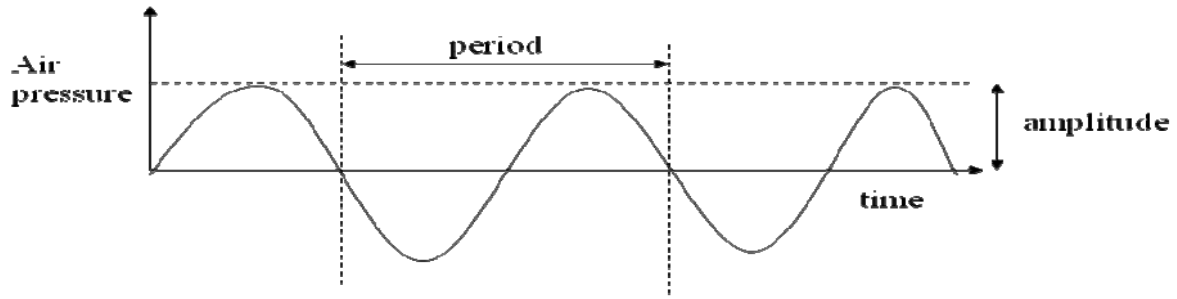
UNIT 2: SOUND /AUDIO SYSTEM (6 HRS.)

1.1 CONCEPTS OF SOUND SYSTEM;

Sound is a physical phenomenon by which matters are vibrated such as a violin string, guitar, and block of woods. As the matter vibrates, pressure variation is created in the surrounding air. This low and high air pressure is propagated through an air in a wave light motion. Which when reaches to human ear, a sound is heard. The propagations of oscillation which is called a wave form.



The pattern of the oscillation is called a waveform. The waveform repeats the same shape at regular intervals and this point is called a period. Since sound wave forms occur naturally, Sound waves are never perfectly smooth or uniformly periodic.

The wave form repeats the same shape at regular intervals which is called period. (one complete cycle). The natural sounds are not perfectly smooth or uniformly period. The sounds which have recognizable periodicity tend to be more musical than non-periodic sounds.

Example of periodic sounds sources are musical instrument, vowel sounds whistling wind, bird songs, etc

Non periodic sounds are coughs, sneezing, rousing water etc.

Sound vs Audio

- The key difference between sound and audio is their form of energy.
- Sound is mechanical wave energy (longitudinal sound waves) that propagate through a medium causing variation in pressure within the medium.
- Audio is made of electrical energy (analog or digital signals) that represent sound electrically.

To put it simply:

- Sound is vibrations through materials, the Action.
- Audio is the End result, the technology to hear sounds coming from natural or human-made sources.
- Sound is a continuous wave that travels through air by measuring the pressure level at a point.

- Microphone is sound field moves according to the varying pressure exerted on it, Transducer convert energy into voltage level (energy of another from – electrical energy)

1.2 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. Represents the number of periods in a second and is measured in hertz (Hz) or cycles per second.

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

Some of the frequency ranges are:

- Infra sound: 0 - 20 Hz
- Human audible sound: 20 Hz - 20KHz
- Ultra sound: 20KHz - 1GHz
- Hyper sound: 1GHz - 10THz

Human audible sound is also called audio or acoustic signals (waves). Speech is an acoustic signal produced by the humans.

1.3 AMPLITUDE

The amplitude of the sound is the measure of the displacement of the air pressure wave from its mean or quiescent state. The greater the amplitude, the louder the sound.

Subjectively heard as loudness. Measured in decibels.

