2.1 Speech Signal and Characteristics

Speech is a pressure waveform that travels from a speaking person to one or more listeners. This signal is typically measured directly in front of the speaker’s mouth, which is the primary output location for the speech. Since the ambient atmosphere in which one speaks imposes a basic pressure, it is actually the variation in pressure caused by the speaker that constitutes the speech signal. The signal is continuous in nature and is very dynamic in time and amplitude, corresponding to the constantly changing status of the vocal tract and vocal cords. Speech can be characterized as a discrete sequence of sound segments called phones, each having certain acoustic and articulatory properties during its brief period of time. Each phoneme imposes certain constraints on the positions for these vocal tract articulators or organs: vocal folds or vocal cords, tongue, lips, teeth, velum, and jaw. Speech sounds fall into two broad classes: (a) Vowels which are responsible to allow unrestricted airflow throughout the vocal tract, and (b) Consonants which control airflow at some point and have weaker force than the vowels.

2.1.1 Articulatory Phonetics and Speech Generation

Speech is generated as one exhales air from the lungs while the articulators move. Sound production process can be recognized as a filtering process in which the vocal tract filter is excited by a speech sound filter. The source either is noisy aperiodic, causing voiced speech, or is periodic, causing unvoiced speech [9]. The source of the periodicity for the former is found in the larynx where vibrating vocal cords interrupt the airflow from the lungs, producing pulses of air. Both the area of the glottal opening between the vocal cords and the volume of the pulse of air can be approximated as half rectified sine waves in time, except that the glottal closure is more abrupt than its opening gesture. This
asymmetry assists the speech signal to be more intense than might otherwise be the case, because abrupt changes in the excitation increase the bandwidth of the resulting speech. Layout of human speech production mechanism is shown in Figure 2.1.

2.1.2 Anatomy and Physiology Speech Generation
The human speech production organs are multi purpose: speech generation, breathing, eating and sensing odors. Thus, in a communication sense, speech production cannot be as optimal an information source as the ear is a receiver. Certain parallels can be made between electronic and human speech communication. Humans minimize effort, in terms of their energy and time while maximizing the perceptual contrast in listeners.

2.1.3 Vocal Tract
The lungs provide the airflow and pressure source for speech, and the vocal cords usually modulate the airflow to create many sound variations. However, it is the vocal tract that is the most important system component in human speech production. It is a tube-like passageway made up of muscles and other tissues, and enables the production of the different sounds that constitute spoken language.

![Figure 2.1 Human speech production mechanism](image)
For most sounds, the vocal tract modifies the temporal and spectral distribution of power in the sound waves, which are initiated in the glottis. In addition, the vocal tract generates some sounds directly. It is the source for obstruent like stop and fricative sounds. Different phones primarily can be distinguished by their periodicity - voiced or unvoiced, spectral shape - mostly which frequencies have major power and duration - longer phones are perceived as having greater stress. The state of the vocal folds usually specifies each phone’s voicing feature choice [9]. By far the most important aspect of speech production is the specification of different phones via the filtering actions of the vocal tract. Because speech perception is dominated by the presence of sound power. The formants are often abbreviated F1, hence F1 means the formant with the lowest frequency. In voiced phones, the formants often decrease in power as a function of frequency due to the general lowpass nature glottal excitation, thus F1 is usually the strongest formant. Displacing the articulators changes the shape of the acoustic tube through which sound passes, and alters its frequency response. After leaving the larynx, air from the lungs passes through the pharyngeal and oral cavities and then exits at the lips. For nasal sounds, air is allowed to enter the nasal cavity by lowering the velum, at the boundary between the pharyngeal and oral cavities.

