3. SPEECH

3.1 Overview

This Chapter is about the nature of speech and describes briefly the processes of human speech production mechanisms. Also, the characteristics of voiced and unvoiced speech, and the technical characteristics of a typical speech signal are described briefly.

3.2 Nature of Speech

Speech is regarded a continuous movement of the human voice mechanisms. There are two types of speech, namely voice and unvoiced. Unvoiced speech is produced when air going through the vocal folds and causes these to vibrate. Voiced speech results when a constriction in the vocal tract causes only turbulent airflow. The result is that a pressure wave forms in front of the lips. Speech is a sampled version of this pressure wave and since it is nearly impossible for a human to reproduce the exact same articulation each time, speech is random in nature. This applies to music since the same music recorded over again will not produce the exact same results. Speech is non-stationary (or quasi stationary or locally stationary) since existence or non-existence of vocal fold vibration causes a change in the distribution of the samples. Speech is regarded locally stationary because there is a limit to how rapid a person can articulate. This implies that a small, isolated (adjoining) portion of a speech signal can be considered as a stationary signal on its own and this ranges from 10 - 20ms. Stationary signals are preferred otherwise a Fourier Transform (FT) of the signal will end up showing frequencies occurring at all time because FT only contains frequency information while no time information is retained. This is why speech is assumed to be locally stationary and processed as a series of isolated stationary signal-fragment signals known as frames.

3.3 Speech Processing

Speech processing extracts the desired information from a speech signal. To process a signal by a digital computer, the signal must be represented in digital form so that it can be used by a digital computer.

3.3.1 Speech signal acquisition

Initially, the acoustic sound pressure wave is transformed into a digital signal suitable for voice processing. A microphone or telephone handset can be used to convert the acoustic wave into an analog signal. This analog signal is conditioned with antialiasing filtering (and possibly additional filtering to compensate for any channel impairments). The antialiasing filter limits the bandwidth of the signal to approximately the Nyquist rate (half the sampling rate) before sampling. The conditioned analog signal is then sampled to form a digital signal by an analog-to digital (A/D) converter. Today's A/D converters for speech applications typically sample with 12 to 16 bits of resolution at 8,000 to 20,000 samples per second.

Oversampling is commonly used to allow a simpler analog antialiasing filter and to control the fidelity of the sampled signal precisely (e.g., sigma-delta converters). In local speaker-verification applications, the analog channel is simply the microphone, its cable, and analog signal conditioning. Thus, the resulting digital signal can be very high quality, lacking distortions produced by transmission of analog signals over telephone lines.

3.3.2 Speech production

There are two main sources of speaker-specific characteristics of speech: physical and learned. Vocal tract shape is an important physical distinguishing factor of speech. The vocal tract is generally considered as the speech production organ above the vocal folds. As shown in Figure 3.1, this includes the following: laryngeal pharynx (beneath epiglottis), oral pharynx (behind the tongue, between the epiglottis and velum), oral cavity (forward of the velum and bounded by the lips, tongue, and palate), nasal pharynx (above the velum, rear end of nasal cavity), and the nasal cavity (above the palate and extending from the pharynx to the nostrils). An adult male vocal tract is approximately 17 cm long.

The vocal folds (formerly known as vocal cords) are shown in Figure 3.1. The larynx is composed of the vocal folds, the top of the cricoid cartilage, the arytenoids cartilages, and the thyroid cartilage (also known as "Adam's apple"). The vocal folds are stretched between the thyroid cartilage and the arytenoid cartilages. The area between the vocal folds is called the glottis.

As the acoustic wave passes through the vocal tract, its frequency content (spectrum) is altered by the resonances of the vocal tract. Vocal tract resonances are called *formants*. Thus, the vocal tract shape can be estimated from the spectral shape (e.g., formant location and spectral tilt) of the voice signal.

