

ADVANCED SPEECH TO TEXT RECOGNITION USING OPEN CV

DivyaBharathy.B,Gowri.B,VishnuPriya.S

Student, Department of Electronics and Communication Engineering ,Vel Tech Engineering College, Chennai,India
email:agalchagowri2330@gmail.com

Mr.Lokesh.S M.E

Assistant professor,Department of Electronics and CommunicationEngineering,Vel Tech Engineering
College,Chennai,India,email:lokesh@veltechengg.com

Abstract : First, cochlear implant performance has remained largely unchanged, despite the number of publications tripling per decade in the last 30 years. Little has been done so far to address a fundamental limitations in the electrode-to-neuron interface, with the electrode size being a thousand times larger than the neuron diameter while the number of electrodes being a thousand times less. In this paper we will research word based, phone based and phonologic-feature-based audio and audio visual speech recognition models. To develop low word error rate speech recognition system using wavelet transform through pattern recognition approach. The aim is to make an intelligent system that can recognize the speech signal. This also includes the feature extracted from the speech signal using discrete wavelet transform. Dynamic Time Warping is used for pattern matching from the stored data base of stored pattern to recognize the word.

Keywords:Back propagation neural network (BPNN),Discretewavelet transform(DWT),Computer version (CV)

Introduction: Historically the sound of languages have been studied at two different levels [1] phonetic components of spoken words [2] Acoustic wave patterns. Language can be broken down into a small number of basic sounds called phonemes. An acoustic wave is a sequence of changing vibration patterns. In sound analysis there are two dimensional patterns called spectrogram, which displays frequency(vertical axis) Vs. time(horizontal axis). The power source for consonants is airflow producing white noise, while the power for vowels is vibrations from the vocal chords. The analog processors generally use filter banks to segment the

voice spectrum. With the advent of digital signal processing and digital computers the beginning of modern automatic speech recognizers(ASR) are used. Voice recognition by computer is used in access control and security system.

LITERATURE SURVEY: Challenges in improving cochlear implant [1] To identify two gaps in cochlear implants that have been limiting their performance and acceptance. The high cost contributes to low market penetration rate, which is about 20% in developed countries and less than 1% in developing countries.It will discuss changes needed in both research strategy and business practice to close the gap between prosthetic and normal hearing as well as that between haves and have-nots.

Modeling electrode place discrimination in cochlear implant stimulation [2] By modeling the CI electrode- to- nerve interface and quantifying electrode discriminability in the model, it address the questions of how many individual channels can be distinguished by CI recipients and the extent to which performance might be improved by inserting electrodes deeper into the cochlear.

Impedance measures during in vitro cochlear [3] an advanced bionics CI array as inserted into a clear, plastic cochlear one electrode contact at a time in a saline bath. These measures were correlated to actual proximity as assessed by microscopy using linear regression models.

Evaluation of different cochlear implants in unilateral hearing [4] Advanced methodologies used for the biomedical signal interpretation allow using cerebral signals to assess important cognitive functions in humans. Thehypothesis was that the lower the number of interconnections engaged, the lower the

cerebral effort induced by the experimental condition. In the present research this index has been applied to test the reaction to the use of different cochlear implant processors[5-11].

PROPOSED SYSTEM:

The general public's understanding of speech recognition comes from things such as HAL 9000 computer in Stanley Kubrick's film 2001: A Space Odyssey. Notice that HAL is a perversion of IBM. At the time of movie's release(1968) IBM was just getting started with a large speech recognition project that led to a very successful large vocabulary isolated word dictation system and several small vocabulary control systems.

The a-to-d conversion is generally accomplished by digital signal processing hardware on the computer's sound card. The typical sampling rate 8000 samples per second is adequate. The spoken voice is considered to be 300-3000 Hz. A sampling rate 8000 gives a Nyquist frequency of 4000 Hz which should be adequate for a 3000 Hz voice signal. The sample resolution is the 8 or 16 bits per second that sound cards can accomplish.

Speaker recognition can be classified into identification and verification. Speaker identification is a process of determining which registered speaker provides a given utterance. Speaker verification is a process of accepting or rejecting the identity claim of the speaker. The figure 1 shows a basic structure's of speaker identification and verification system. The system that we will describe is classified as text-independent speaker identification system since it's task is to identify the person who speaks regardless of what is saying.