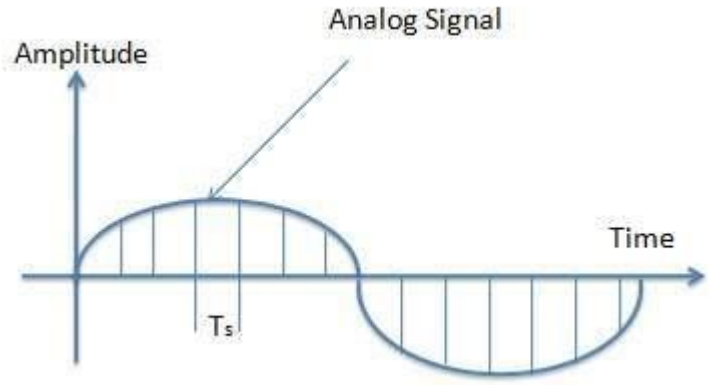
- 0 db - essentially no sound heard
- 35 db - quiet home
- 70 db - noisy street
- 120db - discomfort

1.4 COMPUTER REPRESENTATION OF SOUND

Sound waves are continuous while computers are good at handling discrete numbers. In order to store a sound wave in a computer, samples of the wave are taken. Each sample is represented by a number, the 'code'. This process is known as digitization.

Digitization is a process of converting the analog signals to a digital signal. There are three steps of digitization of sound. These are:

- **Sampling**
- **Quantization**
- **Encoding**



Sampling - Sampling is a process of measuring air pressure amplitude at equally spaced moments in time, where each measurement constitutes a sample.
A sampling rate is the number of times the analog sound is taken per second. A higher sampling rate implies that more samples are taken during the given time interval and ultimately, the quality of reconstruction is better.

To discretize the signals, the gap between the samples should be fixed. That gap can be termed as a **sampling period T_s**.

Sampling Frequency= $1/T_s = f_s$

Where,

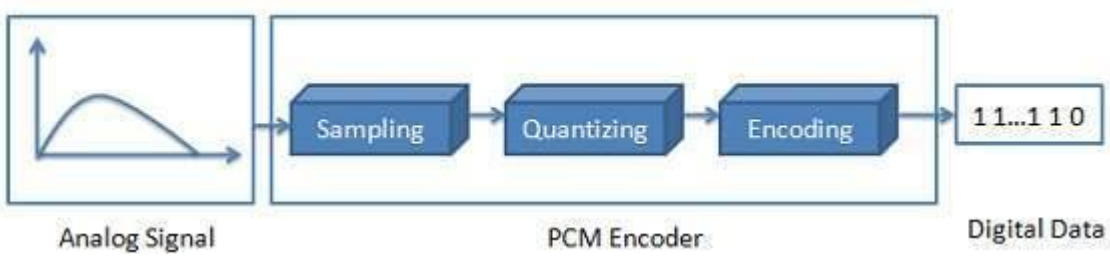
- Ts is the sampling time
- fs is the sampling frequency or the sampling rate

Sampling frequency is the reciprocal of the sampling period. This sampling frequency, can be simply called as **Sampling rate**. The sampling rate denotes the number of samples taken per second, or for a finite set of values.

The sampling rate is measured in terms of Hertz, Hz in short, which is the term for Cycle per second. A sampling rate of 5000 Hz (or 5kHz, which is more common usage) implies that mt uj vu8i 9ikuhree sampling rates most often used in multimedia are 44.1kHz(CD-quality), 22.05kHz and 11.025kHz.

Quantization - Quantization is a process of representing the amplitude of each sample as integers or numbers. How many numbers are used to represent the value of each sample known as sample size or bit depth or resolution. Commonly used sample sizes are either 8 bits or 16 bits. The larger the sample size, the more accurately the data will describe the recorded sound. An 8-bit sample size provides 256 equal measurement units to describe the level and frequency of the sound in that slice of time. A 16-bit sample size provides 65,536 equal units to describe the sound in that sample slice of time. The value of each sample is rounded off to the nearest integer (quantization) and if the amplitude is greater than the intervals available, clipping of the top and bottom of the wave occurs.