Voice verification systems typically use features derived only from the vocal tract. As seen in Figure 3.1, the human vocal mechanism is driven by an excitation source, which also contains speaker-dependent information. The excitation is generated by airflow from the lungs, carried by the trachea (also called the "wind pipe") through the vocal folds (or the arytenoid cartilages). The excitation can be characterized as phonation, whispering, frication, compression, vibration, or a combination of these.

3.4 Designing Effective Speech

Speech applications are like conversations between the user and the computer. Conversations are characterized by turn-taking, shifts in initiative, and verbal and nonverbal feedback to indicate understanding.

A major benefit of incorporating speech in an application is that speech is natural: people find speaking easy; conversation is a skill most master early in life and then practice frequently. At a deeper level, naturalness refers to the many subtle ways people cooperate with one another to ensure successful communication.

An effective speech application is one that simulates some of these core aspects of human-human conversation. Since language use is deeply ingrained in human behavior, successful speech interfaces should be based on an understanding of the different ways that people use language to communicate. Speech applications should adopt language conventions that help people know what they should say next and that avoid conversational patterns that violate standards of polite, cooperative behavior.

3.5 When to Use Speech

A crucial factor in determining the success of a speech application is whether or not there is a clear benefit to using speech. Since speech is such a natural medium for communication, users expectations of a speech application tend to be extremely high.

This means speech is best used when the need is clear for example, when the user's hands and eyes are busy or when speech enables something that cannot otherwise be done, such as accessing electronic mail or an on-line calendar over the telephone.

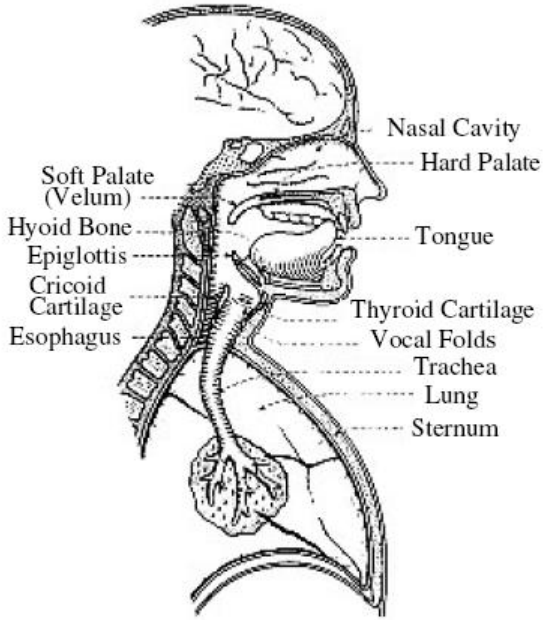

Figure 3.1 Human vocal system [28].

Speech applications are most successful when users are motivated to cooperate. For example, telephone companies have successfully used speech recognition to automate collect calls. People making a collect call want their call to go through, so they answer prompts carefully. People accepting collect calls are also motivated to cooperate, since they do not want to pay for unwanted calls or miss important calls from their friends and family. Automated collect calling systems save the company money and benefit users. Telephone companies report that callers prefer talking to the computer because they are sometimes embarrassed by their need to call collect and they feel that the computer makes the transaction more private.

Speech is well suited to some tasks, but not for others. The tables Table 3.1 and Table 3.2 list characteristics that can help you determine when speech input and output are appropriate choices.

Including speech in an application because it is a novelty means it probably will not get used. Including it because there is some compelling reason increases the likelihood for success.

Table 3.1 When is speech input appropriate?

Use When	Avoid When
- No keyboard is available (e.g., over the telephone, at a kiosk, or on a portable device). - Task requires the user's hands to be occupied so they cannot use a keyboard or mouse (e.g., maintenance and repair, graphics editing). - Commands are embedded in a deep menu structure. - Users are unable to type or are not comfortable with typing. - Users have a physical disability	- Task requires users to talk to other people while using the application. - Users work in a very noisy environment. - Task can be accomplished more easily using a mouse and keyboard

3.6 Challenges

Even if you design an application with speech in mind from the outset, you face substantial challenges before your application is robust and easy to use. Understanding these challenges and assessing the various trade-offs that must be made during the design process will help to produce the most effective interface.

