Asian Journal of Convergence in Technology *Volume II, Issue III*
Issn No.:2350-1146, I.F-2.71

Speech Signal Analysis: A frequency domain approach

Sharada V Chougule,
Department of Electronics and Telecommunication Engg.
Finolex Academy of Management and Technology, Ratnagiri
Maharashtra, India
shardavchougule@gmail.com

Mahesh S Chavan
Department of Electronics Engg.
KIT's College of Engg.Kolhapur
Maharashtra, India
maheshpiyu@gmail.com

Abstract **- Speech is the most natural source of communication for human beings. It can be produced instantaneously when required. When treated as a signal, it is found useful in number of speech related applications performed by machines such as speech recognition, speaker recognition and speech synthesis. The basic requirement using the speech for various applications is to analyze the speech signal and extract the useful characteristics from the same for the specific task. In this paper, frequency domain methods to analyze the speech signal are discussed along with their significance in specific applications. The properties or specific features that can be extracted from frequency domain analysis are also described with the help of mathematical analysis behind the same.**

Index Terms – Speech recognition, Speaker recognition, Speech synthesis

I. INTRODUCTION

Human speech is a multi-disciplinary area including communication, linguistics and computer science. Though occur naturally and easily, it is complicated in nature comprising variety of information in it. Along with conveying the actual message, it also gives the knowledge about the language, emotion and identity of the person indirectly. All this information is embedded in the speech signal in very complex way. Speech signal analysis is necessary to understand the nature and characteristics and is the integral part of any speech related application. Speech analysis can be performed through time domain as well frequency domain. The structure of speech production organs can be better examined, analyzed and modeled through spectral analysis than that of time domain parameters. In this paper various approaches towards the frequency domain analysis of speech signal are discussed along with their pros and cons. The remaining paper is organized as follows: Section II describes the structure of the human speech production organs. In section III, source-filter model of speech production is discussed. Section IV and V presents various approaches towards the frequency domain analysis of speech signal. The paper ends with the conclusion in section VI.

II. HUMAN SPEECH PRODUCTION MECHANISM

Speech is a sequence of sounds intended to convey some message to the listener. Speech production begins with the formation of ideas in speaker's mind which are represented in the form of words and sentences applying the rules of the language. Lungs, vocal folds and vocal tract are the three main groups of speech production organs. While speaking, the lungs act as an energy source which causes the air pressure to move through trachea (wind-pipe) towards the vocal folds. The tensed vocal folds within the larynx (complex system of cartilages, muscles and ligaments) are caused to vibrate due to air pressure (according to Bernoulli oscillation). Vocal folds chops the air to create quasi-periodic pulses (during voiced sounds), which are "spectrally shaped" by the vocal tract. The vocal tract starts from the larynx and end at the lips and the nasal cavity. The position of various articulators such as jaw, tongue, lips determine the type of sound to be produced. Fig.1 shows the elements of human speech production organs and the nature of spectral contents at each stage.

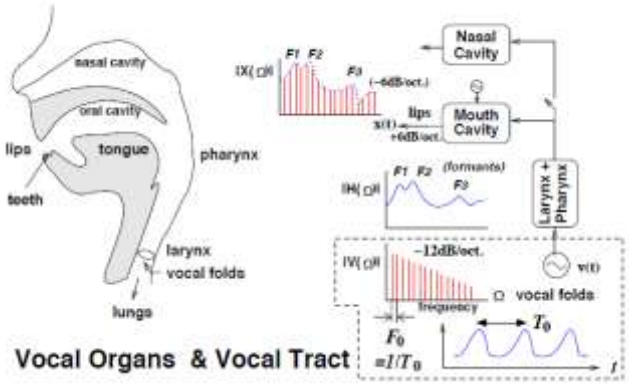

Fig.1. Speech production organs and subsequent spectrum analogy [2]