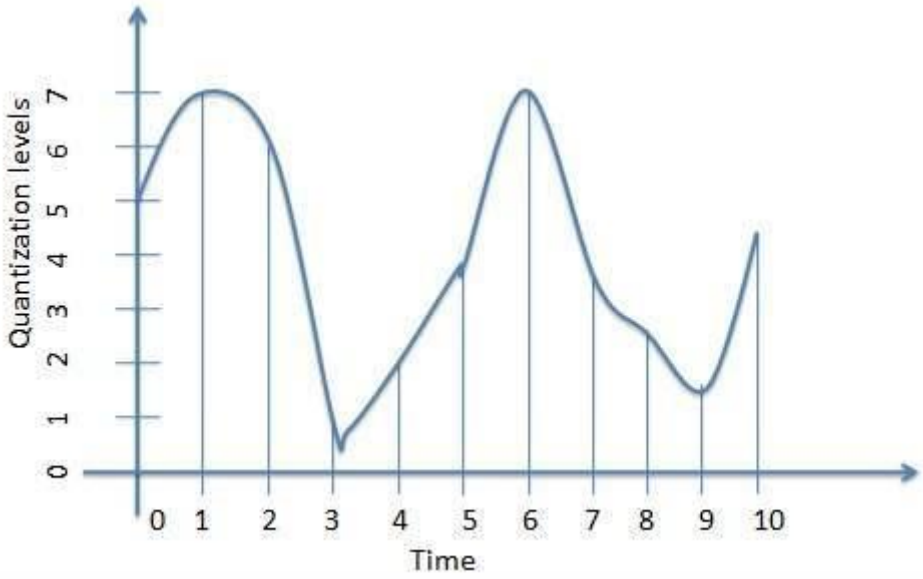
Encoding - Encoding converts the integer base-10 number to a base-2 that is a binary number. The output is a binary expression in which each bit is either a 1(pulse) or a 0(no pulse).



1.4.1 Quantization of Audio

Quantization is a process to assign a discrete value from a range of possible values to each sample. Number of samples or ranges of values are dependent on the number of bits used to represent each sample. Quantization results in stepped waveform resembling the source signal.

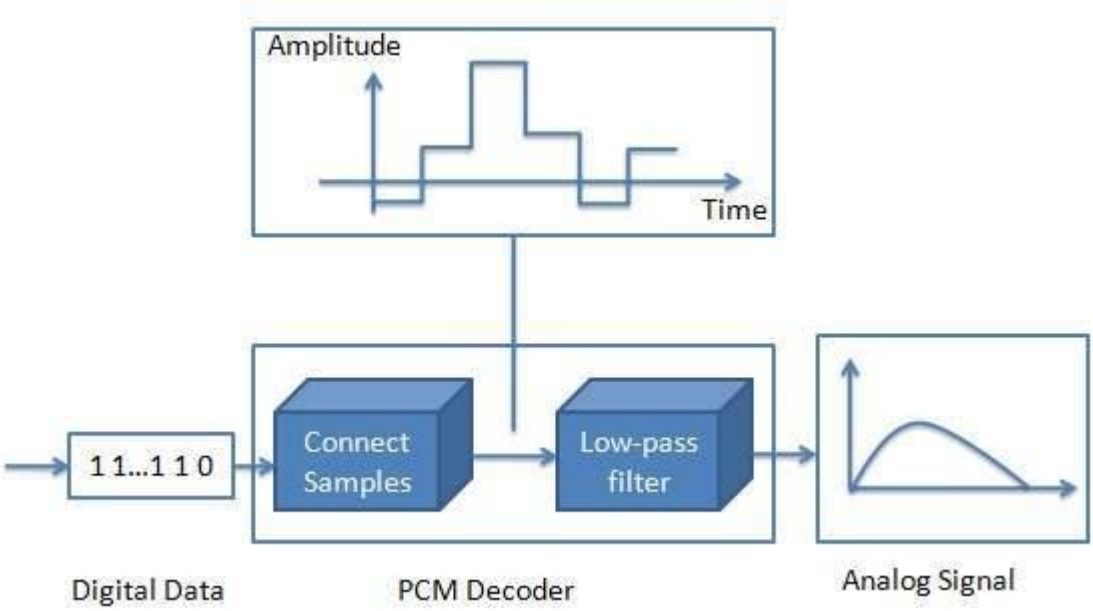
- **Quantization Error/Noise** - The difference between sample and the value assigned to it is known as quantization error or noise.
- **Signal to Noise Ratio (SNR)** - Signal to Ratio refers to signal quality versus quantization error. Higher the Signal to Noise ratio, the better the voice quality. Working with very small levels often introduces more error. So instead of uniform quantization, non-uniform quantization is used as companding. Companding is a process of distorting the analog signal in controlled way by compressing large values at the source and then expanding at receiving end before quantization takes place.