Table 3.2 When is speech output appropriate?

Use When	Avoid When
- Task requires the user's eyes to be looking at something other than the screen (e.g., driving, maintenance and repair). - Situation requires grabbing users attention. - Users have a physical disability (e.g., visual impairment). - Interface is trying to embody a personality.	- Large quantities of information must be presented. - Task requires user to compare data items. - Information is personal or confidential.

3.6.1 Transience: what did you say?

Speech is transient: Once you hear it or say it, it's gone. By contrast, graphics are persistent. A graphical interface typically stays on the screen until the user performs some action.

Listening to speech taxes users' short-term memory. Because speech is transient, users can remember only a limited number of items in a list and they may forget important information provided at the beginning of a long sentence. Likewise, while speaking to a dictation system, users often forget the exact words they have just spoken. Users limited ability to remember transient information has substantial implications for the speech interface design. In general, transience means that speech is not a good medium for delivering large amounts of information.

The transient nature of speech can also provide benefits. Because people can look and listen at the same time, speech is ideal for grabbing attention or for providing an alternate mechanism for feedback. Imagine receiving a notification about the arrival of an e-mail message while working on a spreadsheet. Speech might give the user the opportunity to ask for the sender or

the subject of the message. The information can be delivered without forcing the user to switch contexts.

3.6.2 Invisibility: what can I say?

Speech is invisible. The lack of visibility makes it challenging to communicate the functional boundaries of an application to the user. In a graphical application, menus and other screen elements make most or all of the functionality of an application visible to a user. By contrast, in a speech application it is much more difficult to indicate to the user what actions they may perform, and what words and phrases they must say to perform those actions.

3.6.3 Asymmetry

Speech is asymmetric. People can produce speech easily and quickly, but they cannot listen nearly as easily and quickly. This asymmetry means people can speak faster than they can type, but listen much more slowly than they can read. The asymmetry has design implications for what information to speak and how much to speak. A speech interface designer must balance the need to convey lots of instructions to users with users limited ability to absorb spoken information.

3.6.4 Speech Synthesis Quality

Given that today's synthesizers still do not sound entirely natural, the choice to use synthesized output, recorded output, or no speech output is often a difficult one. Although recorded speech is much easier and more pleasant for users to listen to, it is difficult to use when the information being presented is dynamic. For example, recorded speech could not be used to read people their e-mail messages over the telephone. Using recorded speech is best for prompts that don't change, with synthesized speech being used for dynamic text.

Mixing recorded and synthesized speech, however, is not generally a good idea. Although users report not liking the sound of synthesized speech, they are, in fact, able to adapt to the synthesizer better when it is not mixed with recorded speech. Listening is considerably easier when the voice is consistent.

As a rule of thumb, use recorded speech when all the text to be spoken is known in advance, or when it is important to convey a particular personality to the user. Use synthesized speech when the text to be spoken is not known in advance, or when storage space is limited. Recorded audio requires substantially more disk space than synthesized speech.

3.6.5 Speech Recognition Performance

Speech recognizers are not perfect listeners. They make mistakes. A big challenge in designing speech applications, therefore, is working with imperfect speech recognition technology. While this technology improves constantly, it is unlikely that, in the foreseeable future, it will approach the robustness of computers in science fiction movies.

An application designer should understand the types of errors that speech recognizers make and the common causes of these errors. Unfortunately, recognition errors cause the user to form an incorrect model of how the system works. For example, if the user says "Read the next message," and the recognizer hears "Repeat the message," the application will repeat the current message, leading the user to believe that "Read the next message" is not a valid way to ask for the next message. If the user then says "Next," and the recognizer returns a rejection error, the user now eliminates "Next" as a valid option for moving forward. Unless there is a display that lists all the valid commands, users cannot know if the words they have spoken should work; therefore, if they don't work, users assume they are invalid.

Some recognition systems adapt to users over time, but good recognition performance still requires cooperative users who are willing and able to adapt their speaking patterns to the needs of the recognition system. This is why providing users with a clear motivation to make speech work for them is essential.

