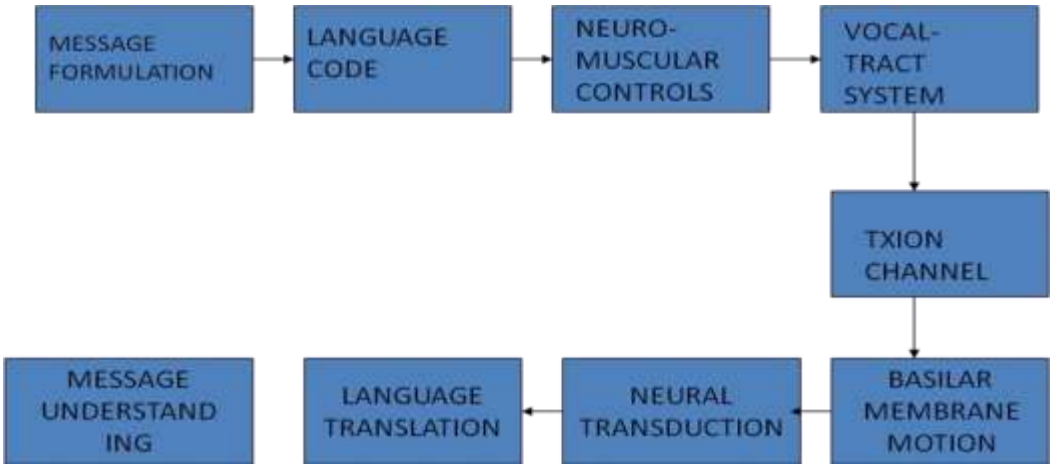
UNIT2

Speech Production block diagram:



Speech communication is the transfer of information from one person to another via speech, which consists of variations in pressure coming from the mouth of a speaker. Such pressure changes propagate as waves through air and enter the ears of listeners, who decipher the waves into a received message. The chain of events from the concept of a message in a speaker's brain to the arrival of the message in a listener's brain is called the speech chain The chain consists of a speech production mechanism in the speaker, transmission through a medium (e.g., air), and a speech perception process in the ears and brain of a listener.

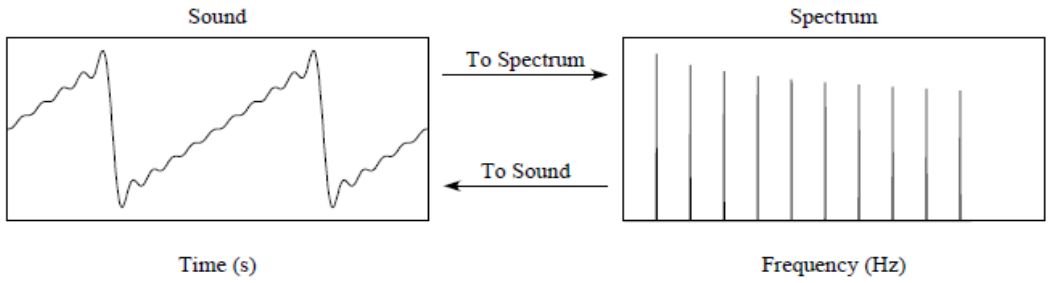
The speaker produces a speech signal in the form of pressure waves that travel from the speaker's head to the listener's ears. This signal consists of variations in pressure as a function of time and is usually measured directly in front of the mouth; the primary sound source (although sound also comes from the nostrils, cheeks, and throat). The amplitude variations correspond to deviations from atmospheric pressure caused by traveling waves. The signal is non stationary (time-varying), changing characteristics as the muscles of the vocal tract contract and relax. Speech can be divided into sound segments, which share some common acoustic and articulatory

properties with one another for a short interval of time. Corresponding to the message to be conveyed, for each sound, there is a positioning for each of the vocal tract articulators: vocal folds (or cords), tongue, Lips, teeth, velum, and jaw. Sounds are typically divided into two broad classes: (a) vowels, which allow unrestricted airflow in the vocal tract, and (b) consonants, which restrict airflow at some point and have a weaker intensity than vowels.