1.4.2 Transmission of Audio

In order to send the sampled digital sound/ audio over the wire that it to transmit the digital audio, it is first to be recovered as analog signal. This process is called de-modulation.

- **PCM Demodulation** - PCM Demodulator reads each sampled value then apply the analog filters to suppress energy outside the expected frequency range and outputs the analog signal as output which can be used to transmit the digital signal over the network.

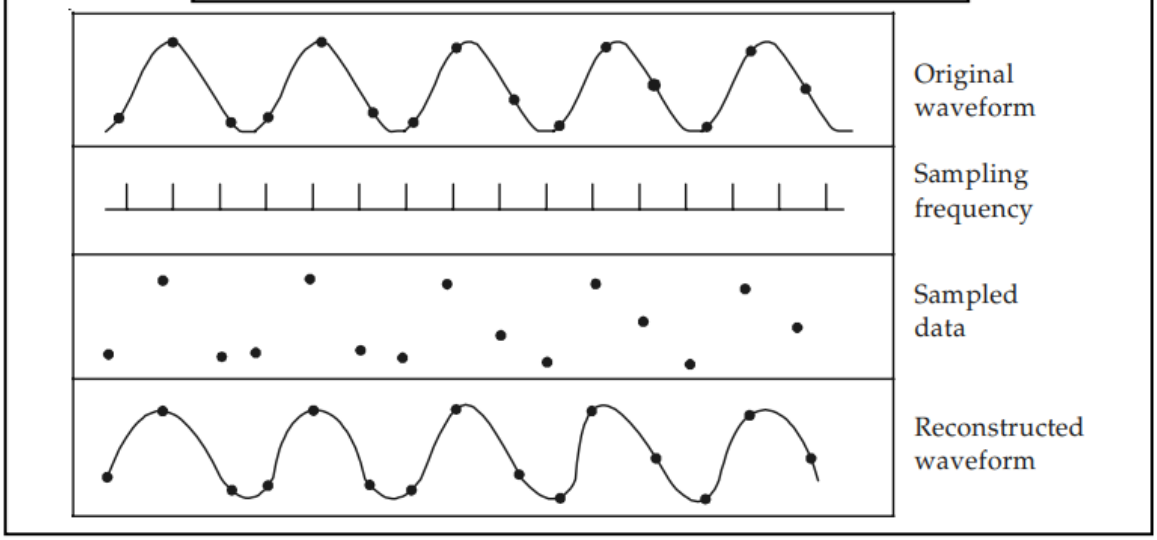


Sound Bit Depth

Sampling rate and sound bit depth are the audio equivalent of resolution and colour depth of a graphic image. A single bit rate and single sampling rate are recommended throughout the work. Bit depth depends on the amount of space in bytes used for storing a given piece of audio information. Higher the number of bytes higher is the quality of sound. Multimedia sound comes in 8-bit, 16-bit, 32-bit and 64-bit formats. An 8-bit has 28 or 256 possible values. A single bit rate and single sampling rate are recommended throughout the work. An audio file size can be calculated with the simple formula:

File Size in Disk = (Length in seconds) × (sample rate) × (bit depth/8 bits per byte).