3.6.6 Recognition: Flexibility vs. Accuracy

A flexible system allows users to speak the same commands in many different ways. The more flexibility an application provides for user input, the more likely errors are to occur. In designing a command-and-control style interface, therefore, the application designer must find

a balance between flexibility and recognition accuracy. For example, a calendar application may allow the user to ask about tomorrow's appointments in ways such as:

- What about tomorrow?
- What do I have tomorrow?
- What's on my calendar for tomorrow?
- Read me tomorrow's schedule.
- Tell me about the appointments I have on my calendar tomorrow.

This may be quite natural in theory, but, if recognition performance is poor, users will not accept the application. On the other hand, applications that provide a small, fixed set of commands also may not be accepted, even if the command phrases are designed to sound natural (e.g., Lookup tomorrow). Users tend to forget the exact wording of fixed commands. What seems natural for one user may feel awkward for another.

3.7 Voiced and Unvoiced Speech

Voiced or unvoiced speech provides some information on the articulation of sound. Vowels are usually voiced. Voiced or unvoiced speech can be detected because voiced signals are quasi periodic while unvoiced speech has a random nature. This is viewed below in Figure 3.2.

3.8 Technical Characteristics and Analysis of the Speech Signal

An engineer looking at (or listening to) a speech signal might characterize it as follows:

• The bandwidth of the signal is 4 kHz.

• The signal is periodic with a fundamental frequency between 80 Hz and 350 Hz

• There are peaks in the spectral distribution of energy at

$$(2n - 1) * 500 \text{ Hz} ; \quad n = 1, 2, 3, . .$$

• The envelope of the power spectrum of the signal shows a decrease with increasing frequency (-6dB per octave)

Many techniques are used to analyze a speech waveform, among them a few important ones are enumerated below.

3.8.1 Bandwidth

The bandwidth of the speech signal is much higher than the 4 kHz stated above. In fact, for the fricatives, there is still a significant amount of energy in the spectrum for high and even ultrasonic frequencies. However, as we all know from using the (analog) phone, it seems that within a bandwidth of 4 kHz the speech signal contains all the information necessary to understand a human voice.

3.8.2 Oscillogram (waveform)

Physically the speech signal is a series of pressure changes in the medium between the sound source and listener. The most common representation of the speech signal is the oscillogram, often called the waveform (see Figure 3.3). In this time axis is the horizontal axis from the left to write and the curve shows how the pressure increases and decreases in the signal. However, a suitable structure is extremely difficult to extract from the mass of information in the intensity waveform. This difficulty motivates us to search for some transformation of the raw intensity waveform into a different representation where the important structure is easier to identify and the enormous amount of variability is reduced.