The spectrum

The spectrum is an invaluable aid in studying differences between speech sounds. Almost all analyses that compare sounds, are based on spectra. The spectrum is a frequency-domain representation of a sound signal; the spectrum gives information about frequencies and their relative strengths.

A spectrum and a sound are different. A sound you can hear, a spectrum not. The spectrum is a (mathematical) construct to represent a sound for easier analysis. One makes calculations with a spectrum, one visualizes aspects of a spectrum but you can not hear it or touch it. Only after you have synthesized the sound from the spectrum, can you listen to the sound. The reason for the popularity of the spectrum is that it is often easier to work with than the sound. When the spectrum is calculated from a sound, a mathematical technique called Fourier analysis is used. A Fourier analysis finds all the frequencies in the sound and their amplitudes, i.e. their strengths. There is no information loss in the spectrum: we can get the original sound back from it by Fourier synthesis. These two transformations, analysis and synthesis, that are each others inverse, are visualized in figure 7.1. On the left we see a very small part of a sound as a function of time and on the right the sound as a function of frequency. The top arrow going from the sound to the spectrum, labelled “To Spectrum”, visualizes the Fourier analysis. The bottom arrow, labelled “To Sound”, visualizes Fourier synthesis. Although intuitively the **spectrum is a simple object, i.e. a representation of the frequency content of a signal,** the mathematics to calculate the spectrum from a sound is not simple. The main causes for mathematical complications are first of all the finite duration of the sound and secondly the fact that sounds are sampled in the time domain.



The spectrum is not a simple object like a mono sound but a complex one. Complex has a double meaning in this respect. The first meaning of complex is “composed of two or more parts”. There are two parts in a spectrum: one part represents the amplitudes of all the frequencies and the other part the phases of the frequencies. The other meaning of complex is the mathematical one from “complex number”.1 This is about how the two aspects of a frequency, its amplitude and its phase, are represented. To visualize a

complete spectrum we would need three dimensions: one for frequency, one for amplitude and one for phase. Three dimensional representations are difficult, we therefore limit ourselves to **the most popular two dimensional representation: the amplitude spectrum, where vertically amplitude is displayed in decibel and horizontally frequency in hertz.** Often the amplitude spectrum is visualized in text books in two different ways: as a line spectrum with vertical lines, or as an amplitude spectrum where instead of showing the vertical lines, the tips of the lines are connected. In the sequel we will show that what is visualized as a line spectrum only occurs for very special sound signals. In Praat the amplitude spectrum is always drawn, although for special combinations of tone frequencies and tone durations, the amplitude spectrum may have the appearance of a line spectrum. The most important reason for the popularity of the amplitude spectrum is that the human ear is not very sensitive to the relative phases of the components of a sound, the relative amplitudes of the component are of far more importance

the amplitude spectrum of pure tones. A pure tone can be described mathematically as a function of time as