Figure 3.1: It is difficult to reconstruct the original waveform if the sampling frequency is very low.



Bit Rate refers to the amount of data, specifically bits, transmitted or received per second. It is Notes comparable to the sample rate but refers to the digital encoding of the sound. It refers specifically to how many digital 1s and 0s are used each second to represent the sound signal. This means the higher the bit rate, the higher the quality and size of your recording. For

instance, an MP3 file might be described as having a bit rate of 320 kb/s or 320000 b/s. This indicates the amount of compressed data needed to store one second of music.

$$\text{Bit Rate} = (\text{Sample Rate}) \times (\text{Bit Depth}) \times (\text{Number of Channels})$$

1.4.3 Types of Digital Audio File Formats

There are many different types of digital audio file formats that have resulted from working with different computer platforms and software. Some of the better known formats include:

WAV

WAV is the Waveform format. It is the most commonly used and supported format on the Windows platform. Developed by Microsoft, the Wave format is a subset of RIFE RIFF is capable of sampling rates of 8 and 16 bits. With Wave, there are several different encoding methods to choose from including Wave or PCM format. Therefore, when developing sound for the Internet, it is important to make sure you use the encoding method that the player you're recommending supports.

AU

AU is the Sun Audio format. It was developed by Sun Microsystems to be used on UNIX, NeXT and Sun Sparc workstations. It is a 16-bit compressed audio format that is fairly prevalent on the Web. This is probably because it plays on the widest number of platforms.

RA

RA is Progressive Networks RealAudio format. It is very popular for streaming audio on the Internet because it offers good compression up to a factor of 18. Streaming technology enables a sound file to begin playing before the entire file has been downloaded.

AIFF

AIFF or AFF is Apple's Audio Interchange File Format. This is the Macintosh waveform format. It is also supported on IBM compatibles and Silicon Graphics machines. The AIFF format supports a large number of sampling rates up to 32 bits.

MPEG

MPEG and MPEG2 are the Motion Picture Experts Group formats. They are a compressed audio and video format. Some Web sites use these formats for their audio because their compression capabilities offer up to a factor of at least 14:1. These formats will probably become quite.

2 MUSIC AND SPEECH

The relationship between music and computers has become more and more important, especially considering the development of MIDI (Music Instrument Digital Interface) and its important contributions in the music industry today. The MIDI interface between electronic musical instruments and computers is a small piece of equipment that plugs directly into the computer's serial port and allows the trans- mission of music signals. MIDI is considered to be the most compact interface that allows full-scale output.

2.1 MIDI BASIC CONCEPTS

MIDI is a standard that manufacturers of electronic musical instruments have agreed upon. It is a set of specifications they use in building their instruments so that the instruments of different manufacturers can, without difficulty, communicate musical information between one another.

A MIDI interface has two different components:

Hardware connects the equipment. It specifies the physical connection between musical instruments, stipulates that a MIDI port is built into an instrument, specifies a MIDI cable (which connects two instruments) and deals with electronic signals that are sent over the cable.

A **data format** encodes the information traveling through the hardware. A MIDI data format does not include an encoding of individual samples as the audio format does. Instead of individual samples, an instrument- connected data format is used. The encoding includes, besides the instrument specification, the notion of the beginning and end of a note, basic frequency and sound volume. MIDI data allow an encoding of about 10 octaves, which corresponds to 128 notes.

The MIDI data format is digital; the data are grouped into MIDI messages. Each MIDI message communicates one musical event between machines. These musical events are usually actions that a musician performs while playing a musical instrument. The action might be pressing keys, moving slider controls, setting switches and adjusting foot pedals.

When a musician presses a piano key, the MIDI interface creates a MIDI message where the beginning of the note with its stroke intensity is encoded. This message is transmitted to another machine. In the moment the key is released, a corresponding signal (MIDI message) is transmitted again. For ten minutes of music, this process creates about 200 Kbytes of MIDI data, which is essentially less than the equivalent volume of a CD-audio coded stream in the same time.