The airflow velocity at the vocal folds results in periodic oscillations representing time-varying area of the glottis (slit like orifice between the two folds). The time duration of one glottal cycle (T_0) is referred to as *pitch period* and reciprocal of pitch period is called as *fundamental frequency* (F_0). Such periodic oscillations are generally observed during vowel sounds, which may have one to four pitch periods over the duration of sound [1]. The function of vocal tract is to generate perceptually distinct sounds by varying the position of various articulators in oral and nasal path. This is analogous to creating resonances of different frequencies. These resonances are observed as peaks in the spectrum and are usually called as *formant frequencies* or simply *formants* (e.g. F_1, F_2 ,...). Thus location of formants is one of the important characteristics of speech sounds. Practically there is some

variation (in certain range) in the location of formants due to difference in structure of speech production system from person to person. Thus formants can be characteristics of speech ('what is spoken') as well as of the speaker ('who is speaking').

III. Digital Model of Speech Production

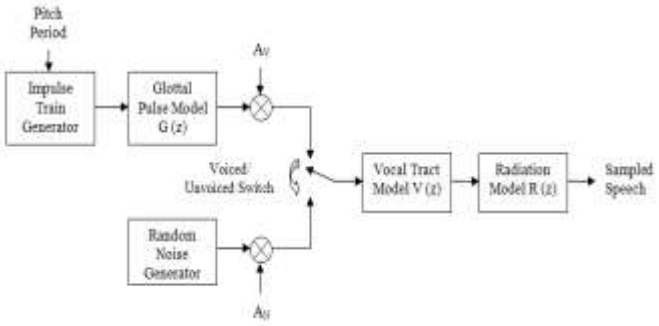

Fig. 2.Digital Model of Speech Production

The process of generating the speech from human speech production system can be characterized in the form of digital model. Each element of human speech production system is represented as a part of linear time varying system. Voiced sounds are periodic in nature. A periodic train of impulses is generated from the impulse train generator with frequency equal to the fundamental frequency (representing the frequency of glottal pulses). Natural glottal pulse waveform could be replaced by a synthetic pulse waveform of the form [3]:

$$g(n) = \frac{1}{2}\left[1 - cos\left(\frac{\pi n}{N_1}\right)\right] \qquad 0 \le n \le N_1$$

$$= cos\left(\frac{\pi(n - N_1)}{2N_2}\right) \qquad N_1 \le n \le N_2$$

$$= 0 \qquad otherwise \qquad (1)$$

The nature of resultant glottal flow waveform is shown in Fig.2. As the sequence g(n) has finite length, its z-transform G(z) is all-zero system, creating a low-pass filtering effect in frequency domain. Random noise generator provides the excitation for unvoiced sounds. A_V and A_N are the gain parameters.

Considering speech as a quasi-stationary signal, the relation between glottal airflow velocity input and vocal tract airflow velocity output can be approximated as a linear filter, whose characteristics depends on the nature (shape and size) of vocal tract. Under ideal condition, the vocal tract can be modeled as a concatenation of lossless tubes of different areas and lengths. The resonances (formants) of these tubes are created due to different vocal tract configuration during speaking. These resonances are observed as peaks in the spectrum, which can be modeled by an all-pole system V(z), in which each pair of complex conjugate pole corresponds to the respective formant. As observed in Fig. 3, there are three formants approximately at 750 Hz, 2000 Hz and 2900 Hz.

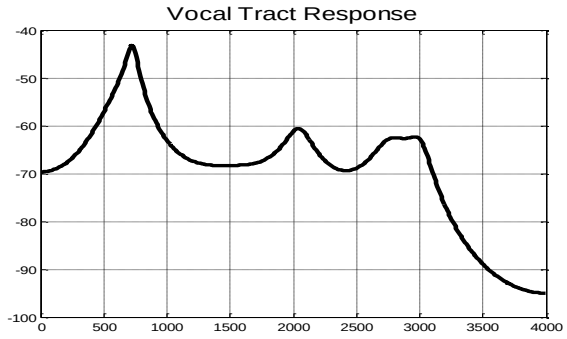

Fig.3. Resonances of the vocal tract

The generalized form of vocal tract transfer as an all-pole system is given by:

$$V(z) = \frac{A}{\prod_{k=1}^{N}(1 - c_k z^{-1})\,(1 - c_k^* z^{-1})} \qquad (2)$$

The lossless tube model ignores all the losses except the losses at the glottis and lips. Out of these two, the effect of pressure at the lips is taken as radiation loss, where the pressure is related to the volume velocity equivalent to high-pass filtering operation. The radiation model can then be a first order system (all-zero) characterized by:

$$R(z) = R_0(1 - z^{-1}) \qquad (3)$$

In most of the cases, the radiation model and vocal tract model are combined to form a single system.