Figure 2.2 Schematized diagram of vocal apparatus
The velum is kept in a raised position for most speech sounds, blocking the nasal cavity from receiving air. During nasal sounds as well as during normal breathing, the velum lowers to allow air through the nostrils. In the speech production process, it is helpful to abstract the important features of the physical system in a manner which ultimately leads to a realistic yet tractable mathematical model. Figure 2.2 shows such a more physically realistic schematic diagram of the vocal system. For completeness the diagram includes the sub glottal system composed of the lungs, bronchi and trachea, a mechanical model of the vocal cords, including mass, spring and damping components and a variable area set of tubes that model the vocal tract configuration. The sub glottal system serves as a source of energy for the production of speech. The mechanical model of the vocal cords provides the excitation signal for the vocal tract. The resulting speech signal is simply the acoustic wave that is radiated from this system when air is expelled from the lungs and the resulting flow of air is shaped accordingly by the time varying vocal tract [9]. The vocal tract and nasal tract are shown in Figure 2.2 as tubes of nonuniform cross sectional area. As sound generated as discussed above, propagates down these tubes, the frequency spectrum is shaped by the frequency selectivity of the tube. This effect is very similar to the resonance effects observed. In the context of speech production, the resonance frequencies of the vocal tract tube are called formant frequencies or simply formants. The formant frequencies are depend upon the parameter like shape and dimensions of the vocal tract; every individual shape is characterized by a set of formant frequencies. Different sounds are generated by varying the shape of the vocal tract. Thus, the spectral properties of the speech signal vary with time as the vocal tract shape changes.

Vocal tract might be assumed as a sequence of cylinders, each having a variable cross-sectional area. In the vocal tract, the tongue, the lower teeth and the lips undergo significant movements during speech production. In contrast, the upper and rear boundaries of the vocal tract are relatively fixed but with diverse composition [10]. The nasal cavity consists of many passages lined with mucous tissue and has no movable structures. Its large interior surface area significantly attenuates speech energy. The opening between the nasal and pharyngeal cavities controls the amount of acoustic coupling between the cavities and hence the amount of energy leaving the nostrils.
Increased heat conduction and viscous losses cause formant bandwidths to be wider in nasals than for other sonorants. During speech production, if either the vocal tract or glottis is completely closed for a time, airflow ceases and typically no sound emerges. The class of phones called stops or plosives uses such closures lasting several tens of milliseconds. Immediately prior to closure and at closure release, stops have acoustics that vary depending on the closure point: in the glottis, in the vocal tract or the lips.

2.1.4 Larynx and Vocal Folds

The vocal folds are important for speech production because normal breathing creates little audible sound, as the air expelled by the lungs passes mostly unobstructed throughout the vocal tract. Sound is generated in the vocal tract only when the airflow path is narrowly or totally constricted, effectively interrupting the airflow, thus creating either turbulent noise or pulses of air. Most speech originates in the larynx, which consists of four cartilages: thyroid, cricoids, arytenoids, and epiglottis joined by ligaments and membranes [10]. The passageway between the lungs and the vocal tract is called the trachea which divides into two bronchial tubes towards the lungs. A nonspeech organ called the epiglottis protects the larynx from food and drink. The vocal folds inside the larynx are typically about 15 mm long, have a mass of about 1 gram each, and have amplitude vibrations of about 1 mm. When one breathes normally, the vocal folds are far enough apart to avoid sound creation, although increased airflow during exercise leads to loud whispers. If airflow is strong enough and the folds are close enough, such turbulent noise occurs at the glottis.

This speech is called whisper or aspiration and corresponds to the phoneme /h/. Similar noise can be generated in the same way higher up in the vocal tract at a narrow constriction either between the tongue and the palate (the roof of the mouth) or with the lips and teeth. The latter noises are called fricative sounds. The main difference between aspiration and frication lies in the range of frequencies: broadband for whisper and high-frequency only for frication, because each noise source excites mostly the portion of the vocal tract in front of the source constriction. In frication, shorter cavities correspond to higher frequency resonances than with aspiration. All such noise sounds are aperiodic.
due to the random nature of their turbulent source. Air leaving the lungs during sonorant phones is interrupted by the quasiperiodic closing and opening of the vocal folds. The rate of this vibration is called the fundamental frequency. It is often abbreviated F0, in contrast to the formants F1, F2,... although F0 is not a resonance power in sonorants appears primarily at harmonic multiples of F0 but these harmonics have widely varying intensity depending on the formants [10]. The fundamental period or pitch period of voiced speech, \( T_0 = 1/F_0 \), corresponds to the time between successive vocal fold closures. 

\( T_0 \) has a large dynamic range but its average value is proportional to the size of the vocal folds. To physically cause the vocal folds to vibrate, they must be close together and the lungs must generate sufficient pressure using the diaphragm so that the difference between the pressure below the glottis and that above is large enough. Voicing is unlike other rapid movements in speech which are due to voluntary muscle actions and are thus limited to low frequencies.