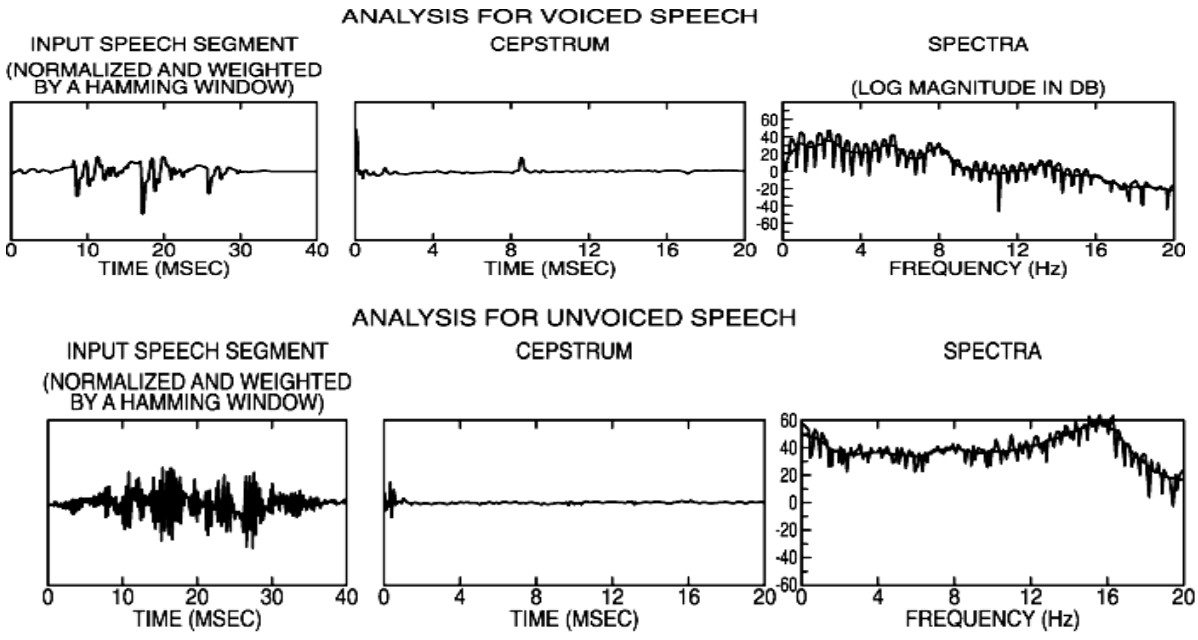

Figure 3.2 Voiced Speech and Unvoiced Speech.[27]

3.8.3 Fundamental frequency (pitch)

Another representation of the speech signal is the one produced by a pitch analysis. Speech is normally looked upon as a physical process consisting of two parts: a product of a sound source (the vocal chords) and filtering (by the tongue, lips, teeth etc). The pitch analysis tries to capture the fundamental frequency of the sound source by analyzing the final speech utterance. The fundamental frequency is the dominating frequency of the sound produced by the vocal chords. This analysis is quite difficult to perform. Several algorithms have been developed, but no algorithm has been found which is efficient and correct for all situations. The fundamental frequency is the strongest correlate to how the listener perceives the speaker's accent and stress.

Using voiced excitation for the speech sound will result in a pulse train, the so-called fundamental frequency. Voiced excitation is used when articulating vowels and some of the consonants. For fricatives (e.g., /f/ as in fish or /s/, as in mess), unvoiced excitation (noise) is used. In these cases, usually no fundamental frequency can be detected. On the other hand, the zero crossing rate of the signal is very high. Plosives (like /p/ as in put), which use transient excitation, you can best detect in the speech signal by looking for the short silence necessary to build up the air pressure before the plosive bursts out.

3.8.4 Spectrum

According to general theories each periodical waveform may be described as the sum of a number of simple sine waves, each with a particular amplitude, frequency and phase. The spectrum gives a picture of the distribution of frequency and amplitude at a moment in time (see Figure 3.4). The horizontal axis represents frequency, and the vertical axis amplitude. If we want to plot the spectrum as a function of the time we need a way of representing a three-dimensional diagram, one such representation is the spectrogram. As shown in Figure 3.5, various speakers have peaks at certain frequencies, resulting in varied speech qualities.

