal the concept of time varying Fourier representation is used. However, the temporal properties of speech signal such, as energy, zero crossing, correlation etc are assumed constant over a short period. That is its characteristics are short-time stationary. Therefore, using hamming window, Speech signal is divided into a number of blocks of short duration so that normal Fourier transform can be used. In this work, the Mel frequency Cepstrum Coefficient (MFCC) feature has been used for designing a text independent voice to sign language conversion system.

The sign language conversion system is an independent block in lieu of the above mentioned block, where the above mentioned block is just involved in speech recognition part, however the sign language conversion system employs a Lab VIEW program to output the necessary sign.

II. PRINCIPLE OF SPEECH RECOGNITION

A. Using Speech Recognition Algorithms

A voice analysis is done after taking an input through microphone from a user. The design of the system involves manipulation of the input audio signal. At different levels, different operations are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel Cepstrum analysis and Recognition (Matching) using the Hidden Markov Model algorithm of the spoken word, display of the necessary multimedia animation in American Sign Language.

The speech algorithms consist of two distinguished phases. The first one is training sessions, whilst, the second one is referred to as operation session or testing phase as described in figure 1 [2].

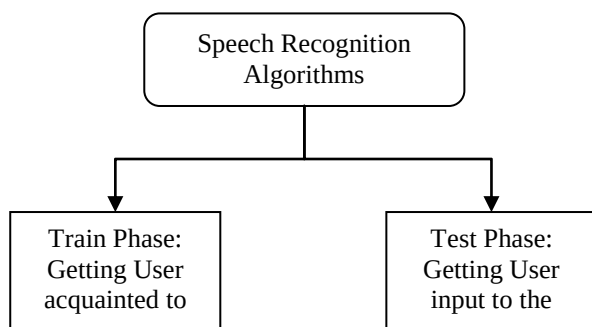


Fig.1

B. Calculation of Mel Frequency Coefficients

Perceptual experiments say human ear concentrates on certain regions rather than using whole of the spectral envelop.

Mel-Frequency analysis of speech is based on human perception of these experiments. In other words, in Mel Frequency Cepstrum Coefficients (MFCC) is based on known variation of the human ear's critical bandwidth with frequency

It is observed that human ear acts as a smart filter, concentrating on only certain frequency Components. These filters are non-uniformly spaced on the frequency axis i.e more filters are present in the low frequency regions, less no. of filters are present in high frequency regions

As shown in figure[3].

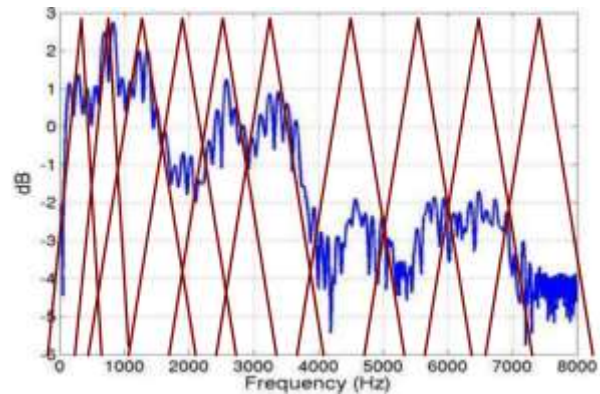


Fig.2

Steps involved in calculating Mel Frequency Cepstral Coefficients':[4]

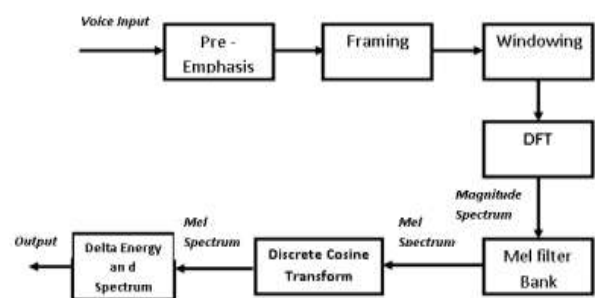


Fig.3

Step 1: Pre-emphasis [5]

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y[n] = X[n] - 0.95 X[n-1]$$

Let's consider $a = 0.95$, which make 95% of any one sample is presumed to originate from previous sample.

Step 2:Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M ($M < N$). Typical values used are $M = 100$ and $N = 256$.

