Chapter 2

Sound and Audio

2.1 Basic Sound Concept

Sound is perhaps the most important element of multimedia. It is meaningful "speech" in any language, from a whisper to a scream. It can provide the listening pleasure of music, the startling accent of special effects or the ambience of a mood setting background. Sound is the terminology used in the analog form, and the digitized form of sound is called as audio.

Sound is a physical phenomenon produced by the vibration of matter. The matter can be almost anything: a violin string or a block of wood, for example. As the matter vibrates, pressure variations are created in the air surrounding it. This alternation of high and low pressure is propagated through the air in a wave-like motion. When the wave reaches our ears, we hear a sound.

Sound Transmission

- Sound is transmitted by molecules bumping into each other.
- Sound is a continuous wave that travels through air.
- Sound is detected by measuring the pressure level at a point.

Receiving

- Microphone in sound field moves according to the varying pressure exerted on it.
- Transducer converts energy into a voltage level (i.e. energy of another form - electrical energy)

Sending

- Speaker transforms electrical energy into sound waves.

Frequency of a sound wave

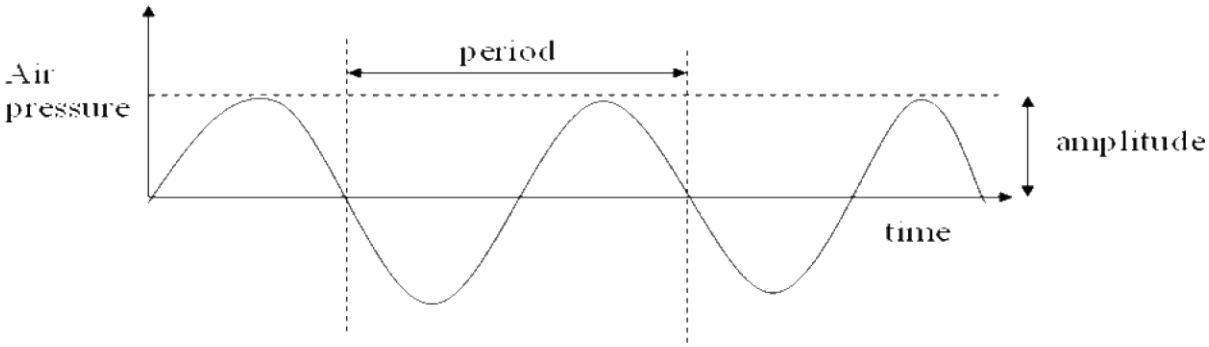

Figure 2.1: Oscillation of an air pressure wave.

The pattern of the pressure oscillation is called a *waveform*. The waveform in repeats the same shape at regular intervals and this portion of the waveform is called a *period*. A waveform with a clearly defined period occurring at regular intervals is called a*periodic waveform*.

Since they occur naturally, sound waveforms are never as perfectly smooth nor as uniformly periodic as the waveform shown in figure 2.1. However, sounds that display a recognizable periodicity tend to be more musical than those that are nonperiodic. Here are some sources of periodic and nonperiodic sounds:

Periodic

- Musical instruments other than unpitched percussion
- Vowel sounds
- Bird songs
- Whistling wind

Nonperiodic

- Unpitched percussion instruments
- Consonants, such as “t,” “f,” and “s”
- Coughs and sneezes
- Rushing water

Frequency

The *frequency* of a sound is the reciprocal value of the period; it represents the number of times the pressure rises and falls, or oscillates, in a second and is measured in *hertz* (Hz) or *cycles per second (cps)*. A frequency of 100 Hz means 100 oscillations per second. A

convenient abbreviation, kHz for *kilohertz*, is used to indicate thousands of oscillations per second: 1 kHz equals 1000 Hz.

The frequency range of normal human hearing extends from around 20 Hz up to about 20 kHz.

- Wavelength is the distance travelled in one cycle

  20Hz is 56 feet, 20KHz is 0.7 in.

Frequency represents the number of periods in a second (measured in hertz, cycles/second).

Human hearing frequency range: 20Hz - 20Khz, voice is about 500Hz to 2Khz.

The frequency range is divided into:

Infrasound	from 0 - 20 Hz
Human range	from 20Hz - 20KHz
Ultrasound	from 20kHz - 1GHz
Hyper sound	from 1GHz - 10THz

Amplitude

A sound also has an *amplitude*, a property subjectively heard as loudness. The amplitude of a sound is the measure of the displacement of air pressure from its mean, or quiescent state. The greater the amplitude, the louder the sound.

- ❖ Amplitude of a sound is the measure of the displacement of the air pressure wave from its mean or quiescent state.
- ❖ Subjectively heard as loudness. Measured in decibels.

  0 db - essentially no sound heard

  35 db - quiet home

  70 db - noisy street

  120db - discomfort

Computer Representation of Audio

The smooth, continuous curve of a sound waveform isn't directly represented in a computer. A computer measures the amplitude of the waveform at regular time intervals to produce a series of numbers. Each of these measurements is called a *sample*. Figure illustrates one period of a digitally sampled waveform.