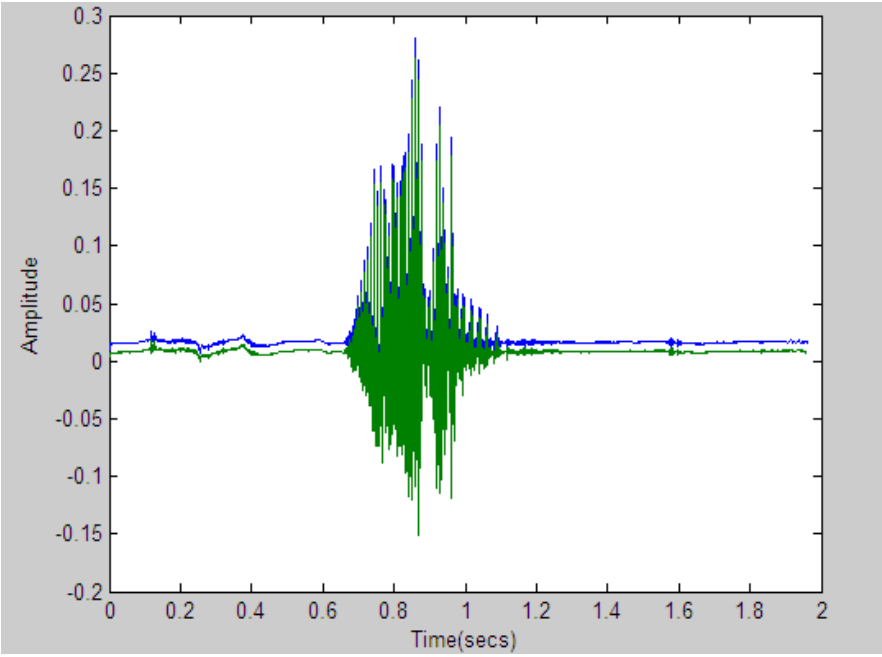

Figure 3.3 A speech signal waveform for sentence "zero".

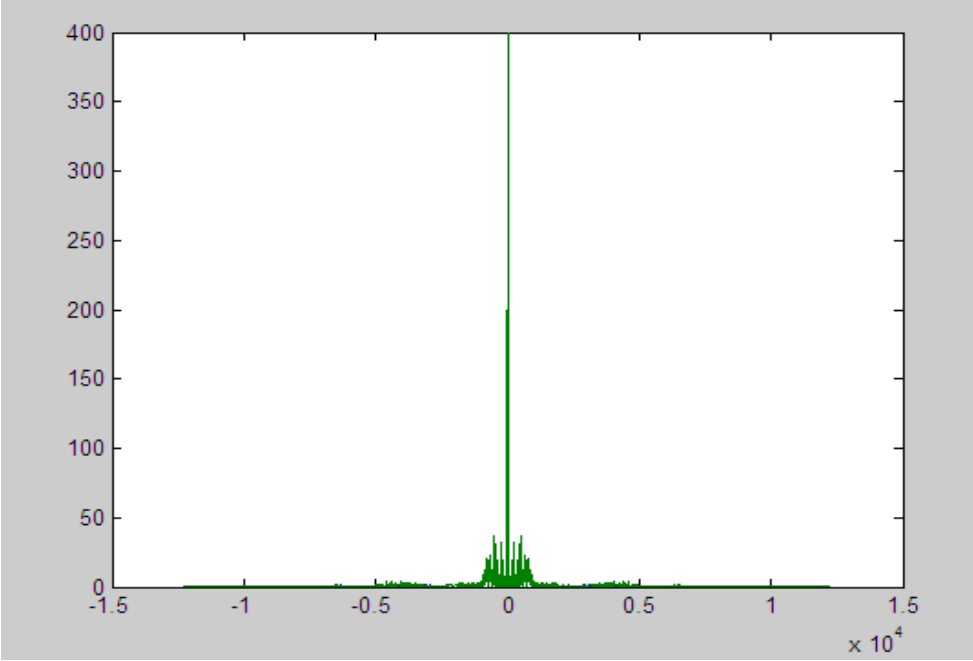

Figure 3.4 Speech spectrum while speaking the sentence "zero".

