Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Frequency Analisys of Speech Signals for Devanagari Script and Numerals Using FFT

Umesh Kumar Gupta
M-Tech Student of Electronics Engg. Department
B.V.D.U. College of engineering, Pune, India.

Dr. R. K. Prasad
Electronics Engg. Department
B.V.D.U. College of Engg. Pune, India.

Abstract: - This paper contains the frequency analysis of spoken Devnagari script and Numerals from the original speech signals. Devnagari vowels and numerals are playing the vital role in pronunciation of any word or counting. Each vowel & number is classified as starting, middle and end according to the duration of occurrences in the word. The Devnagari script having 12-vowels and 34-consonants are used in some Indian language like Hindi and 10 numerals (0-9) are used in mathematics. Sound samples from multiple speakers were utilized to extract different features. Initial processing of data, i.e., normalizing and time-slicing was done using a combination of Simulink and MATLAB. Afterwards, the same tools were used for calculation of Fourier descriptions and correlations. The correlation allowed comparison of the same words or numeral spoken by the same and different speakers. So the frequency has been calculated in statistical manner and generates a table between amplitude and frequencies. Mean and standard deviation such a system can be potentially utilized in implementation of a voice-driven help setup at call centres of commercial organizations operating in India and other foreign region. The implementation, experiments and result discussions are also existence.

Keywords: Spoken Hindi word & numerals, Fourier descriptors, Correlation, Mel Frequency Cepstral Coefficient (MFCC) and Feature extraction.

I. INTRODUCTION

Fundamental frequency estimation has been a popular topic in many fields of research. Such as speech synthesis, speech processing, speaker identification etc. The Devnagari vowels and numerals cannot pronounce two ways but it can be pronounced only one way e.g. Devnagari 12-vowels are classified with the phonetic transcription structure of phonemes according to organ used in produce the sound. Devnagari is based on phonetics principles which are considered as Place of articulation (POA) vowels. These Devnagari vowels having Frequency analysis of speech signals are estimated in noisy environment (original signals) for analysis and synthesis. The original speech signals are unbalanced to adjustment of an interval with help of some feature extraction techniques or use Sound Forge 9.0 software. The initial objective is to estimating the pitch of Devnagari vowels and numerals with noisy environments speech signals. When one looks at a person, car or house, one's brain tries to match the incoming plot with hundreds (or thousands) of plot that are already stored in memory. In the speech recognition research literature, no work has been reported on Devnagari speech processing and numerals. So we consider our work to be the first such attempt in this direction. The process involves extraction of some distinct characteristics of individual words by utilizing Fourier transforms and their correlations. The system is speaker-independent and is moderately tolerant to background noise.

II. DEVNAGARI VOWELS

The 12-Devnagari vowels are categorised as per IPA (International Phonetics Association) as shown in Table-2. These are used for the speech analysis and synthesis purpose. It describes in different categories such as follows:

A. Short Vowels

The short vowel is a single vowel (V) in a short word or syllable, that vowel usually makes a short sound. These short vowels usually appear at the beginning of the word or between two consonants. E.g. the short vowels represent character in Marathi and in Hindi.

B. Long Vowels

The long vowels a short word or syllable ends with a vowel-consonant (VC). The `a' at the end of the word is silent. Long vowels when the word or syllable has a single vowel and the vowel appears at the end of the word or syllable, the vowels usually represent makes the long sound in Hindi.

C. Conjunct Vowels

The conjunct vowels are combination of short and long vowels. These phonemes are produced in Hindi e.g. as shown in Table-2.

D. Nasal Vowel

A nasal vowel is produced with a low tune so that air pressure through nose as well as mouth. The term "nasal" is slightly air pressure which does not come exclusively out of the nose in nasal vowels.

E. Visarg Vowel

The Visarg symbol is used rarely in Devnagari. The visarg is pronounced as the voiceless sound after the vowels. E.g.in Hindi.