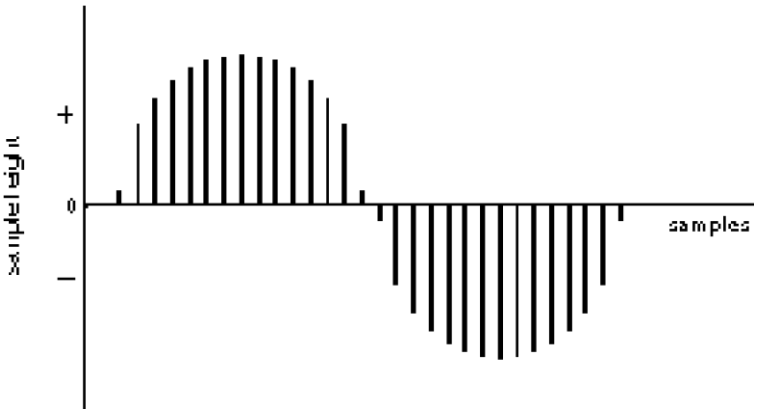

Figure 2.2: Sampled Waveform

Each vertical bar in Figure 2-2 represents a single sample. The height of a bar indicates the value of that sample.

The mechanism that converts an audio signal into digital samples is called an *analog-to digital converter*, or *ADC*. To convert a digital signal back to analog, you need a *digitalto-analog converter*, or *DAC*.

- A transducer converts pressure to voltage levels.
- Convert analog signal into a digital stream by discrete sampling.
- Discretization both in time and amplitude (quantization).
- In a computer, we sample these values at intervals to get a vector of values.
- A computer measures the amplitude of the waveform at regular time intervals to produce a series of numbers (samples).

Sampling Rate

The rate at which a waveform is sampled is called the *sampling rate*. Like frequencies, sampling rates are measured in hertz. The CD standard sampling rate of 44100 Hz means that the waveform is sampled 44100 times per second. This may seem a bit excessive, considering that we can't hear frequencies above 20 kHz; however, the highest frequency that a digitally sampled signal can represent is equal to half the sampling rate. So, a sampling rate of 44100 Hz can only represent frequencies up to 22050 Hz, a boundary much closer to that of human hearing.

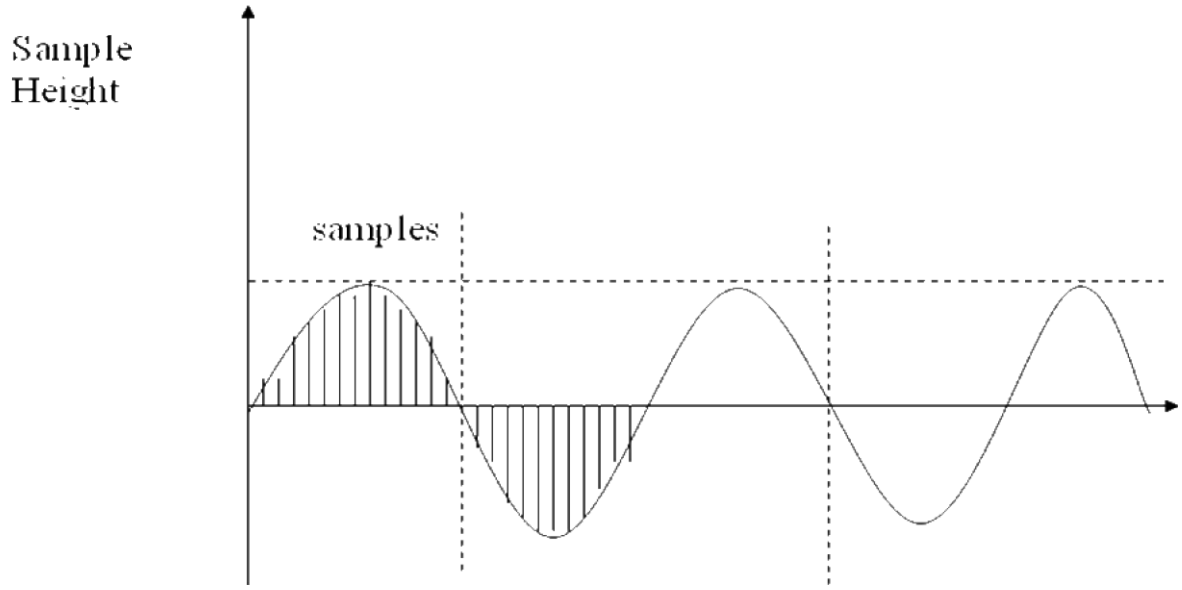

Figure 2.3: Sampling

- ❖ Rate at which a continuous wave is sampled (measured in Hertz)
  - CD standard - 44100 Hz, Telephone quality - 8000 Hz.
- ❖ Direct relationship between sampling rate, sound quality (fidelity) and storage space.