Step 3: Windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as:

If the window is defined as $W(n)$, $0 \leq n \leq N-1$ where

N = number of samples in each frame

$Y[n]$ = Output signal

$X[n]$ = input signal

$W[n]$ = Hamming window, then the result of windowing signal is shown below:

$$Y[n] = X[n] * W[n]$$

Where $w[n]$ is

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

Step 4: Fast Fourier Transform

To convert each frame of N samples from time domain into

Frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse $U[n]$ and the vocal tract impulse response $H[n]$ in the time domain. This statement supports the equation below:

$$Y(\omega) = \text{FFT}[h(t) * X(t)] = H(\omega) * X(\omega)$$

If $X(\omega)$, $H(\omega)$ and $Y(\omega)$ are the Fourier Transform of $X(t)$, $H(t)$ and $Y(t)$ respectively.

Step 5: Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale as shown in figure 4[5] is then performed.

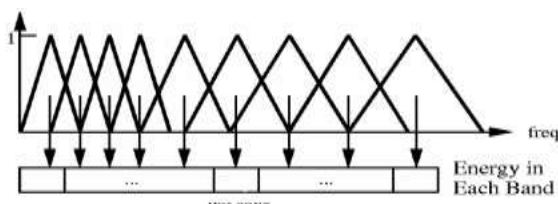


Fig.4 Mel Filter bank

This figure shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [7, 8]. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in HZ: [5]

$$F(\text{mel}) = \lceil 2595 * \log_{10} \left[1 + \frac{f}{700} \right] \rceil$$

Step 6: Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

Step 7: Delta Energy and Delta Spectrum

The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time. 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t_1 to time sample t_2 , is represented at the equation below:

$$\text{Energy} = \sum [x(t) * x(t)]$$

Each of the 13 delta features represents the change between frames in the equation 8 corresponding cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features.

III. METHODOLOGY

Speech recognition works based on the premise that a person voice exhibits characteristics are unique to different speaker. The signal during training and testing session can be greatly different due to many factors such as people voice change with time, health condition, speaking rate and also acoustical noise and variation recording environment via microphone. Table II gives detail information of recording and training session, whilst Figure 5 shows the flowchart for overall voice recognition process, also the final Multimedia Display using LabVIEW

Once the features are extracted after employing the MFCC calculation, the Hidden Markov Model (HMM) algorithm is used to match the extracted features with the features of the dictionary samples, if there is a match then the speech is said to be recognized and the corresponding dictionary sample is said to be the recognized word.

Hidden Markov Model:

The hidden Markov model is used to compare the spoken word with the dictionary of sample words to find the right match, after feature extraction we obtain a set of coefficients exclusive to a given word,

Let $\mathbf{Y}_{1:T} = \mathbf{y}_1, \dots, \mathbf{y}_T$ be the features extracted of the spoken word then the decoder employing the HMM attempts to find the sequence of words $\mathbf{w}_{1:L} = \mathbf{w}_1, \dots, \mathbf{w}_L$ which is most likely to have generated $\mathbf{Y}$. the decoder tries to find

$$\mathbf{w}^* = \arg \max \{P(\mathbf{w}|\mathbf{Y})\}.$$

Each spoken word is broken down into a sequence of sounds; such a sequence is called its pronunciation denoted by $\mathbf{q}^{(w)}_{1:Kw} = \mathbf{q}_1, \dots, \mathbf{q}_{Kw}$. But a single word can have a variety of pronunciations, to accommodate that the likelihood of the word is computed over multiple pronunciations

$$p(\mathbf{Y}|\mathbf{w}) = \sum p(\mathbf{Y}|\mathbf{Q})P(\mathbf{Q}|\mathbf{w}).$$