The increased velocity in the narrow glottis causes local pressure to drop. When it is low enough, the negative pressure forces the vocal folds to close which interrupts the airflow. Then a positive pressure develops in the larynx, forcing the vocal folds open again and the cycle repeats until the diaphragm or vocal cord muscles relax. Major and secondary articulators as noted above, vocal tract organs that move to produce speech sounds are called articulators. The tongue and the lips are the major most important articulators. In secondary roles, the velum and larynx also contribute to speech production. Through its varying glottal opening, the larynx controls airflow into the vocal tract. The larynx can also be raised or lowered which alters the tract length and which in turn raises or lowers respectively formant frequencies.

The jaw may be thought of as a secondary articulator because it helps to position the tongue and lips for most sounds. The lips are a pair of muscular folds that can cause a vocal tract closure or produce a narrow slit at the mouth. The lips can also either round and protrude or spread and retract only the four front upper teeth participate actively in speech production. Immediately behind the teeth is the hard palate the upper wall of the oral tract. Many phones require a constriction between the hard palate and the tongue.
The most important primary articulator is the tongue which has four components: tip, blade, dorsum and root. Owing to the relative rigidity of the upper and rear walls of the vocal tract, speakers rely on the very flexible tongue to provide the mechanism to create the different vocal tract shapes needed to produce the various speech sounds. The tip is fast and agile able to move up and then down within about 100 ms. The dorsum is the surface of the tongue whose frontal portion is the blade, the tongue body positions the dorsum. Most articulators move towards different target positions for each successive phone, starting as ballistic motion and then becoming more focused as the target nears or as other muscle commands are issued for a new target. The peak velocity of an articulator is often linearly related to its actual displacement. In generation of speech always combination of voiced and unvoiced sound is present which is modeled in Figure 2.3.

2.2 Major Features of Speech Articulation

Most languages, including english, can be described in terms of a set of distinctive sounds or phonemes [10]. In particular, for american english, there are somewhere between 39 and 48 phonemes including vowels, diphthongs, semivowels and consonants. Total 48 phonemes of american english along are there with the International Phonetic. It can be seen that the 48 phonemes are divided into 5 broad classes namely,

- vowels and diphthongs are 18
- vowel-like consonants are 4
- standard consonants are 21
- syllabic sounds are 4
- glottal stop is 1

![Figure 2.3 Generation of speech for voiced and unvoiced sound](image-url)
Rather than using the full set of 48 phonemes, it is convenient to work with a reduced set of 39 sounds [10]. Figure 2.4 summaries phonemes in detail.

- vowels are 11
- diphthongs are 4
- semi vowels are 4
- nasal consonants are 3
- voiced and unvoiced stop consonants are 6
- voiced and unvoiced fricatives are 8
- affricate consonants are 2
- whispered sound is 1

Figure 2.4 Phonemes in American English
Main interest in the production of phonemes is the following: the state of the vocal cords means checking about vibration, the degree of any major constriction in the vocal tract, and the location of such constrictions. These correspond to the features of voicing, manner of articulation, and place of articulation, respectively.

Manner of articulation deals with airflow in the vocal tract: whether it flows through the oral and/or nasal cavities and the degree of any major vocal tract constrictions. The manner classes are: vowels including diphthongs, glides, liquids, nasals, fricatives, and stops.

The vowels are the most important and largest class of phonemes; air passes relatively freely at rates of 100–200 cc/s through the pharynx and oral cavities. In some languages, nasalized vowels also allow air through the nasal cavity as well. Vowels have no constrictions narrow enough to cause frication noise turbulent and random airflow or to block the airflow completely [10]. To avoid noise generation, the area of minimum constriction except for the glottis exceeds 0.3 cms.

Glides also called semivowels resemble vowels but have a very high tongue position which causes a narrow vocal tract constriction barely wide enough to avoid frication. In many languages, there is a glide that closely resembles each high vowel in the language (e.g., glide /y/ resembles vowel /iy/ and /w/ resembles /uw/). In practice, glides are simply very high vowels that are difficult to maintain for more than a few tens of milliseconds; thus they may be thought of as transient vowels - vowels, on the other hand, can easily last for hundreds of ms, if desired.

Liquids also resemble vowels, except for use of part of the tongue as a major obstruction in the oral tract, which causes air to deflect from a simple path. For the liquid /l/ also called a lateral, the tongue tip is in contact with the alveolar ridge and causes a division of the airflow into two streams on both sides of the tongue.
Nasal consonants have a lowered velum, which allows airflow into the nasal cavity and through the nostrils while the oral tract is completely closed. Some languages have nasalized vowels, where air flows through both the oral and nasal cavities. Such nasalization lowering the velum often occurs in parts of English vowels but such sounds are not associated with a separate phonemic category because English listeners interpret these sounds as normal free variation when vowels occur adjacent to nasal consonants. Thus the distinction is ‘allophonic’ and not phonemic [10].

