www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issues 8 Aug 2016, Page No. 17590-17592

Effect Of Combination Of Different Features On Speech Recognition For Abnormal Speech

Ms. Yogita A. More[1], Mrs. S. S. Munot(Bhabad)[2]
Dept. Of Electronics And Telecommunication, K.K.W.I.E.E.R, Nashik, India
yogitamore.30@gmail.com,ssb.eltx@gmail.com

***Abstract*- Speech is very important for communication. With the help of speech one can express their thoughts, interact with people easily. Feature Extraction is the very first step in speech recognition. This paper presents the effect of combination of different features on speech recognition. Different features of speech such as spectral, temporal, perceptual and energy. Also accuracy for each feature is calculated and it seems that Mel Frequency Cepstral Coefficients (MFCC) gives better accuracy among all the features. And the accuracy of MFCC is further increased by combination of MFCC with bark.**

Keywords-Speech Recognition, Spectral Features, Temporal Features, Perceptual Features, Energy, Mel Frequency Cepstral Coefficients (MFCC).

1. INTRODUCTION

Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. Speech is easiest way for communication. Many people in world are deaf. Speech communication is a difficult task for them. Speech recognition system is growing area of research. Many researchers are working for solving the difficulties and problems faced by speech disabled people. They have worked for developing systems in phoneme correction, word correction, etc. But research on speech correction is still a challenging area. Speech recognition involves stages such as Pre-processing, Feature Extraction, Classification and Verification[3].

Speech has been very active research field over recent last year. Although a lot of features have been proposed and many classifiers have been employed, but there are few works on comparing these features. Speech is a form of communication in everyday life. It is essential to know how we produce and perceive it and how speech technology may assist in communication. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signal.

In this paper, speech recognition concept is explained. Also feature extraction technique is explained. And in next sections results and conclusion.

2. DATABASE CREATION

Database Creation is the very first step in this system. 10 users database is created from the digit zero to ten. Every user speech is recorded for 10 times. Total 1100 samples are collected.

3. FEATURE EXTRACTION

Theoretically, it should be possible to recognize speech directly from the digitized waveform. However, because of the large variability of the speech signal, it is better to perform some feature extraction that would reduce that variability. Particularly, eliminating various source of information, such as whether the sound is voiced or unvoiced and, if voiced, it eliminates the effect of the periodicity or pitch, amplitude of excitation signal and fundamental frequency etc. Feature extraction is the process of retaining useful information of the signal while discarding redundant and unwanted information or we can say this process involves analysis of speech signal [6]. However, in practice, while removing the unwanted information, on may also lose some useful information in the process [7]. Feature extraction may also involve transforming the signal into a form appropriate for the models used for classification. Feature extraction techniques are classified as follows:

A. Spectral Features

i. Spectral Centroid

The spectral centroid is defined as the center of gravity of the magnitude spectrum of the STFT,

$$c_t = \frac{\sum_{n=1}^{N} n.|M_t[n]|^2}{\sum_{n=1}^{N} |M_t[n]|^2} \quad ....(1)$$

where M_t is the magnitude of the Fourier transform at frame t and frequency bin n.The centroid is a measure of spectral shape and higher centroid values correspond to “brighter" textures with more high frequencies.

ii. Spectral Roll-Off

The spectral roll-off is defined as the frequency Rt below which 85% of the magnitude distribution is concentrated

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 \times \sum_{n=1}^{N} M_t[n] \quad ....(2)$$

iii. Spectral Spread

It describes concentration of the spectrum around the centroid and is define as,

$$Ss_r = \sqrt{\frac{\sum_{k=1}^{N/2}[log_2(F[k]1000) - Sc_r]^2 Pr[k]}{\sum_{k=1}^{N/2} Pr[k]}} \quad ....(3)$$

Where F[k] is the frequency at bin k. Ss_r is spectral centroid, Pr is the power spectrum of the frame.

iv. Mel Frequency Cepstral Coefficients(MFCC)

MFCC is based on the human peripheral auditory system. The 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 *t* 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 logarithmic spacing above 1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

B. Temporal Features

i. Loudness(Energy)

The global energy of the signal x[n] can be computed simply by taking the root average of the square of the amplitude, also called root-mean-square energy (RMS).

$$x_{rms} = \sqrt{\frac{1}{2}\sum_{i=1}^{n} x_i^2} \quad ....(4)$$

ii. Autocorrelation

Related to the time domain feature detector is the autocorrelation method. The autocorrelation of the signal is first formed:

$$r(\tau) = \int_{-\infty}^{\infty} x(t)x(t+\tau)dt \quad ....(5)$$

iii. Log Attack Time

Log attack time is the logarithm of the time it takes from the beginning of a sound signal to the point in time where the amplitude reaches a first significant maximum.