TABLE I. RANGE OF HUMAN SPEECH

Gender	Fundamental frequency (F0) Min Hz	Fundamental frequency (F0)Max Hz
Male	80	200
Female	150	350

TABLE II. DEVNAGARI VOWELS CLASSIFIED INTO FIVE TYPES

TYPE OF DEVNAGARI VOWELS	1	2	3	4
SHORT	अ	इ	उ	-
LONG	आ	ई	ऊ	-
CONJUN-CT	अ+इ=ए	अ+ई=ऐ	अ+उ=ओ	अ+उ=औ
NASAL	अं	-	-	-
VISARG	अः		-	-

TABLE III. HINDI CHARACTER SET

Vowels	अ आ इ ई उ ऊ ऋ ए ऐ ओ औ अं अः a ā i ī u ū r e ai o au aṅ ah
Gutturals (कवर्ग)	क ख ग घ ङ ka kha ga gha ṅa
Palatals (चवर्ग)	च छ ज झ ञ ca cha ja jha ña
Cerebrals (टवर्ग)	ट ठ ड ढ ण ṭa ṭha ḍa ḍha ṇa
Dentals (तवर्ग)	त थ द ध न ta tha da dha na
Labials (पवर्ग)	प फ ब भ म pa pha ba bha ma
Semi-Vowels	य र ल व ya ra la va
Sibilants	श ष स sa pa sa
Aspirate	ह Ha

III. Speech Modelling Using Average Energy In The Zerocrossing Interval

The speech production model suggests that the energy of the voiced speech is concentrated about 8 kHz, where as in the case of unvoiced speech, most of the energy is found at higher Frequencies. Since high frequency implies high zerocrossing rate and low frequency implies low zerocrossing rate, there is strong correlation between zerocrossing rate and energy distribution with frequency. This motivates us to model the speech signal using average energy in zerocrossing interval of the signal. Consider the speech segment shown in Figure 1. The ZC_i^k shows the ith zerocrossing and ZC_{i+1}^k shows the i+1th zerocrossing of kth observation window. The time interval between these two points is called ith zerocrossing interval T_i^k in the kth observation window.

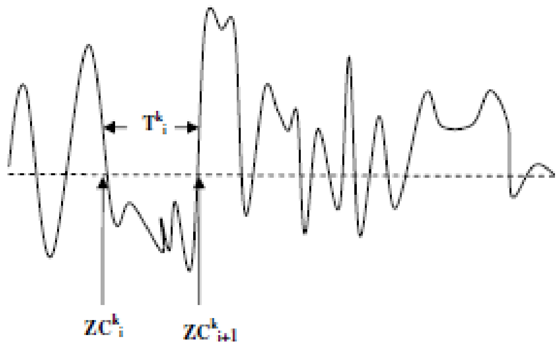

FIGURE 1. SPEECH SEGMENT IN KTH OBSERVATION

The average energy in the ith zerocrossing interval can be obtained by the expression:-

$$\mathbf{E_t^k = \frac{1}{T_t^k}\int_{Zc_i^k}^{Zc_{i+1}^k} X^2\ (t)dt}$$

E_i^k Is the average energy of the signal in $T_i^{k\ th}$ zerocrossing interval of *k*th observation window and *X(t)* is the instantaneous signal amplitude. The aim of the present study is to find a robust coefficient for speech recognition application using the average energy in the zerocrossing interval (AEZI). An XY plot is generated by plotting index number of zero crossing intervals along X axis and Average Energy in the Zerocrossing Interval (AEZI) along Y axis. Figure 1 represents the average energy in the zerocrossing interval *vs* index number of the zerocrossing interval for the Hindi script.

IV.DATA ACQUISITION AND PROCESSING

One of the obvious methods of speech data acquisition is to have a person speak into an audio device such as microphone or telephone. This act of speaking produces a sound pressure wave that forms an acoustic signal. The microphone or telephone receives the acoustic signal and converts it into an analog signal that can be understood by an electronic system. Finally, in order to store the analog signal on a computer, it must be converted to a digital signal.