$$s(t) = a \sin(2\pi f t),$$

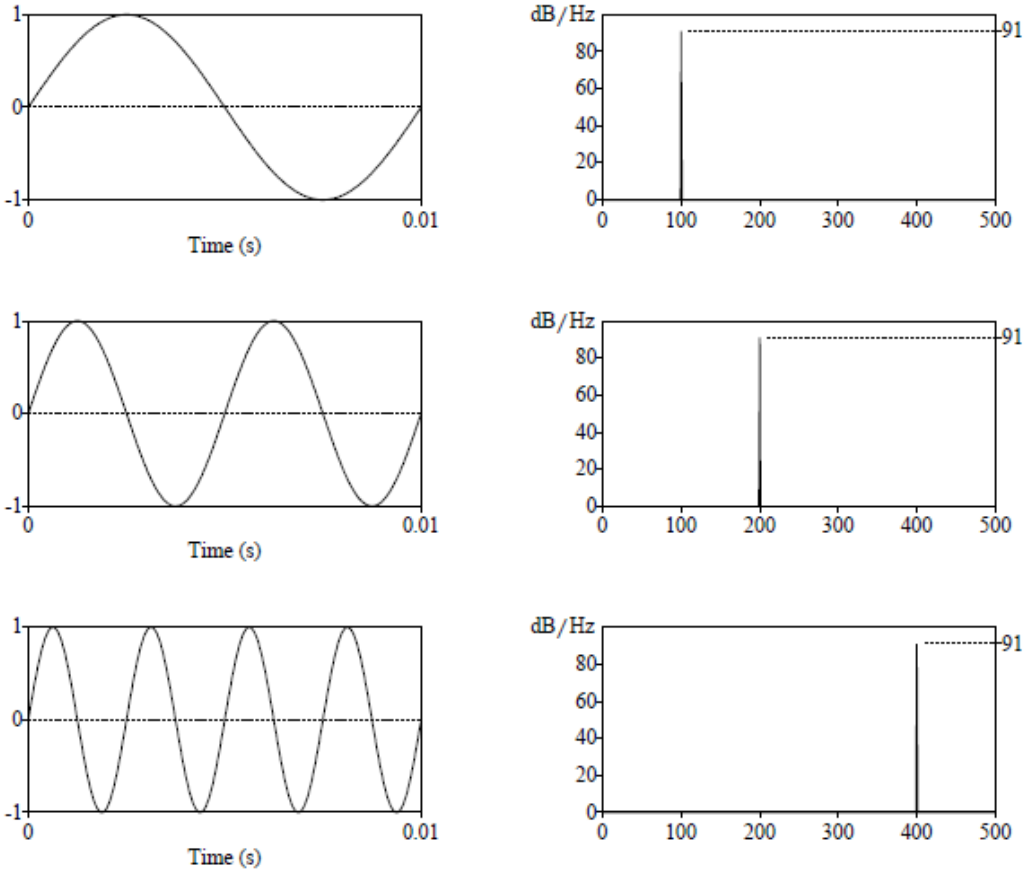


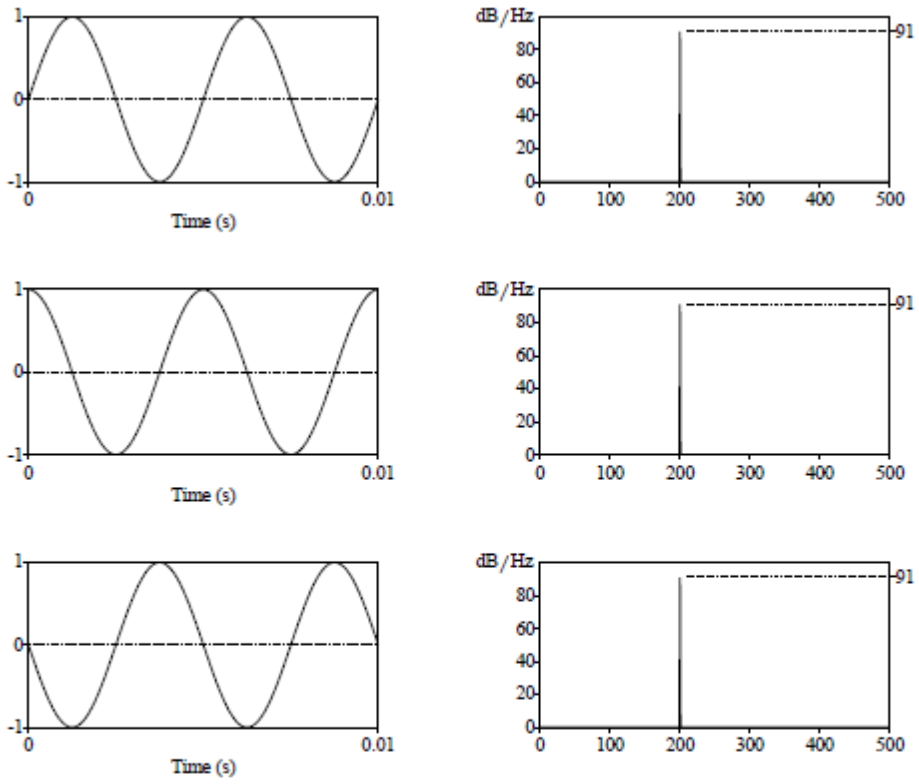
Figure 7.2.: In the left column from top to bottom the first 10 ms of 1 s duration pure tones with frequencies 100, 200 and 400 Hz. The right column shows the amplitude spectrum of each tone.