If a musical instrument satisfies both components of the MIDI standard, the instrument is MIDI device (e.g. a synthesizer), capable of communicating with other MIDI devices through channels. The MIDI standard specifies 16 channels. A MIDI device (musical instrument) is mapped to a channel. Music data, transmitted through a channel, are reproduced at the receiver side with the synthesizer instrument. The MIDI standard identifies 128 instruments, including noise effects (e-g, telephone, air craft), with unique numbers. For example, 0 is for the Acoustic Grand Piano, 12 for the marimba, 40 for the violin, 73 for the flute, etc.

Some instruments allow only one note to be played at a time, such as the flute. 0ther instruments allow more than one note to be played simultaneously, such as the organ. The maximum number of simultaneously played notes per channel is a main property of each synthesizer. The range can be from 3 to 16 notes per channel. To tune a MIDI device to one or more channels, the device must be set to one of the MIDI reception modes. There are four 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.

2.2 COMMON MIDI DEVICES:

There are many types of MIDI devices. They play different role in making music.

Sound generators: It synthesizes the sound. It produces an audio signal that becomes sound when fed into a loud speaker. It can change quality of sound by varying the voltage oscillation of the audio. Sound generation is done in 2-ways:

1. Storing acoustic signals as MIDI data in advance
2. Creating acoustic signals synthetically

Microprocessor: Microprocessor communicates with the keyboard to know which notes the musician is playing. Microprocessor communicates with the control panel to know what commands the musician wants to sent to the microprocessor. The microprocessor then specifies note and sound commands to the sound generators (i.e. microprocessor sends and receives the MIDI message).

Keyboard: It affords the musician's direct control of the synthesizer. Pressing keys means signalling microprocessor what notes to play and how long to play them. Keyboard should have at least 5 octaves and 61 keys.

Control panel: Controls those function that are not directly concerned with notes and duration. Control panel includes a slider, a button and a menu.

Auxiliary controllers: Gives more control over the notes played on keyboard. Pitch bend and modulation are the 2 common variables on the synthesizer.

Memory: Stores patches for the sound generation and settings on the control panel. Drum machine: Specialize in percussion sounds and rhythms.

Master keyboard: Increases the quality of the synthesizer keyboard, Guitar Synthesizer, Drum pad controllers, Guitar controllers and many more.

Channel: MIDI supports upto 16 different channels. We can send off a MIDI event to any of those channels which are later synchronized by the sequencer.

Sequencer: Sequencer is the important MIDI device. It is used as storage server for generated MIDI data. It is also used as music editor. Musical data are represented in musical notes. Sequencer transforms the notes into MIDI message.

Track: It is a sequence of MIDI events.

2.3 MIDI MESSAGES:

MIDI messages are used by MIDI devices to communicate with each other and to determine what kinds of musical events can be passed from device to device.

Structure of MIDI messages:

- MIDI message includes a **status byte** and up to two **data bytes.**
- **Status byte**
  - The most significant bit of status byte is set to 1.
  - The 4 low-order bits identify which channel it belongs to (four bits produce 16 possible channels).
  - The 3 remaining bits identify the message.
- **Data Byte:** The most significant bit of data byte is set to 0.