Quantization

Just as a waveform is sampled at discrete times, the value of the sample is also discrete. The *quantization* of a sample value depends on the number of bits used in measuring the height of the waveform. An 8-bit quantization yields 256 possible values; 16-bit CD quality quantization results in over 65000 values. As an extreme example, Figure 2-3 shows the waveform used in the previous example sampled with a 3-bit quantization.

This results in only eight possible values: .75, .5, .25, 0, -.25, -.5, -.75, and -1.

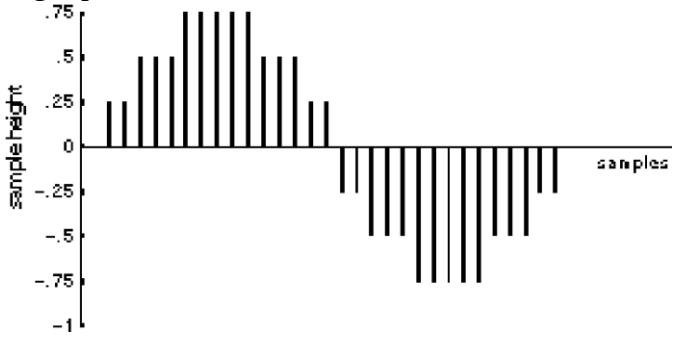

Figure 2.4: Three - bit quantization

Nyquist Sampling Theorem

If a signal f(t) is sampled at regular intervals of time and at a rate higher than twice the highest significant signal frequency, then the samples contain all the information of the original signal.

Example**:**

- ❖Actual playback frequency for CD quality audio is 22050 Hz**.**
- ❖Because of Nyquist Theorem - we need to sample the signal twice, therefore sampling frequency is 44100 Hz.

Data Rate of a Channel

Noiseless Channel

Nyquist proved that if any arbitrary signal has been run through a low pass filter of bandwidth H, the filtered signal can be completely reconstructed by making only 2H (exact) samples per second. If the signal consists of V discrete levels, Nyquist's theorem states:

*Max data rate = 2 *H log_2 V bits/sec*

- Noiseless 3kHz channel with quantization level 1 bit cannot transmit binary signal at a rate exceeding 6000 bits per second.

Noisy Channel

Thermal noise present is measured by the ratio of the signal power S to the noise power N (signal-to-noise ratio S/N).

Max data rate - H log_2 (1+S/N)

Audio Formats

Audio formats are characterized by four parameters:

- ❖*Sample rate:* Sampling frequency
- ❖*Encoding:* audio data representation
  - μ-law encoding corresponds to CCITT G.711 - standard for voice data in telephone companies in USA, Canada, Japan
  - A-law encoding - used for telephony elsewhere.
  - A-law and μ-law are sampled at 8000 samples/second with precision of 12bits, compressed to 8-bit samples.

- Linear Pulse Code Modulation (PCM) - uncompressed audio where samples are proportional to audio signal voltage.

❖ *Precision:* number of bits used to store audio sample

- μ-law and A-law - 8-bit precision, PCM can be stored at various precisions, 16-bit PCM is common.

❖*Channel:* Multiple channels of audio may be interleaved at sample boundaries.

❖ *Available on UNIX*

- au (SUN file format), wav (Microsoft RIFF/waveform format), al (raw a-law), u (raw u-law) …

❖ *Available on Windows-based systems (RIFF formats)*

- wav, midi (file format for standard MIDI files), avi

❖ *RIFF (Resource Interchange File Format)*

- tagged file format (similar to TIFF). Allows multiple applications to read files in RIFF format

❖*RealAudio, MP3 (MPEG Audio Layer 3)*

Digital Audio

Digital audio is created when a sound wave is converted into numbers – a process referred to as digitizing. It is possible to digitize sound from a microphone, a synthesizer, existing tape recordings, live radio and television broadcasts, and popular CDs. You can digitize sounds from a natural source or prerecorded.

Digitized sound is sampled sound.

2.2 Basic Music (MIDI) Concepts

The relationship between music and computer has become more and more important, especially considering the development of MIDI (Musical Instrument Digital Interface) and its important contributions in the music industry today.

MIDI (Musical Instrument Digital Interface) is a communication standard developed for electronic musical instruments and computers. MIDI files allow music and sound

synthesizers from different manufacturers to communicate with each other by sending messages along cables connected to the devices.

Creating your own original score can be one of the most creative and rewarding aspects of building a multimedia project, and MIDI (Musical Instrument Digital Interface) is the quickest, easiest and most flexible tool for this task.

The process of creating MIDI music is quite different from digitizing existing audio. To make MIDI scores, however you will need sequencer software and a sound synthesizer. The MIDI keyboard is also useful to simply the creation of musical scores. An advantage of structured data such as MIDI is the ease with which the music director can edit the data.