Our conclusion is that the amplitude spectra of pure tones with equal amplitudes show peaks of equal heights.

This shows that the amplitude and frequency of pure tones are displayed independently of each other in the amplitude spectrum. The frequency determines the position of the line on the frequency axis and the amplitude the height of the line, i.e. its spectral amplitude. Amplitude and frequency are two independent aspects of a pure tone.

What would happen to the amplitude spectrum if the pure tones didn't start at a time where the amplitude is zero? To model this, the sine function of equation is not sufficient because this one always starts with an amplitude of zero at time t = 0. However, an extra parameter in the argument of the sine can change this behaviour. This parameter is called the phase.

$$s(t) = a \sin(2\pi f t + \phi),$$



.: In the left column from top to bottom the first 10 ms of 1 s duration pure tones with frequency 200 Hz and phases of 0, $\pi/2$, and π. The right column shows the amplitude spectrum of each tone.

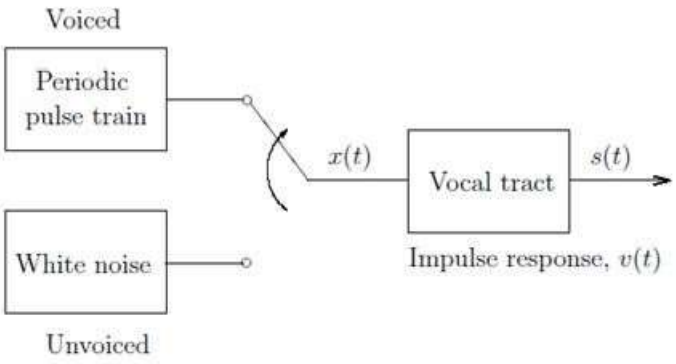
We conclude from this figure that the phase of the tone has no influence on the amplitude spectrum, only sound amplitude and sound frequency matter.

By mixing a sine function and a cosine function we can have any start value we want. In section A.1.3 we show that a mixture of a sine and a cosine function with the same argument is equivalent to a sine with a phase. We translate equation to frequencies and write

$$a \cos(2\pi f t) + b \sin(2\pi f t) = c \sin(2\pi f t + \theta),$$

where the new amplitude is $c = \sqrt{a^2 + b^2}$ and the phase $\theta = \arctan(b/a)$.

Explain the source filter model of Human Speech production system.



Source-Filter Model of Speech Production

Sound is variations in air pressure. The creation of sound is the process of setting the air in rapid vibration. The model of speech production has two major components:

• **Excitation**: It describes how air is set in motion. According to the type of excitation, we have two types of sounds.

Voiced sounds: Periodic air pulses pass through vibrating vocal chords.
Unvoiced sounds: Force air through a constriction in vocal tract, producing turbulence.

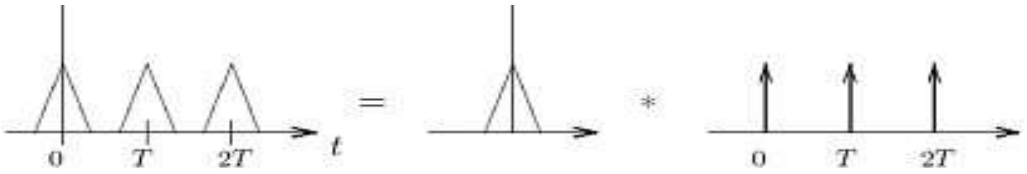
• **Vocal tract**: Guides air.
This system is illustrated in the figure above. Its upper part (the production of voiced sounds) is very much akin to playing a guitar. You produce a sequence of impulsive excitations by plucking

the strings, and then the guitar converts it into music. The strings are sort of like the vocal cords, and the guitar's cavity plays the same role as the cavity of the vocal tract.