Classification of MIDI messages:

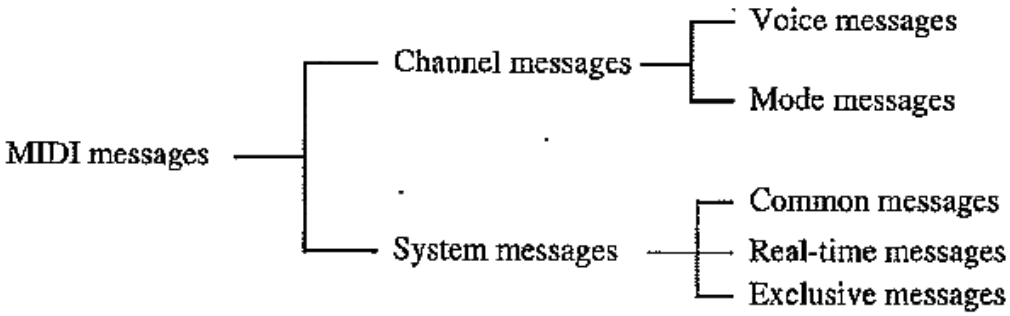


Figure 1: MIDI message taxonomy

Channel Message: Since, channel message are specified, the channel messages go only to specified devices. There are 2 types of channel messages:

- **Channel voice messages:** Sends actual performance data between MIDI devices, describing keyboard action, controller action and control panel changes. E.g. note on, Note off, channel pressure, control change etc.
- **Channel mode messages:** Determine the way that a receiving MIDI device responds to channel voice messages. E.g. local control, All note off, Omni mode off etc.

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

- **System real time messages:** These messages are short and simple (one byte). It synchronizes the timing of MIDI devices in performance. To avoid delay, they are sent in the middle of other messages. E.g. System reset, Timing clock i.e. MIDI clock etc.
- **System common messages:** Commands that prepare sequencer and synthesizer to play a song. E.g. song select, tune request etc.
- **System exclusive messages:** Allows MIDI manufacturers to create customized MIDI messages to send between their MIDI devices.

2.4 MIDI AND SMPTE TIMING STANDARDS

MIDI reproduces traditional note length using MIDI clocks, which are represented through timing clock messages. Using a MIDI cock, a receiver can synchronize with the clock cycles

of the sender. For example, a MIDI clock helps keep separate sequencers in the same MIDI system playing at the same tempo. When a master sequencer plays a song, it sends out a stream of Timing Clock' messages to convey the tempo to other sequencers. The faster the Timing Clock messages come in, the faster the receiving sequencer plays the song. To keep a standard timing reference, the MIDI specifications state that 24 MIDI clocks equal one quarter note.

As an alternative, the SMPTE timing standard (Society of Motion Picture and Television Engineers) can be used. The SMPTE timing standard was originally developed by NASA as a way to mark incoming data from different tracking stations so that receiving computers could tell exactly what time each piece of data was created. In the film and video version promoted by the SMPTE, the SMPTE timing standard acts as a very precise clock that stamps a time reading on each frame and fraction of a frame, counting from the beginning of a film or video. To make the time readings precise, the SMPTE format consists of **hours: minutes: seconds: frames: bits** (e.g., 30 frames per second), uses a 24-hour clock and counts from 0 to 23 before recycling to 0. The number of frames in a second differs depending on the type of visual medium. To divide time even more precisely, SMPTE breaks each frame into 80 bits (not digital bits). When SMPTE is counting bits in a frame, it is dividing time into segments as small as one twenty-five hundredth of a second.

Because many film composers now record their music on a MIDI recorder, it is desirable to synchronize the MIDI recorder with video equipment. A SMPTE synchronizer should be able to give a time location to the MIDI recorder so it can move to that location in the MIDI score (pre-recorded song) to start playback or recording But MIDI recorders cannot use incoming SMPTE signals to control their recording and playback. The solution is a MIDI/SMPTE synchronizer that converts SMPTE into MIDI, and vice versa. The MIDI/SMPTE synchronizer lets the user specify different tempos and the exact points in SMPTE timing at which each tempo is to start, change, and stop. The synchronizer keeps these tempos and timing points in memory. As a SMPTE video deck plays and sends a stream of SMPTE times to the synchronizer, the synchronizer checks the incoming time and sends out MIDI clocks at a corresponding tempo.