The data in this paper is acquired by speaking Hindi Word and numeral into a microphone connected to Windows-7 based PC. The data is saved into '.wav' format files by the using of MATLAB. The sound files are processed after passing through a (Simulink) filter, and are saved for further analysis such as FFT. We recorded the data form speakers who spoke the same word set, i.e. Devnagari Script & numerals.

In general, the digitized speech waveform has a high dynamic range, and can suffer from additive noise. So first, a Simulink model was used to extract and analyze the acquired data; see Fig. 2.

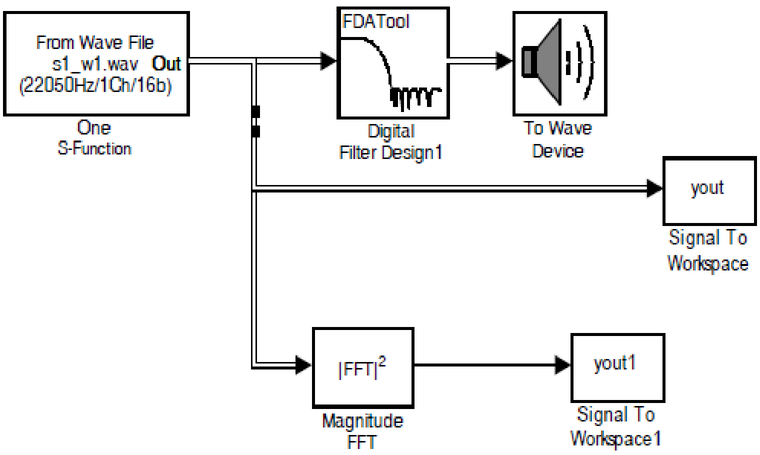

Figure2. Simulink Model For Analyzing Hindi Data And Numerals

The Simulink model, as shown in Fig. 2, was developed for performing analysis such as standard deviation, mean, autocorrelation, magnitude of FFT, data matrix correlation. We also tried a few other statistical techniques.

We would also like to mention that we had started our experiments by using Simulink, but soon found this GUI-based tool to be somewhat limited because we did not find it easy to create multiple models containing variations among them. This iterative and variable-nature of models eventually led us to MATLAB's (text-based) **.m** files. We created these files semi-automatically by using a Hindi-language script; the script was developed specifically for this purpose.

Three main data pre-processing steps were required before the data could be used for analysis.

A. Pre-emphasis

By pre-emphasis, we imply the application of a *normalization* technique, which is performed by dividing the speech data vector by its highest magnitude.

B. Data Length Adjustment

FFT execution time depends on exact number of the samples (*N*) in the data sequence $[x_K]$, and that the execution time is minimal and proportional to $N*log_2(N)$, where *N* is a power of two. Therefore, it is often useful to choose the data length equal to a power of two.

C. Endpoint Detection

The goal of endpoint detection is to isolate the word to be detected from the background noise. It is necessary to trim the word utterance to its tightest limits, in order to avoid errors in the modeling of subsequent utterances of the same word. As we can see from the upper part of Fig. 2, a threshold has been applied at both ends of the waveform. The front threshold is normalized to a value that all the spoken numbers trim to a maximum value. These values were obtained after observing the behaviour of the waveform and noise in a particular environment. We can see the difference in frequency characteristics of the words.

D. Fourier Transform

The MATLAB algorithm for the two dimensional FFT routine is as follows:

fft2(x) =fft (fft (x));

Thus the two dimensional FFT is computed by first computing the FFT of x, that is, the FFT of each column of x, and then computing the FFT of each row of the result. Note that as the application of *fft2* command produced even symmetric data, we only show the lower half of the frequency spectrum in our graphs.

E. Correlation

Calculations for correlation coefficients of different speakers were performed. As expected, the cross-correlation of the same speaker for the same word did come out to be 1. The correlation matrix of a spoken number was generated in a three-dimensional form for generating different simulations and graphs.