A MIDI file format is used in the following circumstances:

- Digital audio will not work due to memory constraints and more processing power requirements
- When there is high quality of MIDI source
- When there is no requirement for dialogue.

Computer Representation of Music

MIDI (Music Instrument Digital Interface) is a standard that manufacturers of musical instruments use so that instruments can communicate musical information via computers.

The MIDI interface consists of:

- *Hardware* - physical connection b/w instruments, specifies a MIDI port (plugs into computers serial port) and a MIDI cable.
- *Data format* - has instrument specification, notion of beginning and end of note, frequency and sound volume. Data grouped into MIDI messages that specify a musical event.
- *An instrument that satisfies both is a MIDI device (*e.g. synthesizer)
- *MIDI software applications* include a music recording and performance applications, musical notations and printing applications, music education etc.

MIDI Reception Modes:

- *Mode 1: Omni On/Poly*
- *Mode 2: Omni On/Mono*
- *Mode 3: Omni Off/Poly*

- *Mode 4: Omni Off/Mono*

The first half of the mode name specifies how the MIDI device monitors the incoming MIDI channels.

If Omni is turned on, the MIDI device monitors all the MIDI channels and responds to all channel messages, no matter which channel they are transmitted on.

If Omni is turned off, the MIDI device responds only to channel messages sent on the channel(s) the device is set to receive.

The second half of the mode name tells the MIDI device how to play notes coming in over the MIDI cable.

If the option Poly is set, the device can play several notes at a time.

If the mode is set to Mono, the device plays notes like a monophonic synthesizer-one note at a time.

MIDI Devices

The heart of any MIDI system is the MIDI synthesizer device. Most synthesizer have the following common components:

- *Sound Generators:*

  Sound generators do the actual work of synthesizing sound; the purpose of the rest of the synthesizer is to control the sound generators.

- *Microprocessors:*

  The microprocessor communicates with the keyboard to know what notes the musician is playing, and with the control panel to know what commands the musician wants to send to microprocessor. The microprocessor sends and receives MIDI message.

- *Keyboard:*

  The keyboard affords the musician's direct control of the synthesizer.

- *Control Panel:*

  The control panel controls those functions that are not directly concerned with notes and durations (controlled by the keyboard).

- *Auxiliary Controllers:*
  Auxiliary controllers are available to give more control over the notes played on the keyboard. Two very common variables on a synthesizer are pitch bend and modulation.

- *Memory:*
  Synthesizer memory is used to store patches for the sound generators and settings on the control panel.

MIDI Messages

MIDI messages transmit information between MIDI devices and determine what kinds of musical events can be passed from device to device.

MIDI messages are divided into two different types:

(1) Channel Messages:

Channel messages go only to specified devices. There are two types of channel messages:

(i) *Channel Voice Messages:*

Send actual performance data between MIDI devices. Example: Note On, Note Off, Channel Pressure, Control Change etc.

(ii) *Channel Mode Messages:*

Determine the way that a receiving MIDI device responds to channel voice messages. Example: Local Control, All Notes Off, Omni Mode Off etc.

(2) System Messages:

System messages go to all devices in a MIDI system because no channel numbers are specified. There are three types of system messages:

(i) *System Real-time Messages:*

System real time messages are very short and simple, consisting of only one byte. They carry extra data with them. Example: System Reset, Timing Clock etc.

(ii) *System Common Messages:*

System common messages are commands that prepare sequencers and synthesizers to play a song. Example: Song Select, Tune Request etc.

(iii) System Exclusive Messages:

System exclusive messages allow MIDI manufacturers to create customized MIDI messages to send between their MIDI devices.

MIDI Software

The software applications generally fall into four major categories:

(i) Music recording and performance applications:

- Provides functions such as recording of MIDI messages editing and playing back the messages in performance.

(ii) Musical notation and printing applications:

- Allows writing music using traditional musical notation.
- Print the music on paper for live performance or publication.

(iii) Synthesizer patch editors and librarians:

- Allow information storage of different synthesizer patches in the computer's memory and disk drives.
- Editing of patches in the computer.

(iv) Music education applications:

- Teach different aspects of music using the computer monitor, keyboard and other controllers of attached MIDI instruments.

The main issue in current MIDI-based computer music systems is interactivity.

The processing chain of interactive computer music systems can be conceptualized in three stages:

- *The sensing stage,* when data are collected from controller reading gesture information from human performers on stages.
- *The processing stage,* when the computer reads and interprets information coming from the sensors and prepares data for the response stage.
- *The response stage,* when the computer and some collection of sound producing devices share in realizing a musical output.

2.3 Speech

Speech can be "perceived", "understood" and "generated" by humans and also by machines.

A human adjusts himself/herself very efficiently to different speakers and their speech habits.

The brain can recognize the very fine line between speech and noise.

The human speech signal comprises a subjective lowest spectral component known as the pitch, which is not proportional to frequency.

The human ear is most sensitive in the range from 600 Hz to 6000 Hz.