LAT = log10(attacktime)

iv. Temporal Centroid

On the same principle of spectral centroid, the temporal centroid is the temporal center of mass of the features rather than spectrum.

$$centroid_i = \frac{\sum_{t=0}^{N-1} x_i(t) * t}{\sum_{t=0}^{N-1} x_i(t)} \quad ....(6)$$

Where xi(t) represents the value of feature i at the frame t, N is the number of frames inside a texture window.

v. Zero Crossing Rate

Zero crossing rate is the number of time domain waveform sign-change rate within a frame.

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \quad ....(7)$$

4. RESULTS

Figure 1 shows accuracy of all the features and figure 2 shows the combination of features with MFCC.

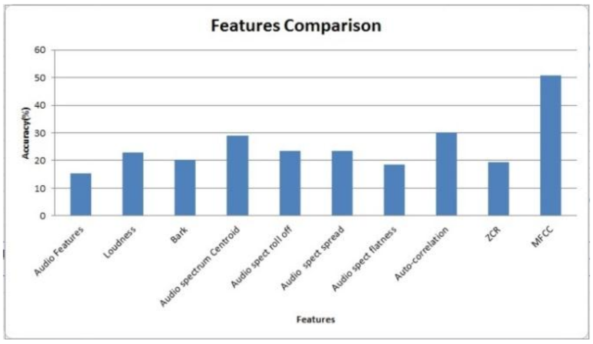

Figure 1: All Features

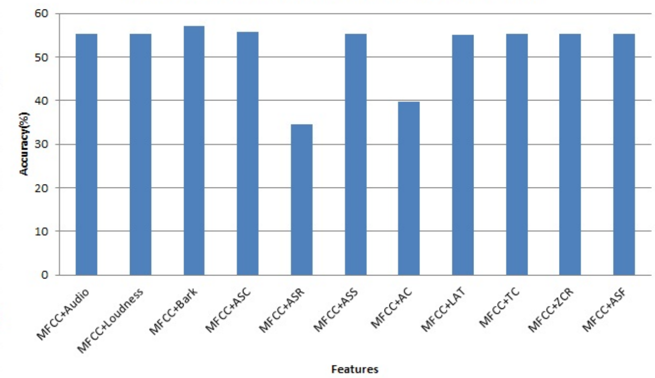

Figure 2: All Features Combination With MFCC

5. CONCLUSION

Feature extraction is the very first step in speech recognition. In this paper, various features are explained with their accuracy. Among all the features Mel Frequency Cepstral Coefficients (MFCC) gives highest accuracy(50.81%) with hamming window. Therefore, further this accuracy is increased by combining each feature with MFCC. MFCC with Bark gives 57% accuracy.

6. REFERENCES

[1]C.Jeyalakshmi,Dr.KrishnamurthLV,Dr.A.Revathi, “Speech Recognition of Deaf and Hard of Hearing People Using Hybrid Neural Network," *2nd International Conference on Mechanical and Electronics Engineering (ICMEE 2010).*

[2] JYoucef TABET Mohamed BOUGHAZI ,”Speech synthesis techniques. A survey," 2011 7th International Workshop on Systems”, *Signal Processing and their Applications (WOSSPA).*

[3]Zul_qar Ali, Mansour Alsulaiman, Ghulam Muhammad,Irraivan Elamvazuthi,”Vocal fold disorder detection based on continuous Speech by using MFCC and GMM," *2013 IEEE GCC Conference and exhibition, November 17-20, Doha,Qatar.*

[4]Ananthi. S, Dhanalakhsmi. P “Survey about Speech Recognition and Its Usage for Impaired (Disabled) Persons",*International Journal of Scientific Engineering Research Volume 4, Issue 2,February-2013.*

[5] Rohit Kumar, “A Genetic Algorithm for Unit Selection based Speech Synthesis,"*International Conference On Spoken Languages Processing(ICSLP),Oct 2004.*

[6] Tao Jiang,ZhiyongWu, Jia Jia, Lanhong Cai, “Perceptual Clustering Based Unit Selection Optimisation For Concatenative Text-To-Speech Synthesis," *IEEE,2012.*

[7] Masatsune Tamura, Norbert Braunschweiler, Takehiko Kagoshima, and Masami Akamine, “Unit Selection Speech Synthesis Using Multiple Speech Units At Non-adjecent Segment For Prosody And Waveform Generation," *IEEE,2010.*

[8] Shahidhar G. Koolagudi,Deepika Rastogi and K. Sreenivasa Rao ,”Identification Of Language Using Mel-Frequency Cepstral Coe_cients(MFCC)," *SciVerse ScienceDirect, International Conference On Modelling, Optimisation and Computing(ICMOC).*