All phonemes in the preceding four classes (vowel, glide, liquid and nasal) are part of a more general manner class called sonorant. They are all voiced and relatively strong in power. The other general class of phonemes is called obstruent and is comprised of stops and fricatives, which are noisy and relatively weak, with the primary acoustic excitation at a major vocal tract constriction.

Stops or plosives employ a complete closure in the oral tract, which is then released. Air continues to flow through the glottis throughout a stop. The velum is raised throughout a stop to prevent nasal airflow during oral closure. Pressure builds up behind the oral closure and is then abruptly released. Air flows through the increasing orifice (at a decreasing speed as the opening grows over a few tens of ms). The initial intense burst of noise upon oral tract opening is called an explosion and is effectively a brief fricative. Prior to actual periodicity, an interval of noisy aspiration typically occurs with a duration called the voice onset time or VOT. The initial portion of the VOT is frication, produced at the opening constriction, whereas the longer remainder is aspiration noise, created at the constricting glottis.

In voiced stops, on the other hand, vocal folds may continue to vibrate throughout the stop or start to vibrate right after the burst. One difference between voiced and unvoiced stops is that the vocal folds are more widely separated during the vocal tract closure for unvoiced stops and start to adduct only at the release, hence the longer VOT for unvoiced stops.
Like other phones, stops usually last on the order of 80 ms. Unlike other sounds, stops are sometimes very brief when they occur between two vowels. An alveolar stop followed by an unstressed vowel in the same word often becomes a flap, where the tongue tip maintains contact with the palate for as little as 10 ms. While stops have a complete occlusion in the vocal tract, fricatives use a narrow constriction instead. To generate noise, fricatives need sufficient pressure behind the constriction with a narrow passage. This causes sufficiently rapid airflow to generate turbulence at the end of the constriction. Most speech sounds vowels, liquids, nasals and fricatives each have a specific articulatory position and can be maintained over several seconds, if desired. Stops, on the other hand, are transient or dynamic consonants which have a sequence of articulatory events.

Glides may also be considered as transient phonemes, owing to the difficulty of sustaining them. These sounds are actually phoneme sequences: each diphthong consists of a vowel followed by a glide, and each affricate consists of a stop followed by a fricative. There are phonological conventions for considering these as individual phonemes, owing to distributional restrictions and durational phenomena.

### 2.3 Properties and Characteristics of Speech Signal

Speech signal have following inherent property:

- In a linear manner it can be noticed that speech is a sequence of continually changing sounds
- The properties of the speech signal are highly dependent on the sounds that are produced
- The properties of the speech signal are highly dependent on the context in which the sounds are produced. It implies that way in which the sounds which generally take place before and after the current sound. This consequence is called speech sound co-articulation and it is the result of the vocal mechanism anticipating following sounds while producing the current sound, thereby changing the sound properties of the current sound. Some of the parameter of vocal cords like the positions, shapes and sizes of the various articulators like teeth, lips, tongue, jaw,
velum etc. All alter slowly over time, thereby producing the required speech sounds.

2.3.1 Time and Frequency Domain Characteristics of Speech
Analyzing speech in the time domain often requires simple calculation and interpretation. Among the relevant features found readily in temporal analysis are waveform statistics, power and F0. The frequency domain, on the other hand, provides the mechanisms to obtain the most useful parameters in speech analysis [11]. Most models of speech production assume a noisy or periodic waveform exciting a vocal-tract filter. The excitation and filter can be described in either the time or frequency domain, but they are often more consistently and easily handled spectrally.

2.3.2 Waveforms
Time-domain speech signals are also called speech waveforms. They show the acoustic signals or sounds radiated as pressure variations from the lips while articulating linguistically meaningful information. The amplitude of the speech waveform varies with time in a complicated way, including variations in the global level or intensity of the sound. The probability density function of waveform amplitudes, over a long time average, can be measured on a scale of speech level expressed as sound dB. This function has a form close to a double sided (symmetric) exponential at high amplitudes and is close to Gaussian at low amplitudes. The entire probability density function can be approximated by a sum of exponential and Gaussian functions.