V. Related Work

This section of paper We will represent the works such as implement an experimental, speaker dependent, real-time for the Hindi language (Devnagari Script).Words using the Dynamic Time Warp (DTW) technique. The presented work emphasized on template-based recognizer approach using linear predictive coding with dynamic programming computation based recognizers in isolated task.

A. *Standard MFCC*

Mel cepstral feature extraction is used in some form or another in virtually every state of the art speech and Frequency analysis system. First, speech samples are divided into overlapping frames. The usual frame length is 25 ms and the frame rate is 10 ms. each frame is usually processed by pre-emphasis filter to amplify higher frequencies. In the next step count the voiced samples and then take the Fourier spectrum is computed for the signal. A Mel spaced bank of filters is then applied to obtain a vector of log energies. Usually 20 to 40 filters are used depending on application. The output of the filter-bank is then converted to cepstral coefficients by using discrete cosine transform (DCT), where only the first 12 coefficients are retained for computing the feature vector. Finally the feature vector consists of 39 values including the 12 cepstral coefficients with one energy.

B. Extended MFCC

Thirteen extra triple delta features are added in standard 39 MFCC features forming a feature vector of 52 values. These 52 values are then reduced to 39 by applying any feature reduction technique. These techniques are based on linear transformation schemes like principal component analysis (PCA), linear discriminate analysis (LDA) and Hetroscedastic linear discriminate analysis (HLDA). HLDA, first proposed by N. Kumar has been widely used for various feature combination techniques. It maximizes the likelihood of all the training data in the transformed space and each training sample contributes equally to the objective function. We have used HLDA for feature reduction and this procedure is named extended MFCC as shown in Figure 3.

C. Robust Features

In noisy environments when training and testing conditions are severely mismatched, these features cannot work well. Therefore, feature domain signal processing methods are applied to enhance the distorted speech. Spectral subtraction is widely used as a simple technique to reduce additive noise in the spectral domain, In order to eliminate the convolutive channel effect and noise distortion.

D. Gaussian Mixture HMM

In this method continuous density hidden Markov models are used to match the phonetic information of speech signal with the feature vectors derived at front end. Multivariate Gaussian mixtures are used to calculate the likelihood of observation vectors (i.e. spectral features). Representation of phonetic information, HMM topology and number of Gaussian mixtures are the key issues for the implementation of these statistical techniques.

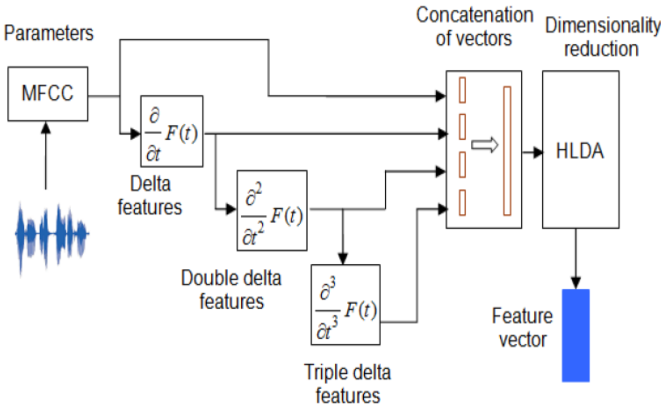

FIGURE 3. EXTENDED MFCC

VI.ANALYSIS & RESULTS

We observed that Fourier descriptor feature was independent for the spoken Devnagari Script and numerals with the combination of the Fourier transform and correlation technique commands used in MATLAB, a high accuracy recognition system can be realized. Recorded data was used in Simulink model for introductory analysis.