At the highest level, all speaker recognition systems contain two main modules(refer to fig. 1); feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from

the voice signal that can be later used to represent each speaker. Feature matching involves actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input once from a set of known speakers .

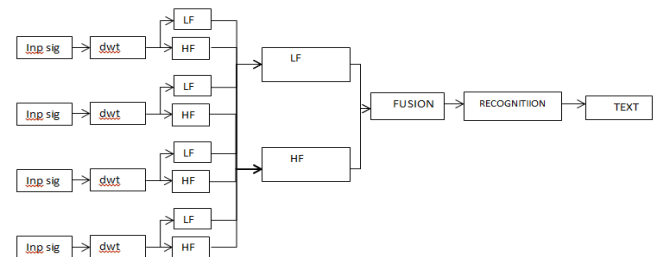


Figure 1: Block diagram of signal extraction

Speaker recognition is a difficult task. Automatic speaker recognition works based on premise that a person's speech exhibits characteristics that are unique to the speaker. An example of speech signal is given in figure 2. However this task has been challenged by the highly variant of input speech signals.

The purpose of this model is to convert the speech waveform, using digital signal processing(DSP) tools, to a set of features(at a considerably lower information rate) for further analysis. This is often referred as a signal processing front end.

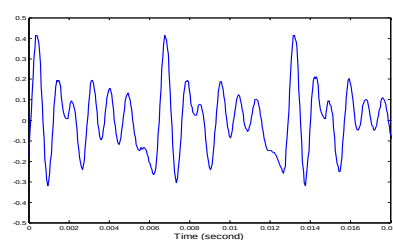


Fig:2 Example of Speech signal

A block diagram of structure of an MFCC is given in fig 3. The speech input is typically recorded at a sampling rate about 1000 HZ. This sampling frequency was chosen to minimize the effects of aliasing in the a-to-d conversion. These sampled signals can capture all frequencies upto 5 KHZ which cover most energy of sounds that are generated by units.

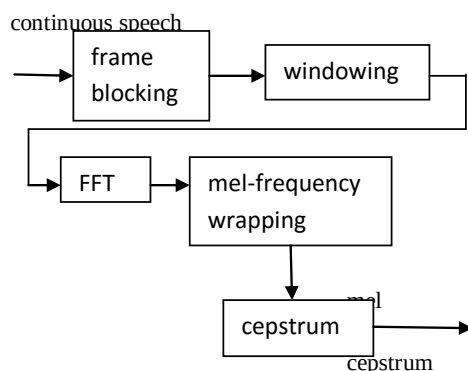


Figure 3:Block diagram of MFCC Processor

The next processing step is fast fourier transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the discrete fouriertransform(DFT), which is defined on the set of N samples $\{X_n\}$, as follows:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \quad k = 0,1,2,\dots, N-1$$

Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency F measured in HZ a subjective pitch is measured on a scale called the 'Mel' scale. The Mel frequency scale is a linear frequency spacing below 1000 HZ and a logarithmic spacing above 1000 HZ.

A useful way of thinking about this Mel-wrapping filter bank is to view each filter as a histogram bin in the frequency domain. The number of Mel spectrum coefficients K is typically chosen as 20.

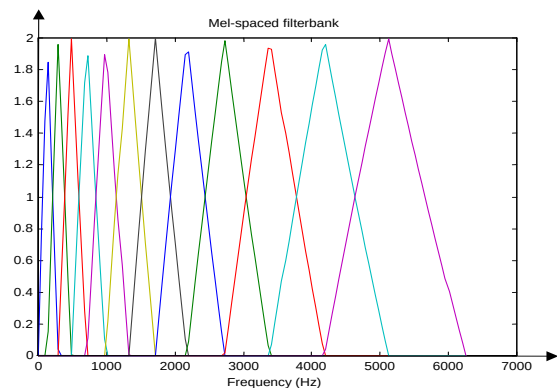


Figure 4: An example of mel-spaced filter bank

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called pattern recognition. The figure 4 drawn above describes about the waveforms obtained in the mel-frequency filters, the goal of pattern recognition is to classify objects of interest into one of the number of categories or classes. The object of interest are generically called patterns and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our cases is applied on extracted features, it can be also referred to as feature matching.