2.3.3 Fundamental Frequency
Under detailed examination, a speech waveform can be typically divided into two categories.

- A quasi-periodic part which tends to be repetitive over a brief time interval.
- A noise-like part which is of random shape.

For the quasi-periodic portion of the speech waveform, the average period is called a fundamental period or pitch period. Its inverse is called the fundamental frequency or
pitch frequency, and is abbreviated F0. The fundamental frequency corresponds to vocal cord vibrations for vocalic sounds of speech. F0 in a natural speech waveform usually varies slowly with time. It can be 80 Hz or lower for male adults and above 300 Hz for children and some female adults. F0 is the main acoustic cue for intonation and stress in speech, and is crucial in tone languages for phoneme identification.

2.3.4 Overall Power

The overall power of the speech signal corresponds to the effective sound level of the speech waveform averaged over a long-time interval [12]. In a quiet environment, the average power of male and female speech waveforms measured at 1 cm in front of a speaker’s lips is about 58 dB. Male speech is on average about 4.5 dB louder (greater power) than female speech. Under noisy conditions, one’s speech power tends to be greater than in a quiet environment. Further, not only the overall power and amplitude is increased, but also the details of the waveform changes in a complicated way. In noisy environments a speaker tends to exaggerate articulation in order to enhance the listener’s understanding, thereby changing the spectrum and associated waveform of the speech signal.

2.3.5 Overall Frequency Spectrum

While the spectral contents of speech change over time, if the discrete-time Fourier transform (DFT) of the speech is taken in waveform over a long-time interval, it can be estimate the overall frequency range that covers the principal portion of the speech power. Such information is important for the design of speech transmission systems since the bandwidth of the systems depends on the overall speech spectrum rather than on the instantaneous speech spectrum. When such an overall frequency spectrum of speech is measured in a quiet environment, it is found that the speech power is concentrated mainly at low frequencies. Over 80% of speech power lies below 1 kHz. Beyond 1 kHz, the overall frequency spectrum decays at a rate of about \(-12\) dB per octave. Above 8 kHz, the speech power is negligible. If the long time fourier transform analyzes only a quasi-periodic portion of the speech waveform, it can be seen that the frequency components in
harmonic relations and integer multiples of a common frequency. This common frequency also called the lowest harmonic component, is the pitch.

2.3.6 Short Time Energy
Speech is dynamic or time varying. Sometimes, both the vocal tract shape and pertinent aspects of its excitation may stay fairly constant for dozens of pitch periods. On the other hand, successive pitch periods may change so much that their name ‘period’ is a misnomer. Since the typical phone averages only about 80 ms in duration, dynamic coarticulation changes are more [12]. Figure 2.5 shows energy plot of utterance ‘what she said; using windowing function.

![Image of energy plots](image)

Figure 2.5 Energy plots of different values

The short time energy of the speech signal gives a convenient representation that reflects these amplitude variations. In general short time energy can be represented as

$En = \sum_{m=-\infty}^{\infty} [x(m)w(n - m)]^2$  \hspace{1cm} (2.1)

This turn of phrase can be written as
Where \( h(n) = w^2(n) \). The selection of the impulse response \( h(n) \) or evenly the window determines the nature of the short time energy representation. In any event, speech analysis usually presumes that signal properties change relatively slowly over time. This is most valid for short time intervals of a few periods at most. During such a short-time window of speech, one extracts parameters or features, which each represent an average over the duration of the time window. As a result of the dynamic nature of speech, it is necessary to divide the signal into many successive windows or analysis frames, allowing the parameters to be calculated frequently enough to model dynamic vocal-tract features. Window size is critical to good modeling. Long vowels may allow window lengths up to 100 ms with minimal loss of detail due to the averaging, but stop explosions require much shorter windows to avoid excess averaging of rapid spectral transitions. In a compromise, typical windows last about 20-30 ms, since one does not know a priori what sound one is analyzing. Figure 2.6 describes how for whole speech waveform window shifts to cover total time in predefined window size.

Windowing means multiplication of a speech signal \( s(n) \) by a window \( w(n) \) which yields a set of speech samples \( x(n) \) weighted by the shape and duration of the window. By successive shifting of \( w(n) \) in time, with spacing corresponding to a suitable update.

\[
E_n = \sum_{m=-\infty}^{\infty} [x^2(m)h(n-m)]
\]

(2.2)

Figure 2.6 Speech signal with three superimposed windows, offset from the time origin by \( 2N \), \( 3N \) and \( 4N \) samples.
(frame) interval (often 10 ms), it may be examined that any or all of \( s(n) \) via the sliding window. Typically \( w(n) \) is a smooth function, so as to properly consider the samples of \( s(n) \) under evaluation. The simplest window has a rectangular shape \( r(n) \) [13].