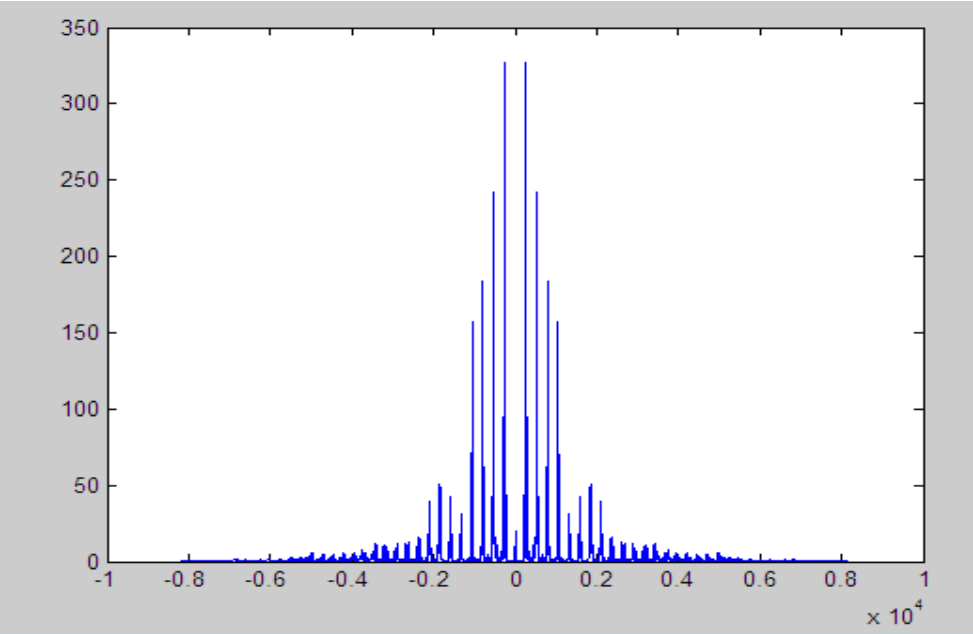

Figure 3.5 Spectrum of another user speaking the same sentence.

3.8.5 Spectrogram

In the spectrogram the time axis is the horizontal axis and the frequency is the vertical axis (see Figure 3.6). The third dimension, amplitude, is represented by shades of darkness. Spectrogram can be considered as a number of a spectrums in a row, looked upon "from above", and where the highs in the spectra are represented with dark spots in the spectrogram. From the picture it is obvious how different the speech sounds are from spectral point of view. The voiced sounds appear more organized. The spectrum highs (dark spots) actually from horizontal bands across the spectrogram. These bands represents frequencies where the shape of the mouth gives resonance to sound. The bands are called formants. The positions of the formants are different sounds.

In Figure 3.7, the areas containing the highest level of energy are displayed in red. As can be seen in the plot, the red area is located between 0.6 and 0.9 seconds. The plot also shows that most of energy is concentrated in the lower frequencies (between 0 Hz and 1 kHz).

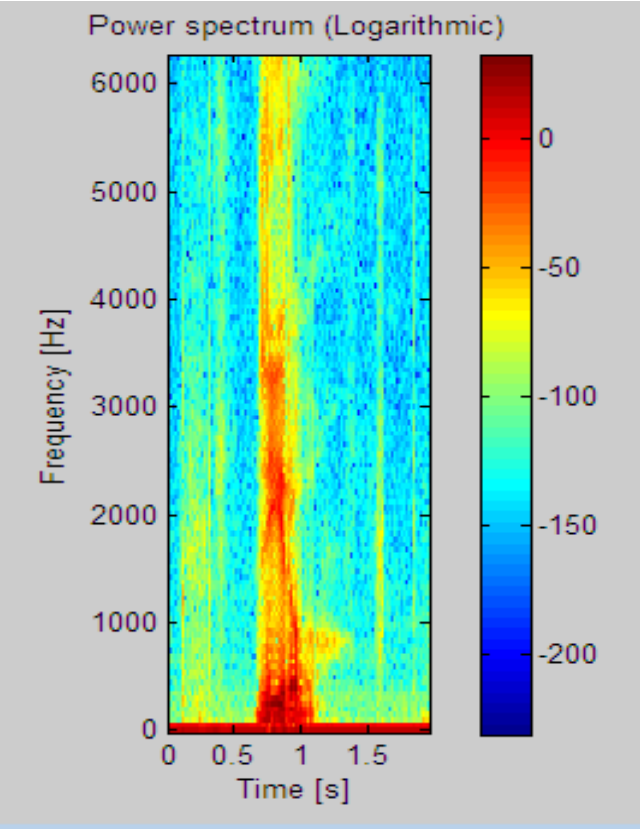

Figure 3.6 A speech spectrogram of an user speaking the sentence "zero"