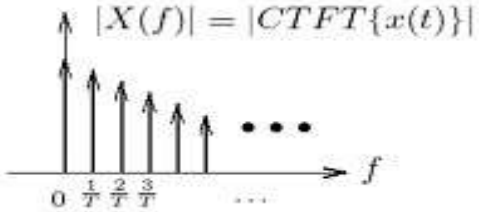
A periodic pulse train excitation is illustrated in Figure. The period *T* is called the *pitch period,* and 1*/T* is called the *pitch frequency*.

For male: *T* _ 8*ms)* pitch _ 125*Hz*; For
female: *T* _ 4*ms)* pitch _ 250*Hz*.

• Different voiced sounds are produced by changing the shape of the vocal tract. Hence this system is time-varying.
However, it is slowly varying as the changes occur slowly compared to the pitch period.



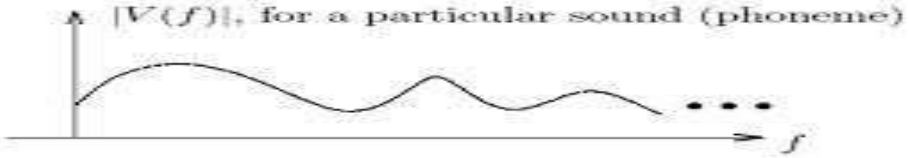
(a) Periodic pulse train.



$\frac{1}{T}$ = pitch frequency

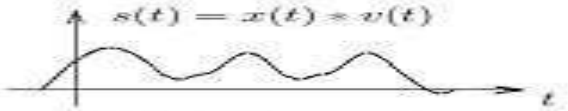
(b) Frequency response.

Figure Time domain and frequency domain perspectives of voiced sounds.

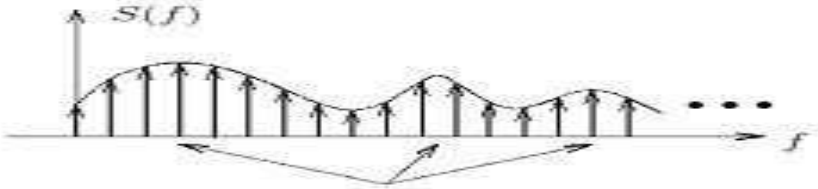


Magnitude response of the vocal tract.

(a) Voice tract transfer function.



(b) Resulting sound wave.



(c) Magnitude response of the resulting sound wave.

The vocal tract as an LTI filter.

In other words, each sound is approximately periodic, but different sounds are different periodic signals. This implies that we can model the vocal tract as an LTI filter over short time intervals. Moreover, since the vocal tract is a cavity, it resonates. **That is, when a wave propagates in a**

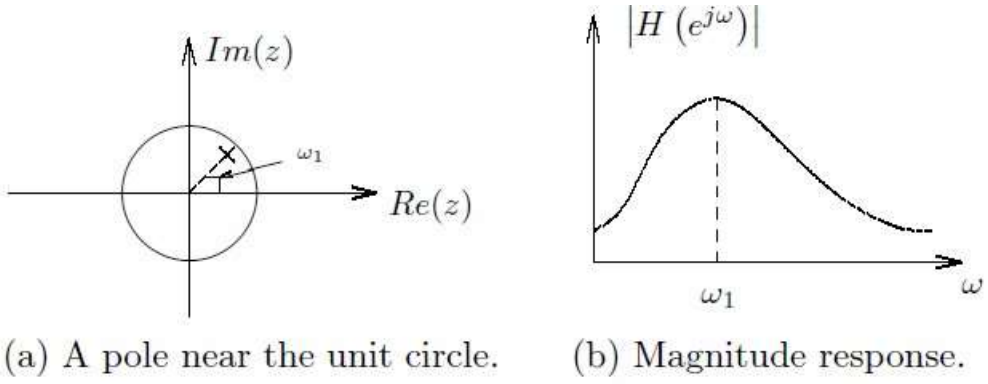
cavity, there is a set of frequencies which get amplified. They are called natural frequenciesof the resonator, and depend on the shape and size of the resonator.