![Common time window with duration normalized to unity](image)

**Figure 2.7 Common time window with duration normalized to unity**

\[
w(n) = r(n) = \begin{cases} 1 & \text{for } 0 \leq n \leq N - 1 \\ 0 & \text{otherwise} \end{cases} \quad (2.3)
\]

This gives equal weight to all samples of \( s(n) \), and limits the analysis range to \( N \) consecutive samples. A common alternative to above mathematical step is the Hamming window as shown in Figure 2.7, a raised cosine pulse:

\[
w(n) = h(n) = \begin{cases} 0.54 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right) & \text{for } 0 \leq n \leq N - 1 \\ 0 & \text{otherwise} \end{cases} \quad (2.4)
\]

or the quite similar Hanning window [14]. Tapering the edges of \( w(n) \) allows its periodic shifting at the update frame rate along \( s(n) \) without having large effects on the resulting speech parameters due to pitch period boundaries.
2.3.7 Spectrogram

A spectrogram is a time-varying spectral representation for forming an image that shows how the spectral density of a signal varies with time. Spectrograms are used to recognize phonetic sounds [15].

The spectrogram is arrangement of graph with two geometric dimensions: the horizontal axis indicates time, the vertical axis indicates frequency; a third dimension indicating the amplitude of a particular frequency at a particular time is represented by the intensity or colour of each point in the image. There are many verities of format: occasionally the vertical and horizontal axes are switched, so time runs up and down; sometimes the amplitude is shown as the height of a 3D exterior as an alternative of color or strength. The axes of frequency and amplitude might be either linear or logarithmic, depending on for what purpose graph is being used. Acoustic would usually be represented with a logarithmic amplitude axis, and frequency would be linear to highlight harmonic relationships, or logarithmic to emphasize musical or tonal relationships.

Spectrograms are usually created in one of two ways: approximated as a filterbank that results from a series of bandpass filters or calculated from the time signal using the short-time Fourier transform (STFT). These two processes actually form two different quadratic Time-Frequency Distributions but are similar under some conditions. The STFT is usually a digital process which creates spectrogram. In the time domain, digitally sampled data is broken up into chunks which usually overlap and fourier transformed to calculate the magnitude of the frequency spectrum for each chunk. Vertical line in the image can be represented by each chunk and it a measurement of magnitude versus frequency for a exact point in time. The time plots are then placed side by side to form the image or a three-dimensional surface. The spectrogram of a signal s(t) can be estimated by computing the squared magnitude of the STFT of the signal s(t).

\[
\text{spectrogram (t,w) = } |\text{STFT}(t,w)|^2
\]

Spectrograms are helpful in overcoming speech defects and in speech training for the portion of the population that is extremely deaf. The observation of phonetics and speech
synthesis are often provided through the use of spectrograms. By reversing the process of producing a spectrogram, it is possible to create a signal whose spectrogram is an random image. This method can be used to secrete a picture in a piece of audio and has been employed by several electronic music artists. Spectrograms can be used to examine the results of passing a test signal through a signal processor such as a filter in order to check its presentation. High definition spectrograms are used in the expansion of RF and microwave systems.

From the formula above, it appears that a spectrogram contains no information about the exact phase of the signal that it represents. For this reason, it is not possible to reverse the process and generate a copy of the original signal from a spectrogram, though in situations where the exact initial phase is unimportant, it may be possible to generate a useful approximation of the original signal. In fact, there is some phase information in the spectrogram, but it appears in another form, as time delay which is the dual of the instantaneous frequency. Steps for the spectrogram shown below.

![Figure 2.8 Take FFT by windowing of total speech signal](image)
Generation of spectrogram is very important systematic procedure for measuring all the parameters of the speech signal at a time. Figure 2.8 shows one detail speech signal. That speech signal is segmented through the process of FFT application. Total time domain speech signal should be given partition and by FFT each small sample group is processed.

In the next step as shown in Figure 2.9 individual frequency is plotted which have been taken by function FFT. Each separate partition is about giving information of the individual frequency and its magnitude in that window. It's similar to give speech signal to the spectrogram.