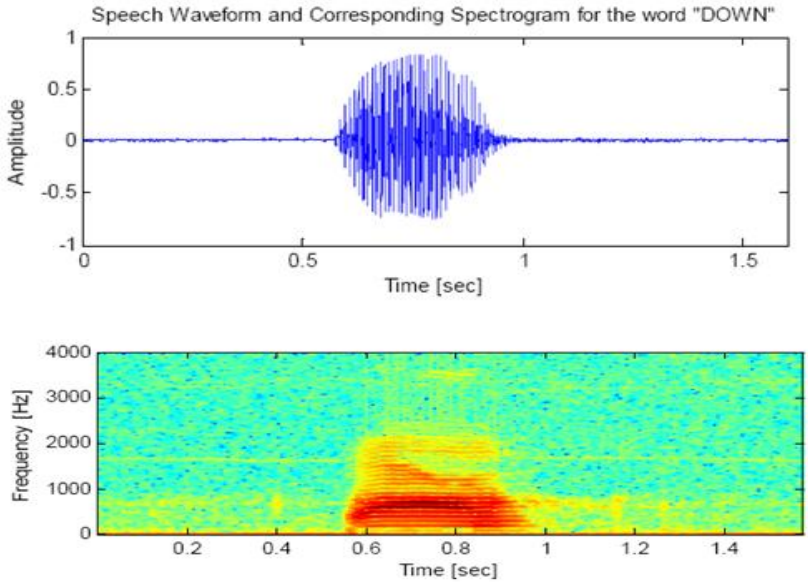

Figure 3.7 Speech waveform (top plot) and associated spectrogram (bottom plot) of the word "down". [22]

3.8.6 Cepstrum

The general problem of fundamental frequency estimation is to take a portion of a signal and to find the dominant frequency of repetition. Difficulties arise from:

- That not all signals are periodic.
- Those that are periodic may be changing in fundamental frequency over the time of interest.
- Signals may be contaminated with noise, even with periodic signals of other fundamental frequencies.
- Signals that are periodic with interval T are also periodic with interval 2T, 3T etc, so we need to find the smallest periodic interval or the highest fundamental frequency.
- Even signals of constant fundamental frequency may be changing in other ways over the interval of interest.

A reliable way of obtaining an estimate of the dominant fundamental frequency for long, clean, stationary speech signals is to use the cepstrum. The cepstrum is a Fourier analysis of the logarithmic amplitude spectrum of the signal. If the log amplitude spectrum contains many regularly spaced harmonics, then the Fourier analysis of the spectrum will show a peak corresponding to the spacing between the harmonics: i.e. the fundamental frequency. Effectively we are treating the signal spectrum as another signal, then looking for periodicity in the spectrum it self.

The cepstrum is a method of speech analysis based on a spectral representation of the signal. One way to think of a speech is as a signal being filtered by the mouth cavity. Assuming that the actual speech (S) one tries to produce is the same for all people, the signal that comes out of the mouth in to a data recorder is that signal filtered by the person's voicebox and throat. If we let v represent this filtering, we can write what we record, r, as the convolution of v and s,

$$r = v*s;$$

We need to deconvolve the vocal tract response and the source signal, thus obtaining the fundamental frequency of the speech. If we move to the frequency domain we would have:

$$R = V S$$

Where R, V, S are the Fourier transforms of r, v, and s respectively. Since we agreed that s is the same for all people, to be able to extract v (or V), there is a need to take logarithm on both sides to separate the variables.

$$\log R = \log V + \log S$$

Thus, an optimal thing for us to compare from sample to sample is the log R quantity instead of just R because the V and S information are combined additively instead of multiplicatively. This type of analysis known as a cepstral analysis.

3.9 Summary

This Chapter has described briefly the technical characteristics and the nature of human speech. It is shown that the speech can be analyzed and displayed in various forms, such as amplitude-time, frequency-time, and power spectrum, and it is also describes the production of the human speech and when to use speech in our life.