Speech signal have two properties which can be used in speech processing:

- *Voice speech signals* show during certain time intervals almost periodic behavior.
- *The spectrum of audio signals* shows characteristic maxima, which are mostly 3-5 frequency bands.

Speech Generation

Generated speech must be understandable and must sound natural. The requirement of understandable speech is a fundamental assumption, and the natural sound of speech increases user acceptance.

Basic Notions

- The lowest periodic spectral component of the speech signal is called the *fundamental frequency*. It is present in a voiced sound.
- A *phone* is the smallest speech unit, such as the *m* of *mat* and *b* of *bat* in English, that distinguish one utterance or word from another in a given language.
- *Allophones* mark the variants of a phone. For example, the aspirated *p* of *pit* and the unaspirated *p* of *spit* are allophones of the English phoneme *p*.
- The *morph* marks the smallest speech unit which carries a meaning itself. Therefore, *consider* is a morph, but *reconsideration* is not.
- A *voiced sound* is generated through the vocal cords. *m, v* and *l* are examples of voiced sounds. The pronunciation of a voiced sound depends strongly on each speaker.

- During the generation of an *unvoiced sound*, the vocal cords are opened. *f* and *s* are unvoiced sounds. Unvoiced sounds are relatively independent from the speaker.

Exactly, there are:

- *Vowels* - a speech sound created by the relatively free passage of breath through the larynx and oral cavity, usually forming the most prominent and central sound of a syllable (e.g., *u* from hunt);
- *Consonants* - a speech sound produced by a partial or complete obstruction of the air stream by any of the various constrictions of the speech organs (e.g., voiced consonants, such as *m* from *mother*, fricative voiced consonants, such as *v* from *voice*, fricative voiceless consonants, such as *s* from *nurse*, plosive consonants, such as *d* from daily and affricate consonants, such as *dg* from knowledge, or *ch* from *chew*).

Reproduced Speech Output

There are two way of speech generation/output performed by *time-dependent sound concatenation* and a *frequency-dependent sound concatenation*.

Time-dependent Sound Concatenation

Individual speech units are composed like building blocks, where the composition can occur at different levels. The individual phones are understood as speech units. The individual phones of the word *crumb*. It is possible with just a few phones to create an unlimited vocabulary

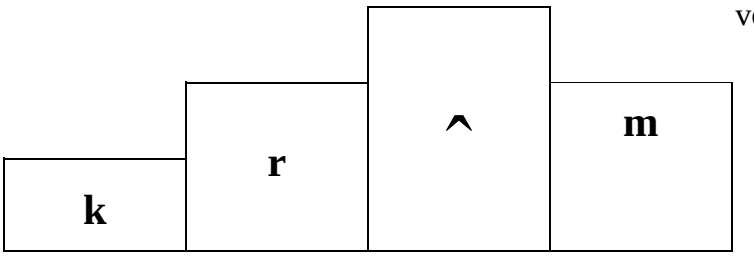

Figure 2.5: Phone sound concatenation.

Two phones can constitute a diphone (from di-phone). Figure 2.6 shows the word *crumb*, which consist of an ordered set of diphones.

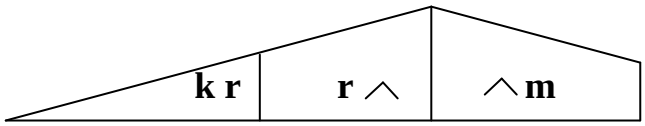

Figure 2.6: Diphone sound concatenation.

To make the transition problem easier, syllables can be created. The speech is generated through the set of syllables. Figure 2.7 shows the syllable sound of the word *crumb.*

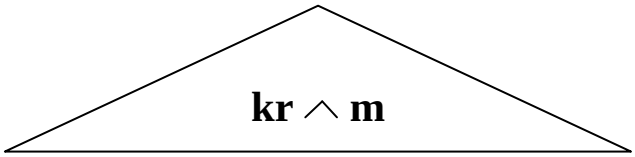

Figure 2.7: Syllable sound.

The best pronunciation of a word is achieved through storage of the whole word. This leads towards synthesis of the speech sequence (Figure 2.8).