In the next task rotate each values of frequency plot by 90 degree as can be seen in the Figure 2.10. Now individual frequency can be plotted in term of one continuous graph with three axis definition. Three axis included information about time, frequency and magnitude. Plot of the individual frequency component in the FFT mode can be seen in the Figure 2.11.

Figure 2.12 shows combined RGB layout of the spectrogram. Total spectrogram shows very important information of amplitude in terms of the depth of color. Generally in developed application it is necessary to take speech waveform in the real time. Moreover also recording of the noise is also required. So for the analysis of speech and noise signal in the time domain with loudness intensity spectrogram is must which has been shown later on in the next incoming chapters.

2.3.8 Short-time average zero-crossing rate

Speech analysis that attempts to estimate spectral features usually requires a Fourier transform (or other major transformation, such as linear prediction). However, a simple measure called the zero-crossing rate (ZCR) provides basic spectral information in some application at low cost. For a speech signal $s(n)$, a zero-crossing takes place whenever $s(n) = 0$, i.e., the waveform crosses the time axis (changes algebraic sign). Total scheme for evaluation of average zero crossing rate is visualized in Figure 2.13. It is showing how from $x(n)$ zero crossing rate $Zn$ is calculated.
Figure 2.9 Plot individual frequency of each windowed FFT

Figure 2.10 Rotation of 90 degree of windowed plot
Figure 2.11 Individual frequency band representation of with individual spectrogram

Figure 2.12 Three axis representation of speech signal X axis shows time, Y axis shows frequency and color of the segment shows magnitude of signal for that duration and frequency
Taking the simple case of a sinusoid (instead of speech), ZCR (measured as zero-crossings/s) yields two zero-crossings/period, and thus its $F_0 = \text{ZCR}/2$. For all narrowband signals (e.g., sinusoids), the ZCR can accurately measure the frequency where power is concentrated. Most short-time processing methods (both temporal and spectral) yield a parameter sequence in the form of a dynamic signal

$$P(n) = \sum_{m=-\infty}^{\infty} T[s(m)]w(n-m)$$

(2.6)

where the speech $s(n)$ is subject to a possibly nonlinear transformation $T$ and is weighted by the window $w(n)$ to limit the time range examined. The desired parameter $P(n)$ as specified by the nature of $T$ appears as a signal with the original sampling rate, representing some speech characteristic averaged over the window duration. $P(n)$ is the convolution of $T[s(n)]$ and $w(n)$. Since $w(n)$ usually behaves as a lowpass filter, $P(n)$ is a smoothed version of $T[s(n)]$.

Thus, equation serves to calculate the ZCR with

$$T[s(n)] = 0.5[\text{sign}(s(n)) - \text{sign}(s(n-1))]$$

(2.7)

Where the algebraic sign of $s(n)$ will be

$$\text{sign}(s(n)) = \begin{cases} 
1 & \text{for } s(n) \geq 0 \\
-1 & \text{otherwise}
\end{cases}$$

(2.8)

and $w(n)$ is a rectangular window scaled by $1/N$ where $N$ is the duration of the window to yield zero-crossings/sample. The ZCR can be significantly decimated for data reduction purposes. Like speech power, the ZCR changes relatively slowly with vocal tract movements. The ZCR is useful for estimating whether speech is voiced. Voiced speech has mostly low-frequency power, owing to a glottal excitation spectrum that falls off at
about −12 dB per octave. Unvoiced speech comes from broadband noise excitation exciting primarily high frequencies, owing to the use of shorter vocal tracts (anterior to the constriction where noise is produced). Since speech is not narrowband, the ZCR corresponds to the average frequency of primary power concentration. Figure 2.14 shows organization of zero crossing rate for voiced, unvoiced and voiced fricatives in clear view.

Thus high and low ZCR about 4900 and 1400 crossings/s correspond to unvoiced and voiced speech, respectively. For sonorant sounds, the ZCR follows F1 well, since F1 has more energy than other formants except in low back vowels, where F1 and F2 are close, and ZCR is thus usually above F1. Voiced fricatives, on the other hand, have bimodal spectra, with voicebar power at very low frequency and frication energy at high frequency; hence, ZCR in such cases is more variable. Unlike short-time power, the ZCR is quite sensitive to noise, especially any low-frequency bias that may displace the zero-amplitude axis.