The discrete time model thus formed is useful to analyze the speech signal in order to investigate the characteristics of excitation model (such as pitch period, voiced/unvoiced classification) and vocal tract mode (e.g. formants) imparting spectral shapes or peaks.

IV. Frequency Domain Analysis of Speech Signals

Frequency domain analysis of speech signal is mostly performed using Fourier analysis, power spectrum, spectral envelop detection and speech spectrogram. Taking Fourier transform of the entire speech signal provides only gross information about the frequency components present in the signal without giving any timing information (when a particular frequency component is present). Short-Time Fourier Transform (STFT) gives better representation of time-varying frequency components.

A. Short-Time Fourier Transform (STFT)

Speech is slowly varying signal assumed to be quasi-stationary i.e. when observed over a short time interval of 10-20 ms its characteristics are almost remain relatively constant.

This leads to *short-time analysis*, in which speech signal is divided into short segments called *frames* using a convenient window function (mostly tapered window such as Hamming window). Adjacent windows are overlapped (30-50% overlap) to avoid spectral leakage. The STFT of windowed frame is given as:

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m} \qquad (4)$$

Where $X_n(e^{j\omega})$ gives the short time spectrum of speech signal reflecting the time varying properties of the speech signal and $w(n-m)$ is the shifted window sequence, which slides over the entire speech signal $x(m)$. T

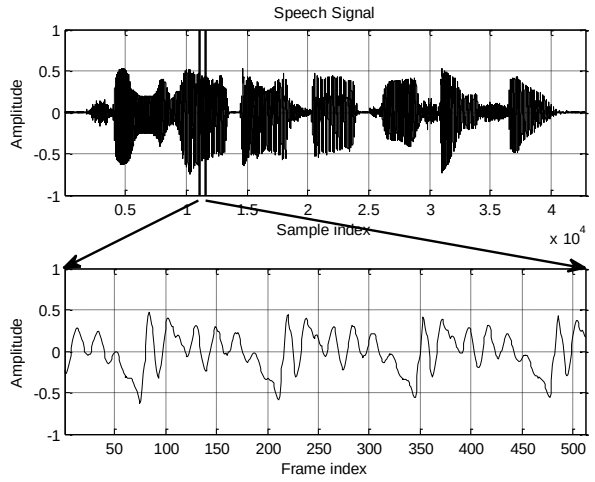

Fig.4. Speech signal showing voiced and unvoiced frame

As shown in Fig.4, the details of information in short frames of speech signal varies depending upon nature of speech sounds.

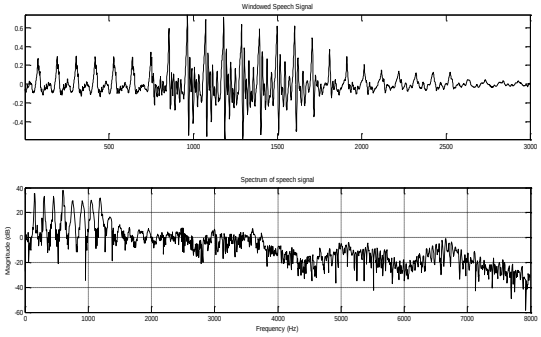

Fig.5 Windowed speech frame and its spectrum

The power spectrum of the speech signal is also the useful analysis technique, especially for obtaining peaks in the spectrum. As discussed in section III, these peaks represent the formants of the speech sound, which is the characteristic of vocal tract. There should be a proper compromise between window length and spectral details. Selecting a smaller window (3-5ms) may give rise to poor spectral resolution, whereas spectrum may suffer from timing resolution using a larger window (100-300 ms). As observed in Fig. 5 and Fig.6, a proper window size will explore both the low frequency details (envelop of the spectrum) characterizing the vocal tract and high frequency detail like pitch and its harmonics relating the excitation source. Shape of window function (except rectangular window) does not affect the characteristics much. Hamming window is good choice considering main lobe width and peak side lobe amplitude.