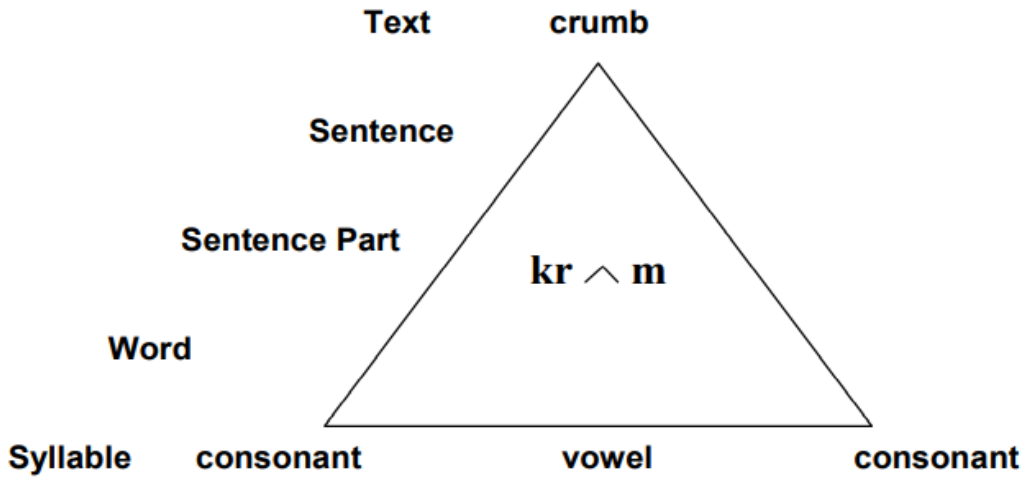

Figure 2.8 Word sound concatenation.

Frequency-dependent Sound Concatenation

Speech generation/output can also be based on a frequency-dependent sound concatenation, e.g., through a formant-synthesis. Formants are frequency maxima in the spectrum of the speech signal. Formant synthesis simulates the vocal tract through a filter.

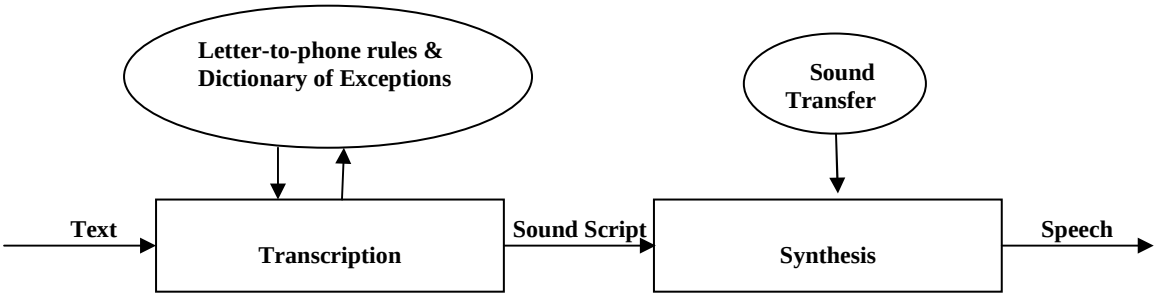

Figure 2.9: Components of a speech synthesis system with time-dependent sound concatenation.

In the first step, transcription is performed, in which text is translated into sound script. Most transcription methods work here with letter-to-phone rules and a Dictionary of Exceptions stored in a library. The generation of such a library in work-extensive, but using the interactive control the user it can be improved continuously. The user recognizes the formula deficiency in the transcription and improves the pronunciation manual.

In the second step, the sound script is translated into a speech signal. Time or frequency dependent concatenation can follow. While the first step is always a software solution, the second step is most often implemented with signal processors or even dedicated processors.

Speech Analysis

Speech analysis/input deals with the research areas shown in Figure 2.10:

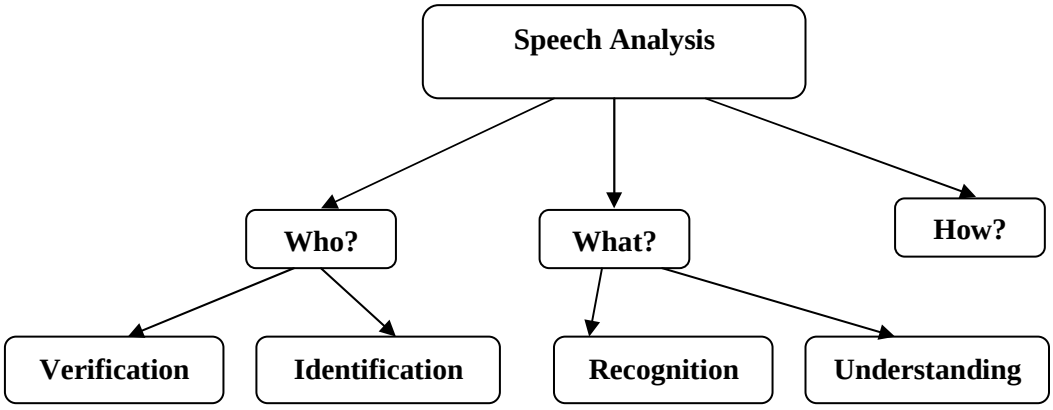

Figure 2.10: Research areas of speech analysis.

- Human speech has certain characteristics determined by a speaker. Hence, speech analysis can serve to analyze *who* is speaking, i.e. *to recognize a speaker* for his/her *identification* and *verification*.
  - The computer identifies and verifies the speaker using an acoustic fingerprint.
  - An acoustic fingerprint is a digitally stored speech *probe* (e.g., certain statement) of a person.