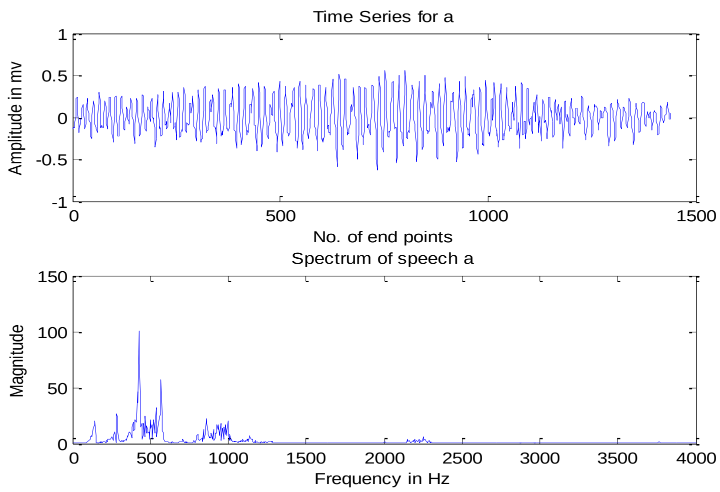

FIGURE 4. THE FFT WAVEFORM OF THE WORD अ IN DEVNAGARI SCRIPT

X = 1500,It's having 1500 numbers of data points. It's denoted by X. and having a 5 peaks values for each & every word same for अ in Devnagari script.

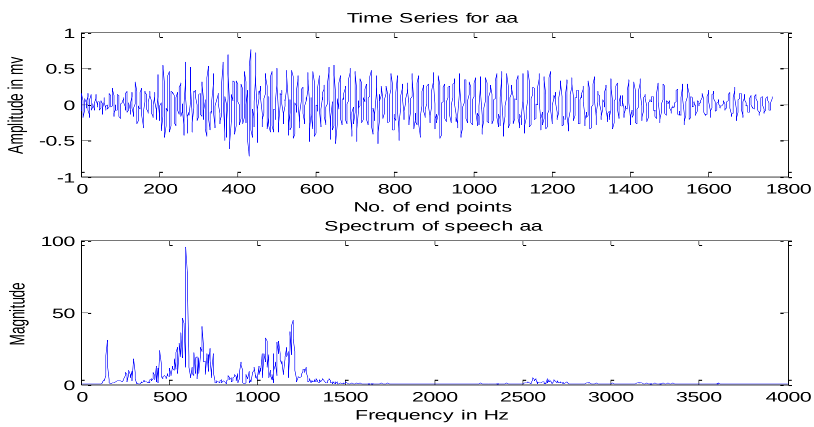

FIGURE 5. THE FFT WAVEFORM OF THE WORD आ IN DEVNAGARI SCRIPT

X = 1800,

It's having 1800 numbers of data points. It's denoted by X. and having a 5 peaks values for each & every word same for आ in Devnagari script.

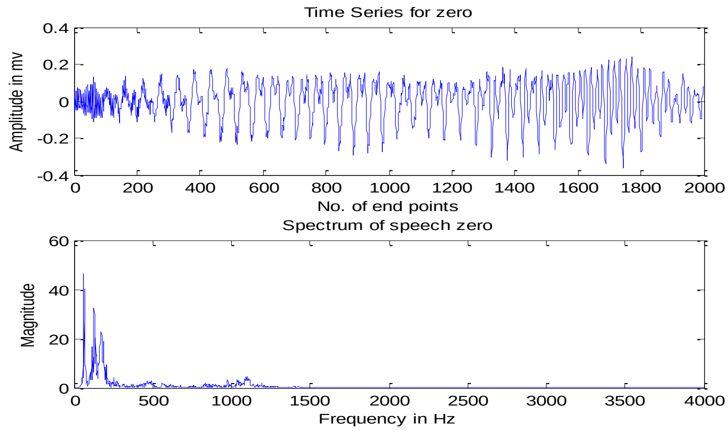

FIGURE 6. THE FFT WAVEFORM OF THE ZERO IN NUMERALS

It's having 2000 numbers of data points. It's denoted by X. and having a 5 peaks values for each & every word same for Zero in Numerals.