Therefore, the magnitude response of the vocal tract for one voiced sound (phoneme) can be modeled as shown in Fig. above. The waveform for this particular phoneme will then be the convolution of the driving periodic pulse train $x(t)$ with the impulse response $v(t)$ as illustrated in figure. The maxima of $|S(f)|$ are called the *formant frequencies* of the phoneme.

• Typically, one formant per 1 kHz.
• Locations are dictated by the poles of the transfer function.

- The first 3 – 4 formants (range: up to 3.5 kHz) are enough for reasonable reconstruction. Thus, sampling at 3.5·2 kHz = 7 kHz is typically enough. Depending on the application, the sampling rate is usually 7 – 20 kHz.

Suppose we discretize speech, and want to model the vocal tract as a digital filter. The following gives a very rough idea of how to do this. If we know the formant frequencies, we could use what we learned about designing frequency selective filters.



(a) A pole near the unit circle. (b) Magnitude response.

Figure 2.44. Poles near the unit circle correspond to large values of $H(e^{j\omega})$.

Poles of $H(z)$ near the unit circle correspond to large values in the magnitude spectrum .
So, we can design an all-pole filter, with poles which are close to the unit circle, corresponding to formant frequencies. The larger the magnitude response at the formant frequency, the closer the corresponding pole(s) to the unit circle.

Practice problem: *A phoneme whose pitch is 100 Hz, is sampled at 6 kHz. It has two formants: a weak one at 500 Hz and a stronger one at 2 kHz. Sketch the approximate polelocations of H(z).*

Solution: *DT frequency* $\omega = 2\pi$ *corresponds to 6 kHz. Therefore,*

100 Hz corresponds to $\frac{2\pi}{6000} \cdot 100 = \frac{2\pi}{60} \Rightarrow D = 60$;

500 Hz corresponds to $\frac{2\pi}{6000} \cdot 500 = \frac{\pi}{6}$;

2000 Hz corresponds to $\frac{2\pi}{6000} \cdot 2000 = \frac{2\pi}{3}$.

The autocorrelation method is a fundamental technique used in spectral analysis for estimating the spectrum of a signal. It's particularly useful for analyzing speech and audio signals due to its ability to capture periodicity and formant structures inherent in such signals. Here's an overview of how the autocorrelation method works:

Autocorrelation Function:

The autocorrelation function of a signal measures the similarity between the signal and a time-delayed version of itself. Mathematically, for a discrete-time signal $x(n)$, the autocorrelation function $R_x(k)$ is defined as:

$$R_x(k)=\sum_{n=0}^{N-1} x(n)x(n+k)$$

Where k is the lag, and N is the length of the signal. This function essentially quantifies the degree of correlation between the signal and its delayed version for different time shifts k.

Power Spectrum Estimation:

The power spectrum of a signal can be estimated from its autocorrelation function. The relationship between the autocorrelation function $R_x(k)$ and the power spectrum $P_x(f)$ of the signal can be described using the Fourier Transform:

$$|2P_x(f)=|F[R_x(k)]|_2$$

Where $[\cdot]F[\cdot]$ denotes the Fourier Transform. This equation shows that the power spectrum of the signal can be obtained by taking the Fourier Transform of the autocorrelation function and then computing the magnitude squared.

Practical Implementation:

In practice, the autocorrelation method for spectral analysis involves the following steps:

1. **Acquisition of Signal**: The speech or audio signal is first acquired.
2. **Windowing**: Typically, the signal is divided into overlapping segments, and each segment is windowed to reduce spectral leakage.
3. **Autocorrelation Computation**: The autocorrelation function is computed for each segment.
4. **Spectrum Estimation**: The power spectrum is estimated by taking the Fourier Transform of the autocorrelation function.
5. **Formant Analysis**: Formants, which represent the resonant frequencies of the vocal tract, can be identified from the peaks in the power spectrum.