- Another task of speech analysis is to analyze what has been said, i.e., to recognize and understand the speech signal itself. Based on speech sequence, the corresponding text is generated. This can lead to a speech-controlled typewriter, a translation system or part of a workplace for the handicapped.
- Another area of speech analysis tries to research speech patterns with respect to *how* a certain statement was said. For example, a spoken sentence sounds differently if a person is *angry* or *calm*.

Speech Recognition

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to set of words.

Speech recognition is the process of converting spoken language to written text or some similar form.

Speech recognition is the foundation of human computer interaction using speech.

Speech recognition in different contexts:

- Dependent or independent on the speaker.
- Discrete words or continuous speech.
- Small vocabulary or large vocabulary.
- In quiet environment or noisy environment.

Natural Language Understanding (NLU) is a process of analysis of recognized words and transforming them into data meaningful to computer. Other words, NLU is a computer-based system that "understands" human language. NLU is used in combination with speech recognition.

Basic Terms and Concepts

- *Utterance (vocal sound)* is any stream of speech between two periods of silence.
- *Pronunciation* is what the speech engine thinks a word should sound like.
- *Grammars* define a domain (of words) within which recognition engine works.
- *Vocabulary (dictionary)* a list of words (utterances) that can be recognized by the speech recognition engine.
- *Training* is the process of adapting the recognition system to a speaker.
- *Accuracy* is the measure of recognizer's ability to correctly recognize utterances.
- *Speaker Dependence*
  - *Speaker dependent system* is designed for only one user (at the time).
  - *Speaker independent system* is designed for variety of speakers.

Types of Speech Recognition

Speech recognizers are divided into several different classes according to the type of utterance that they can to recognize:

- Isolated words,
- Connected words,
- Continuous speech (computer notation)
- Spontaneous (natural) speech
- Voice Verification
- Voice Identification