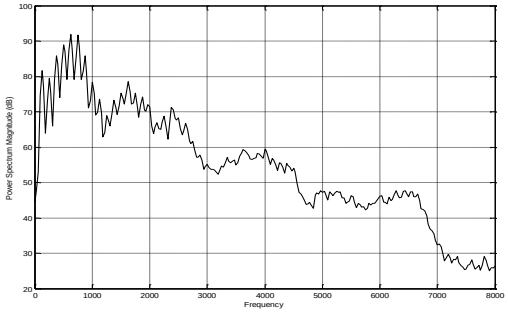

Fig.6. Power spectral density of speech signal of word '*had*'

B. Spectrogram Analysis

Another approach to obtain the properties of speech signal is through Spectrogram. Spectrogram (also called as Sonogram) is similar in principle to spectrum of obtained using Fourier analysis, but the difference is along with frequency it takes in account the time factor simultaneously. It is graphical display of the magnitude of the time-varying spectral characteristics [3]. It also follows the framing and windowing before analyzing the spectrum. It is thus a 3-D view of the spectrum with time displayed on horizontal axis and frequency on vertical axis. The magnitude of frequency components (energy) with respect to time is observed as degree of darkness (more the energy, more the darkness).

There are two types of spectrograms depending upon the size (length) of window function used for framing the signal, namely: Wide-band spectrogram and Narrowband spectrogram. The general form of spectrogram of windowed speech waveform is expressed as:

$$S(\omega,\tau) = \frac{1}{N^2}\left|\sum_{k=-\infty}^{\infty} \tilde{H}(\omega_k)\, W(\omega-\omega_k,\tau)\right|^2 \qquad (5)$$

where $\tilde{H}(\omega) = H(\omega)G(\omega)$

In equation (5), $W(\omega)$ is the spectrum of shifted window function, $\tilde{H}(\omega)$ is the multiplicative spectral component of glottal flow input *g(n)* and time varying system impulse response *h(n)* respectively.

In wide-band spectrogram, spectral analysis is performed on a small segment of windowed speech around 10 ms, giving

broader bandwidth for analysis. This gives rise to better resolution of individual pitch periods and voiced regions, observed as vertical striations in the graphical display. In the counterpart, narrowband spectrogram uses larger window around 50 ms, having a narrow-band for analysis. Narrow-band spectrogram thus better resolve the individual pitch harmonics and gives horizontal striations representing prominent formants. Spectrogram is one of the reliable to estimate the pitch and formants of speech signal.

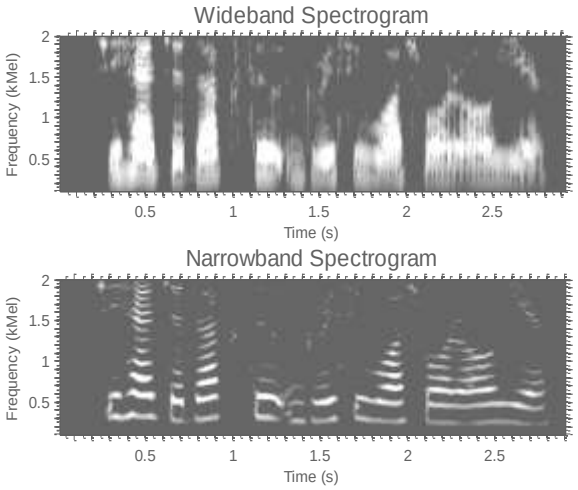

Fig.5. Inverted Wideband and Narrowband Spectrogram

V. Cepstral Analysis

From the source-filter theory of speech signal modelling [4] the speech is considered as the response of linear time-varying system, where excitation is created by vocal folds vibration and vocal tract determines the characteristics of the system. In time domain, speech is represented as convolution operation of these two entities. The objective of *cepstral analysis* is to separate the speech into its source and system components without any a priori knowledge about source and / or system [5]. Convolution operation is transformed as multiplication of in frequency domain their respective spectra. If $e(n)$ is the excitation, $h(n)$ is the impulse response of the system , then speech signal $s(n)$ in time and frequency domain is given by:

$$s(n) = e(n) * h(n) \quad \text{and}$$
$$S(\omega) = E(\omega)H(\omega) \qquad (6)$$