In the proposed system we are using discrete wavelet transform instead of fourier transform in order to split the given signal into high frequency signal and low frequency signal. Here we are using neural networks classifier to differentiate and classify various emotions.

A voice frequency (VF) or voice band is one of the frequencies, within the part of the audio range that is used for the transmission of speech.

In telephony, the usable voice frequency band ranges from approximately 300 HZ to 3400 HZ. It is for this reason that the ultra low frequency band of the electromagnetic between 300 HZ to 3000 HZ is also referred to as voice frequency ,being the electromagnetic energy that represents acoustic energy at baseband. The bandwidth allocate for a single voice frequency transmission channel is usually 4 KHZ including guard bands, allowing a sampling rate of 8KHZ to be used as the basis of the

pulse code modulation for the digital PSTN. However, enough of the harmonic series will be present for the missing fundamentals to create the impression of hearing the fundamental tone.

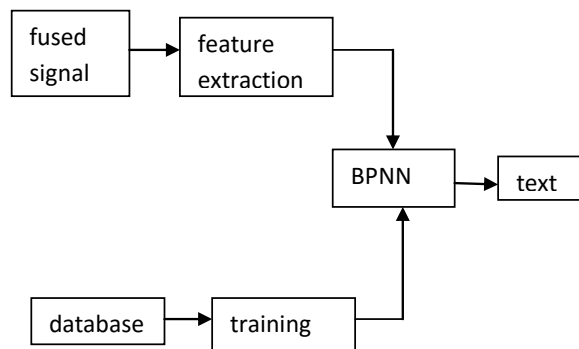


Figure 5: Recognition block

The figure 5 shoes about the recognition block, It recognizes the following ,

Input voice signal, Pre processing, Wavelet transform on low pass signal, Features extraction on mean peak point minimum level, Data set trained voice, BPNN training

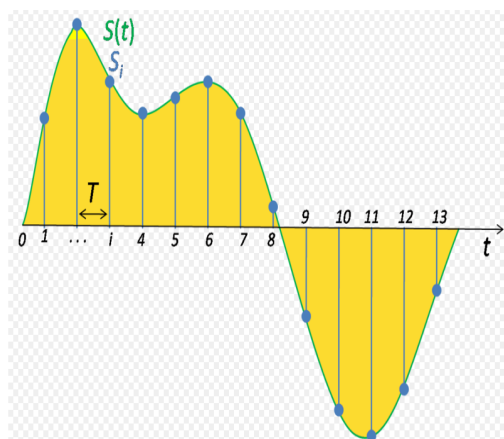


Figure 6:Sampling of signals

Quantization in mathematics and digital signal processing is the process of mapping

a large set of input values to a smaller set such as rounding values to some unit of precision. The term quantization and sampling are important factors for signaling the waves obtained for sampling of signals is shown in figure 6. Quantization is involved to some degree in nearly all digital signal processing as the process of representing the signal in digital form ordinarily involves rounding.

There are different of wavelets. They are, Haar wavelet - it is a bipolar step function. The Haar wavelet is discontinuous in time and it has good localization property in time and frequency domain.

Morlet wavelet - The Morlet wavelet is obtained by multiplying the fourier basis with a Gaussian window.

Mexican-Hat wavelet- Mexican wavelet is a 2-order derivative of the Gaussian function.

Daubechies wavelet- Daubechies wavelet bases are a family of orthonormal compactly supported scaling and wavelet functions that are maximum regularity for a given length of the support of the quadrature mirror filters.

The straight forward convolution implementation of 1-D DWT requires a large amount of memory and large computation complexity. An alternative implementation of the 1-D DWT known as the lifting scheme, provides significant reduction in the memory and computation complexity. Lifting also allows in-place computation of the wavelet coefficients.

Nevertheless the lifting approach computes the same coefficients as the direct filter bank convolution.

Speech signal composed of large numbers of parameters

which indicates emotion content of it. Changes in this parameters indicates changes in the emotions. The acoustic parameters related to speech quality are voice level: signal amplitude, energy and

duration have been shown to be reliable measures of voice level.