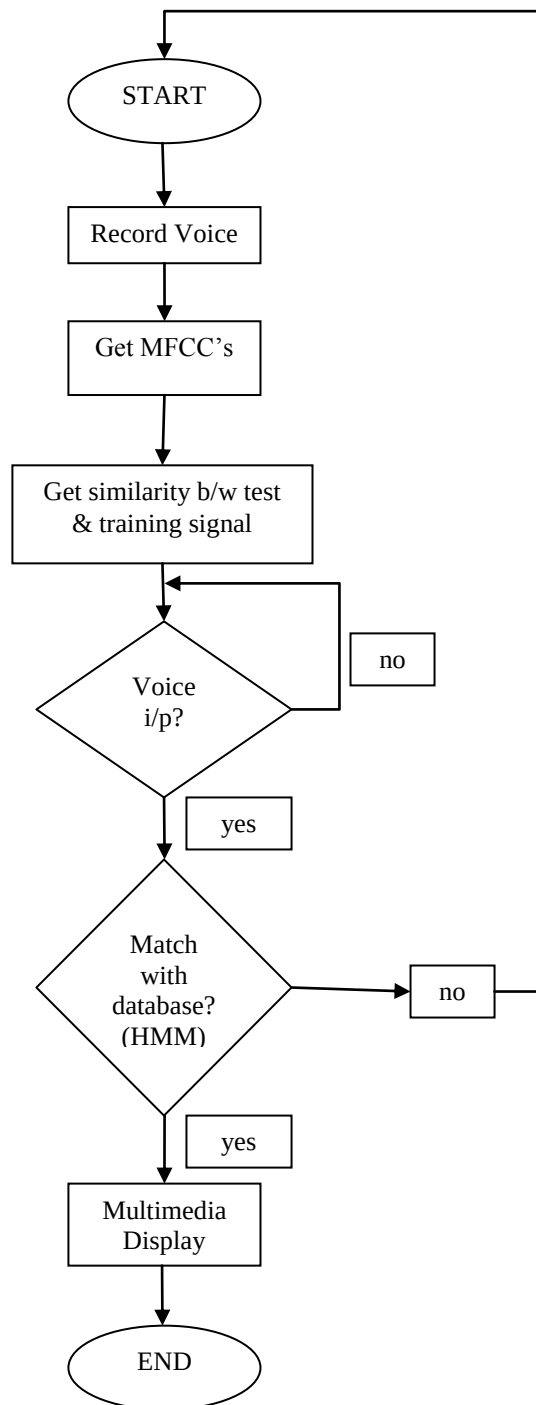


Fig 5.

Each base sound is described by a continuous density HMM with a transitional probability parameter $\{a_{ij}\}$ and an output observational parameters $\{b_j(\cdot)\}$. The HMM makes a transition from one state to the next with a transitional probability of $\{a_{ij}\}$ and subsequently on entering a new state a feature vector is generated using the distribution associated with the state being entered, $\{b_j(\cdot)\}$. This process is performed for each of the base sounds and the output distributions are concatenated to obtain the acoustic likelihood which is given by $p(Y|Q) = \prod p(\theta, Y|Q)$, where $\theta = \theta_0, \dots, \theta_{T+1}$ is a state sequence through the composite

model. This likelihood estimation forms the base of matching the word with the dictionary sample

Part of speech tagging for grammar conversion:

Once the speech is translated to text, for the translated sentence part of speech tagging is performed using Maximum Entropy Markov Model. Each word in the sentence is tagged to a given set of parts of speech such that the total conditional probabilities of observations and labels is maximized. The total conditional probability is broken down into Markov transitional probabilities where the where the probability of transitioning to a particular label depends only on the observation at that position and the previous position's label. Once the probabilities are determined, the sequence of label is determined using the viterbi algorithm.

Once each word in a sentence is tagged to its most probable part of speech, we can convert the sentence to sign language animation with the correct grammar so that the hearing impaired can understand.

IV. RESULTS AND CONCLUSION

This paper has discussed a speech recognition algorithm whose technique was able to identify the particular word spoken by a speaker based on the individual information that was included in the speech signal. The results show that these techniques can be effectively used for speech recognition purposes. Several other techniques involving the use of Artificial Neural networks are in research and would definitely be employed in any future works.

V. ACKNOWLEDGEMENT

To complete this work, we have got valuable suggestions and guidance from Dr. Srinivas, PESIT of this field. The Book Digital Processing of Speech Signals. by L.R. Rabiner and R.W. Schafer is very helpful for beginners' to start research in this field. We are also thankful to all the researchers and enthusiastic authors of several informative blogs/articles which gave us valuable suggestions during this endeavor. We also like to thank PES Institute Of Technology for supporting us throughout the project.

VI. REFERENCES

- [1] LabVIEW is system design software by NI that provides engineers and scientists with the tools you need to create and deploy measurement and control systems. Securely download a fully functional version of LabVIEW Professional for Microsoft Windows (32- and 64-bit) or Mac OS
- [2] <http://www.cse.unsw.edu.au/~waled/phd/html/node38.html>
- [3] Speech Technology: A Practical Introduction, Kishore Prahallad, Carnegie Mellon University & International Institute of Information Technology Hyderabad.

- [4] Feature Extraction and start of Acoustic Modeling (VQ) Stanford University Lectures, by Dan Jurafsky.
- [5] International Journal of Advanced Research in Computer Science and Software Engineering, Speaker Identification using Mel Frequency Cepstral Coefficient and BPNN.
- [6] Journal of Computing, Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques.
- [7] http://mi.eng.cam.ac.uk/~mjfg/mjfg_NOW.pdf