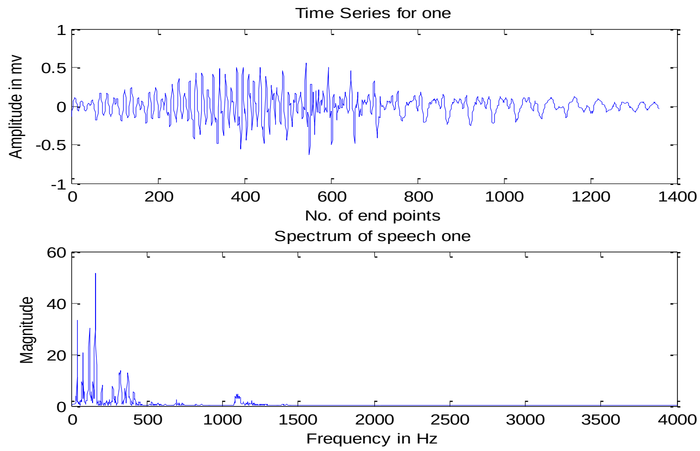

FIGURE 7. THE FFT WAVEFORM OF THE ONE IN NUMERALS

It's having 1400 numbers of data points. It's denoted by X. and having a 5 peaks values for each & every word same for One in Numerals.

VI. Conclusion And Future Work

In conclusion, an efficient, abstract and fast ASR system for regional languages like Hindi is need of the hour. The work implemented in the paper is a step towards the development of such type of systems. The work may further be extended to large vocabulary size and to continuous speech recognition. As shown in results, the system is sensitive to changing spoken methods and changing scenarios, so the accuracy of the system is a challenging area to work upon. Hence, various Speech enhancements and noise reduction techniques may be applied for making system more efficient, accurate and fast.

TABLE.IV. PEAKS AND ITS CORRESPONDING FREQUENCIES

SR.NO	*SPEECH WORD*	*PEAK*		*FREQUENCY IN (HZ)*	
1	*FOR A*	*P1*	*100.9565*	*F1*	*424*
		P2	*71.1906*	*F2*	*429*
		P3	*57.5883*	*F3*	*567*
		P4	*46.6103*	*F4*	*415*
		P5	*37.3028*	*F5*	*434*
2	*FOR AA*	*P1*	*95.4134*	*F1*	*596*
		P2	*77.7759*	*F2*	*601*
		P3	*70.3393*	*F3*	*605*
		P4	*46.3746*	*F4*	*578*
		P5	*44.5413*	*F5*	*120*
3	*FOR ZERO*	*P1*	*46.3055*	*F1*	*62*

		P2	*40.2583*	*F2*	*67*
		P3	*33.3799*	*F3*	*68*
		P4	*32.5196*	*F4*	*129*
		P5	*29.6611*	*F5*	*130*
4	*FOR ONE*	*P1*	*51.6031*	*F1*	*162*
		P2	*46.8251*	*F2*	*164*
		P3	*34.8045*	*F3*	*160*
		P4	*33.1000*	*F4*	*41*
		P5	*32.7313*	*F5*	*40*

References

[1]. S K Husain, Perez Akhter, “Digital Signal Processing, Theory and Worked Examples”, January 2007.

[2]. Samuel D Stearns, Ruth A David, “Signal Processing Algorithms in MATLAB,” Prentice Hall, 1996.

[3] S K Husain, Nighat Jamil, “Implementation of Digital Signal Processing real time Concepts Using Code Composer Studio 3.1, TI DSK TMS 320C6713 and DSP Simulink Blocksets,” IC-4 conference, Indian Navy Engineering College, Goa, Nov. 2007.

[4] M. Habibullah Pagarkar, Lakshmi Gopalakrishnan, et.al. “Language Independent Speech Compression using Devnagari Phonetics”, 2002.

[5] D. O’Shaughnessy, “Interacting with Computers by Voice-Automatic Speech Recognitions and Synthesis”, (Invited Paper), Proceedings of the IEEE, Vol. 91, No. 9, 2003, pp. 1272-1305.