Mean value detection:

- Minimum stem on input voice signal
- Feature average = mean (low pass signal)

Result:

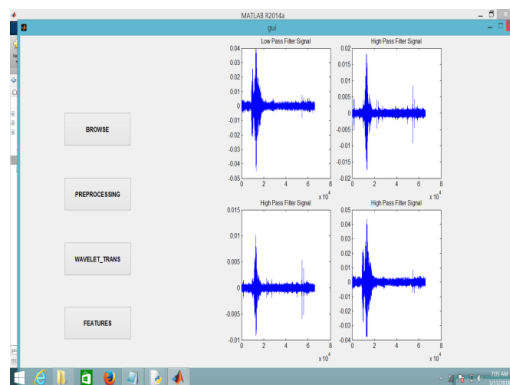


Figure 7: Output Waveform

The feature extraction and recognition process for speech signal is done using discrete fourier transform(DWT) .The waveform for the speech signal is obtained in four steps, in the first step the voice recorded is browsed and then the preprocessing process is done. The processed output signal shown in figure 7 gives the wavelet transform signal. Further the extraction process is obtained which describes the frequency of the signal.

Conclusion: In this work the field of speech emotion recognition is discussed. Most used methods of feature extraction and several classifier performances are viewed. Cochlear implants are now widely considered as the most successful neural prosthesis. The value of pitch frequency can be calculated in each speech frame and the statistics of pitch can be obtained in the whole

speech sample. These statistical values reflect the global properties of characteristic parameters. Emotional contents of a utterance is strongly related with its voice quality. The voice quality can be numerically represented by parameters estimated directly from speech signal. The acoustic parameters related to speech quality are: (1) Voice level: signal amplitude, energy and duration have been shown to be reliable measures of voice level; (2) voice pitch; (3) phrase, phoneme, word and feature boundaries; (4) temporal structures.

References:

- [01] R. J. Briggs, M. Tykocinski, R. Lazsig, A. Aschendorff, T. Lenarz, T. tover, *et al.*, "Development and evaluation of the modiolar researcharray--multi-centre collaborative study in human temporal bones," *Cochlear Implants Int*, vol. 12, pp. 129-39, Aug 2011.
- [02] J. L. Pinyon, S. F. Tadros, K. E. Froud, Y. W. AC, I. T. Tompson, E. N. Crawford, *et al.*, "Close-field electroporation gene delivery using the cochlear implant electrode array enhances the bionic ear," *SciTranslMed*, vol. 6, p. 233ra54, Apr 23 2014.
- [03] B. A. Nayagam, S. S. Backhouse, C. Cimenkaya, and R. K. Shepherd, "Hydrogel limits stem cell dispersal in the deaf cochlea: implications for cochlear plants.
- [04] F. Zeng *et al.*, "Development and evaluation of the neurotron 26-electrode cochlear implant system," *Hearing Research*, vol. 322, pp. 188–199, 2015.
- [05] Dr. Anto Bennet, M, Sankar Babu G, Natarajan S, "Reverse Room Techniques for Irreversible Data Hiding", *Journal of Chemical and Pharmaceutical Sciences* 08(03): 469-475, September 2015.
- [06] Dr. Anto Bennet, M, Sankaranarayanan S, Sankar Babu G, "Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis", *Journal of Chemical and*

Pharmaceutical Sciences 08(03): 571-576, August 2015.

[07] Dr. AntoBennet, M, Suresh R, Mohamed Sulaiman S, "Performance & analysis of automated removal of head movement artifacts in EEG using brain computer interface", Journal of Chemical and Pharmaceutical Research 07(08): 291-299, August 2015.

[08] Dr. AntoBennet, M "A Novel Effective Refined Histogram For Supervised Texture Classification", International Journal of Computer & Modern Technology , Issue 01 ,Volume02 ,pp 67-73, June 2015.

[09] Dr. AntoBennet, M, SrinathR, Raisha Banu A, "Development of Deblocking Architectures for block artifact reduction in videos", International Journal of Applied Engineering Research, Volume 10, Number 09 (2015) pp. 6985-6991, April 2015.

[10] AntoBennet, M & Jacob Raglend, "Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images", Journal of Computer Science, vol. 8, no. 9, pp. 1447-1454, 2012.

[11] AntoBennet, M & Jacob Raglend, "Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering", European Journal of Scientific Research, vol. 66 no.1, pp.120-129, 2011.