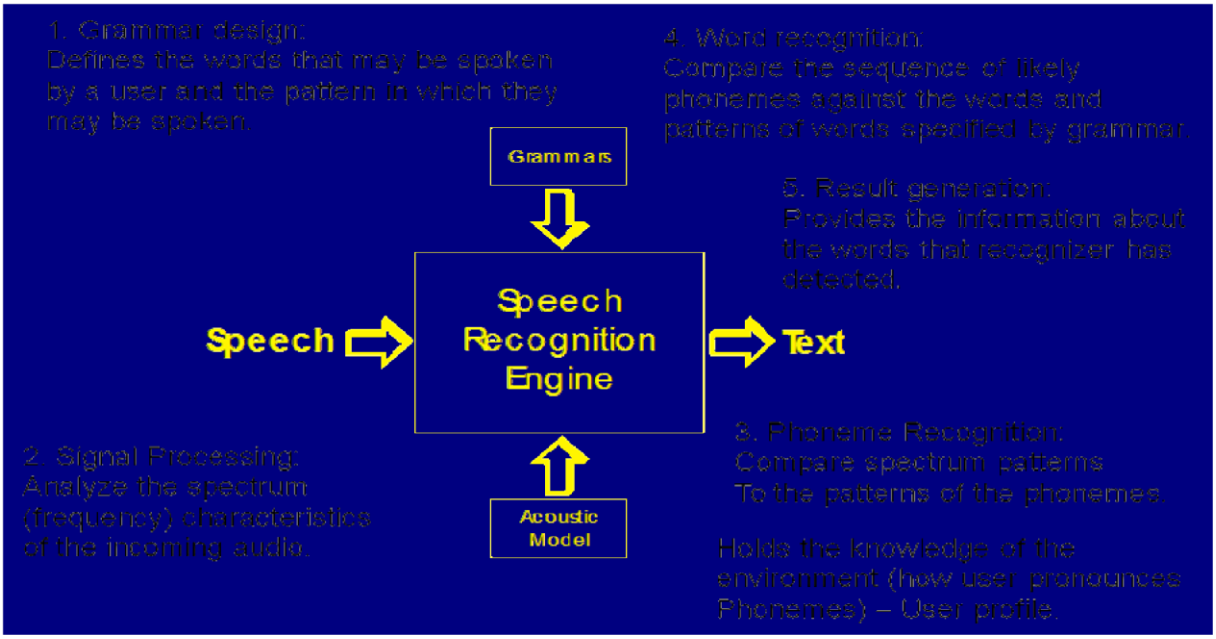

Figure 2.11: Speech Recognition System

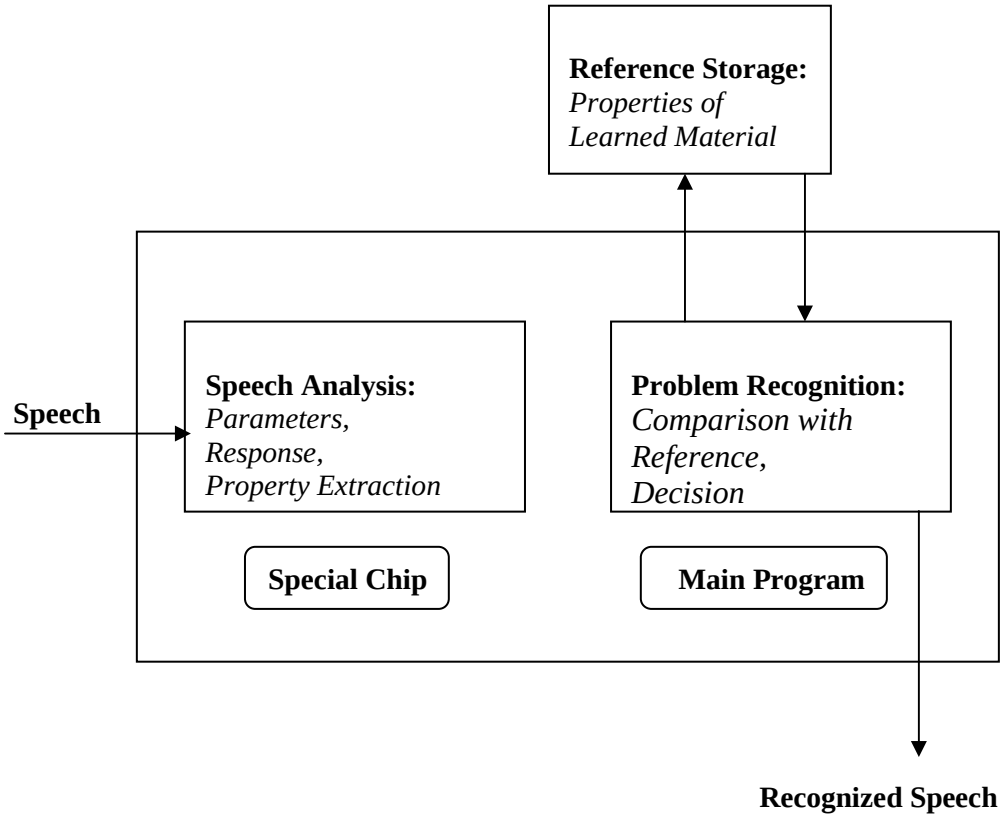

Figure 2.12: Speech recognition system: task division into system components, using the basic principle "Data Reduction Through Property Extraction."

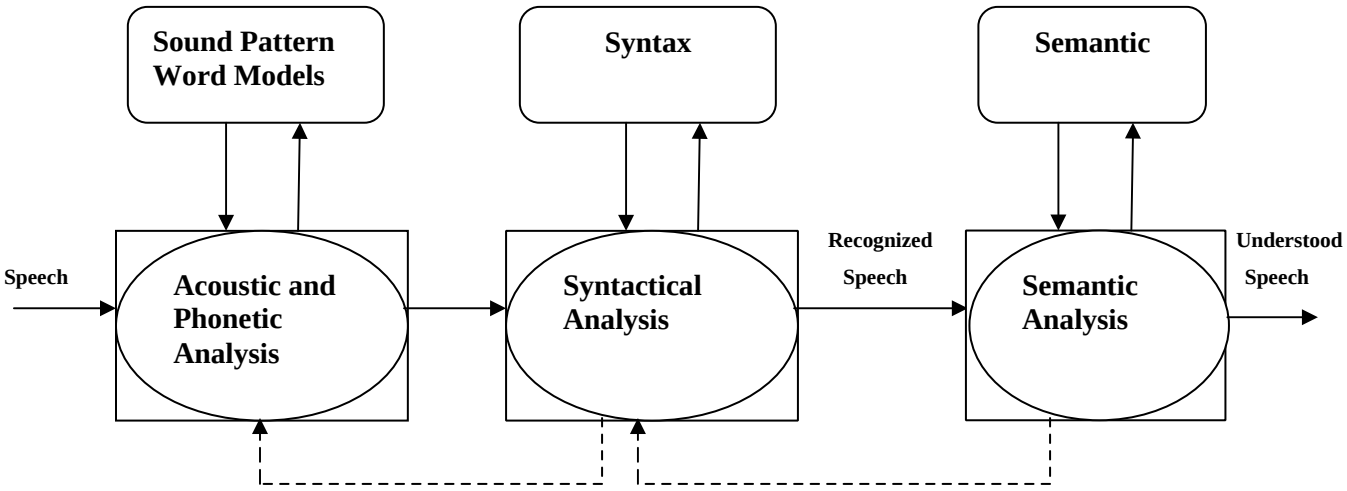

Figure 2.13: Components of speech recognition and understanding.
The first step, the principle is applied to a sound pattern and /or word model. An acoustical and phonetical analysis is performed.

The second step, certain speech units go through syntactical analysis; thereby, the errors of the previous step can be recognized. Very often during the first step, no unambiguous decisions can be made. In this case, syntactical analysis provides additional decision help and the result is a *recognized speech*.

The third step deals with the semantics of the previously recognized language. Here the decision errors of the previous step can be recognized and corrected with other analysis methods. The result of this step is an understood speech.