Taking logarithm of magnitude spectrum gives:

$$\log|S(\omega)| = \log|E(\omega)| + \log\ |H(\omega)| \qquad (7)$$

Here the magnitude spectrum of excitation component and vocal tract component are observed as linearly separable one. Inverse DFT of log spectra transforms the spectra from frequency domain to quefrency domain, also referred as *cepstral* domain. In the quefrency domain the vocal tract components are represented by the slowly varying components concentrated near the lower quefrency region and excitation components are represented by the fast varying components at the higher quefrency region. Thus the *cepstrum* of the speech signal is IDFT of the log spectrum of magnitude spectrum of the speech signal given by [5]:

$$c(n) = IDFT(log|S(\omega)|) \qquad (8)$$
$$= IDFT(log|E(\omega)|) + IDFT(log|H(\omega)|)$$

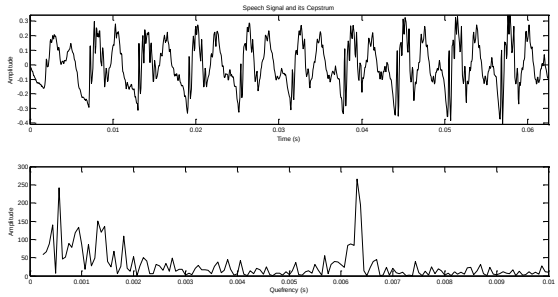

Fig.6. Cepstrum of voiced segment of speech signal

The cepstrum projects all the slowly varying components in log magnitude spectrum to the low frequency region and fast varying components to the high frequency regions. In the log magnitude spectrum, the slowly varying components represent the envelope corresponds to the vocal tract and the fast varying components to the excitation source. As a result the vocal tract and excitation source components get represented naturally in the spectrum of speech.

VI. Conclusion

In this paper, methods for speech analysis in frequency domain are discussed. Short time analysis is essential for speech signal to derive important characteristics of the speech signal. All these methods are based on source-system modeling of speech signal. The characteristics of vocal fold are represented in terms of pitch period or fundamental frequency, glottal flow waveform etc. Resonances of the spectrum or formants are the most prominent features of the vocal tract. All of these features are well distinguished through frequency domain analysis. The usefulness of the extracted features depends upon the end task to be performed. Spectrographic display is useful speech analysis tool, to understand or investigate frequency components of speech sounds with respect to time. Cepstral analysis is a convenient technique for separating source and system parameters of linear speech production system. Variability in same speech sounds is the preferred characteristic in case of identifying an individual from one's voice whereas uniqueness of speech sounds amongst number of speakers is the desirable characteristics for speech recognition. Frequency domain analysis is useful for variety of other applications like

language recognition, emotion recognition, speaker diarization/index.

REFERENCES

[1] Thomas F. Quatieri, Discrete time speech signal processing, Principles and Practice, Pearson Education, 2002.
[2] Hiroshi Shimodaira and Steve Renals: Automatic speech recognition-Lectures Series.
[3] A.E. Rosenberg," Effect of glottal pulse shape on the quality of natural vowels" Journal of Acoustic Soc. Am.Vol.49,No.2, pp.583-590, February 1971.
[4] Lawrence R. Rabiner and Ronald W. Schafer, Digital processing of speech signals, Prentice Hall Innternational,1978J.
[5] *Alan V. Oppenheim and Ronald W. Schafer, "From Frequency to Quefrency: A History of the Cepstrum", IEEE Signal Processing* Magazine, pp.95-110,September 2004.