Advantages and Limitations:

- **Advantages**:
  - It's computationally efficient compared to other spectral analysis methods.
  - It's well-suited for analyzing signals with periodic components like speech.
  - It can provide insights into the spectral characteristics of the signal, including formant frequencies.
- **Limitations**:
  - Accuracy depends on the length of the analyzed segment and the presence of noise.
  - It might not be suitable for signals with non-stationary or rapidly changing spectral content.

Applications:

- Speech analysis and synthesis.
- Audio signal processing, including music analysis and synthesis.
- Speaker recognition and verification.
- Voice activity detection.

In conclusion, the autocorrelation method is a valuable tool for spectral analysis of speech and audio signals, offering insights into their periodic and spectral characteristics, with applications ranging from speech processing to audio engineering.

- **Short-time Fourier transform** (STFT), also known as **time-dependent Fourier transform** of a signal $x[n]$ is defined by

$$X_{\mathrm{STFT}}(e^{j\omega}, n) = \sum_{m=-\infty}^{\infty} x[n-m]\, w[m]\, e^{-j\omega m}$$

where $w[n]$ is a suitably chosen window sequence

- If $w[n] = 1$, definition of STFT reduces to that of DTFT of $x[n]$

$$X_{\mathrm{STFT}}(e^{j\omega}, n) = \sum_{m=-\infty}^{\infty} x[n-m]\, w[m]\, e^{-j\omega m}$$

The **Short-Time Fourier Transform (STFT)** is a powerful technique used in signal processing for analyzing and synthesizing signals with time-varying frequency content, such as audio signals. It provides a way to study the frequency content of a signal as it evolves over time by dividing the signal into short segments and analyzing each segment using the Fourier Transform. Here's a detailed mathematical explanation of the STFT:

Analysis:

Given a continuous-time signal $x(t)$, the STFT is computed by dividing $x(t)$ into short overlapping segments and applying the Fourier Transform to each segment.

1. **Windowing**: First, the signal is divided into short segments or frames. Each segment is multiplied by a window function $w(t)$ to minimize spectral leakage and reduce artifacts due to abrupt changes at the segment boundaries. The windowed segment at time t is denoted as $x_w(t,\tau)=x(t)\cdot w(t-\tau)$, where is the time index of the segment.
2. **Fourier Transform**: The Fourier Transform is then applied to each windowed segment. The Fourier Transform of the windowed segment $x_w(t,\tau)$ is denoted as $X(\omega,\tau)$, where is the frequency index.

$$X(\omega,\tau)=\int_{-\infty}^{\infty} x_w(t,\tau)e^{-j\omega t}dt$$

3. **Short-Time Fourier Transform**: The Short-Time Fourier Transform $X(\omega,\tau)$ is obtained by repeating the above process for different time indices τ. This results in a 2D representation of the signal's frequency content as it evolves over time.

Synthesis:

The synthesis process involves reconstructing the original signal from its Short-Time Fourier Transform representation.

1. **Inverse Fourier Transform**: The inverse Fourier Transform is applied to each windowed segment $X(\omega,\tau)$ to obtain the time-domain representation of each segment.

$$x_w(t,\tau)=\frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega,\tau)e^{j\omega t}d\omega$$

2. **Overlap-Add Technique**: Since adjacent segments overlap, the reconstructed segments are added together with proper weighting to avoid discontinuities at the segment boundaries.

$$x(t)=\sum_{\tau} x_w(t,\tau)$$

Window Functions:

Commonly used window functions include the Hamming, Hanning, and Gaussian windows. These functions determine the shape of the segments and affect the trade-off between frequency resolution and temporal resolution.

Advantages and Limitations:

- **Advantages**:
  - Provides time-frequency analysis of signals.
  - Suitable for analyzing signals with time-varying frequency content.
  - Offers control over temporal and frequency resolution through window parameters.
- **Limitations**:
  - Resolution trade-off between time and frequency domains.
  - Artifacts due to windowing, such as spectral leakage and windowing effects.

In conclusion, the Short-Time Fourier Transform is a versatile tool for analyzing and synthesizing signals with time-varying frequency content, providing valuable insights into the signal's time-frequency characteristics. Proper choice of window functions and parameters is crucial for achieving accurate analysis and synthesis results.