Figure 2.14 Typical zero crossing distribution for voiced, unvoiced fricatives and voiced fricatives
2.4 Speech vs Silence Discrimination using Energy and Zero Crossings

The problem of locating the beginning and end of a speech utterance in a background of noise is of importance in many areas of speech processing. In particular, in automatic recognition of isolated words, it is essential to locate the regions of a speech signal that correspond to each word. A scheme for locating the beginning and end of a speech signal can be used to eliminate significant computation in non-real time systems by making it possible to process only the parts of the input that correspond to speech. The problem of discriminating speech from background noise is not trivial, except in the case of extremely high signal to noise ratio acoustic environments. For such high signal to noise ratio environments, the energy of the lowest level speech sounds exceeds the background noise energy and thus a simple energy measurement suffices.

Algorithm for voice detection is based on two simple time domain representation measurements: energy and zero crossing rate. Usually some difficulties encountered in locating the beginning and end of a speech utterance. In Figure 2.15 it can be seen easily that speech and background noise can be portioned. The frequency content of the speech is radically different from the background noise as can be observed by the sharp increase in zero crossing rate of the waveform. The speech energy at the beginning of the utterance is comparable to the background noise energy.

In the case of word /four/ it is very difficult to precisely identify the beginning point. In general, it is difficult to locate the beginning and end of an utterance if there are:

1. Weak fricatives (/f/ , /θ/ , /h/) at the beginning or end.
2. Weak plosive bursts (/p/ , /t/ , /k/) at the beginning or end.
3. Nasals at the end
4. Voiced fricatives which become devoiced at the end of words.
5. Trailing off of vowel sounds at the end of an utterance.

In spite of the difficulties posed by the above situations, energy and zero crossing rate representations can be combined to serve as the basis of a useful algorithm for locating the beginning and end of a speech signal. A system can be designed such a way that a
speaker utters a word during a prescribed recording interval, and the entire interval is sampled and stored for processing. The purpose of the algorithm is to find the beginning and end of the word so that subsequent processing and pattern matching can ignore the surrounding background noise.

![Figure 2.15 Measurement of average magnitude and zero crossing](image)

The basic representations used are the number of zero crossings per 10 msec frame and the average magnitude computed with a 10 msec window. Both functions are measured for the entire recording interval at a rate of 100 times/sec. It is assumed that the first 100 msec of the interval contains no speech. For specified internal, mean and standard deviation of the average magnitude and zero crossing rate are computed which highlights a statistical characterization of the background noise. By utilizing this statistical properties and the maximum average magnitude in the interval, zero crossing rate and energy thresholds are measured as shown in Figure 2.15. The average magnitude profile is taken to find the duration in which it always greater a very conservative threshold ITU. It is assumed that the beginning and ending points lie outside this interval. Then working backwards from the point at which Mn first exceeded the threshold ITU, the point labeled N1 where Mn first having place below a lower threshold ITL is temporary considered as the beginning point. A equivalent steps are followed to find the tentative endpoint N2. This double threshold procedure ensures that dips in the average magnitude function do not falsely signal the endpoint. At this stage it is reasonably safe to assume that the beginning and ending points are not within the interval N1 th N2. The next step is
to move backwards from N1 matching the zero crossing rate to a threshold IZCT calculated from the statistical of

Figure 2.16 Sequence of average magnitude plots showing how the algorithm performed over a variety of words
the zero crossing rate for the background noise. This is limited to the 25 frames proceeding N1. If the zero crossing rate increases the threshold 3 or more times, the starting point N1 is moved back to the first original point at which the zero crossing threshold was exceeded. Otherwise N1 is declared as the beginning. A similar procedure is followed at the end.

Figure 2.16 shows how the algorithm works. There are 8 plots of the average magnitude function for 8 different words. The markers on each plot show the beginning point and ending point of each word as determined by algorithm. The average magnitude thresholds were sufficient to locate the boundary points. The zero crossing algorithms can be used to determine the ending point due to the final fricative /s/. It should be noted that even though the final /s/ has fairly large average magnitude, since the average magnitude thresholds were set conservatively, this criterion algorithm has been defined. The final /t/ in the word delete was correctly located because of the significant zero crossing rate over the 70msec burst when the /t/ was released. Thus even though the average magnitude and zero crossing rate were very small for about 50 msec on the stop gap, the algorithm can identify the endpoint because of the strength of the burst. If the burst has been weak, the ending point would have been located at the beginning of the stop gap.