2.5 MIDI SOFTWARE:

The software applications generally fall into 4 major categories:

1. **Music recording and performance applications**: Provides function as recording of MIDI messages. Editing and playing the messages in performance.
2. **Musical notations and printing applications:** Allows writing music using traditional musical notation. User can play and print music on paper for live performance or publication.
3. **Synthesizer path editor and librarians:** Allows information storage of different synthesizer patches in the computer's memory and disk drives. Editing of patches in computer.
4. **Music education applications:** Teaches different aspects of music using the computer monitor, keyboard and other controllers of attached MIDI instruments.

Processing chain of interactive computer music systems:

- Sensing stage: Data is collected from controllers reading the gesture information from human performers on stage.
- Processing stage: Computer reads and interprets information coming from the sensors and prepares data for the response stage.
- Response stage: Computer ad some collection of sound producing devices share in realizing a musical output.

3 SPEECH GENERATION

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

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

- Voiced speech signals show during certain time intervals almost periodic behavior. Therefore, we can consider these signals as quasi-stationary signals for around 30 milliseconds.
- The spectrum of audio signals shows characteristic maxima, these maxima, called formants, occur because of resonances of the vocal tract.

Speech Generation:

An important requirement for speech generation is real-time signal generation. With such a requirement met, a speech output system could transform text into speech automatically without any lengthy pre-processing.

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.

3.1 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 the 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 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).

3.2 REPRODUCED SPEECH OUTPUT

The easiest method of speech generation/output is to use pre-recorded speech and play it back in a timely fashion. The speech can be stored as PCM (Pulse Code Modulation) samples. Further data compression methods, without using language typical properties, can be applied to recorded speech.

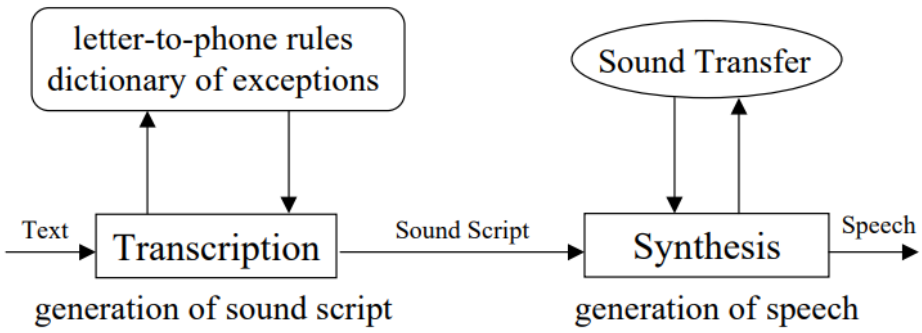
There are two way of speech generation/output performed by time-dependent sound concatenation and a frequency-dependent sound concatenation.

3.2.1 Time-dependent Sound Concatenation

- Individual speech units are composed like building blocks, e.g. phones
- Transitions between speech (coarticulation) units via allophones, i.e. variants of phones depending on previous and following phones
- Creations of syllables as building blocks for words and sentences
- Prosody, i.e. stress and melody course of a spoken phrase. Problem: Prosody is often context dependent.

3.2.2 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.
- Individual speech elements (e.g., phones) are defined through the characteristic values of the formants. Similar problems to the time-dependent sound concatenation exist here. The transitions, known as co-articulation, present the most critical problem. Additionally, the respective prosody has to be determined.
- New sound-specific methods provide a sound concatenation with combined time and frequency dependencies. Initial results show that new methods generate fricative and plosive sounds with higher quality.
- Human speech can be generated using a multi-pole lattice filter. The first four or five formants, occurring in human speech are modeled correctly with this filter type.
- Using speech synthesis, an existent text can be transformed into an acoustic signal. The typical components of a speech synthesis system with time-dependent concatenation:



Step 1: Generation of a Sound Script

Transcription from text to a sound script using a library containing (language specific) letter - to -phone rules. A dictionary of exceptions is used for word with a non -standard pronunciation.

Step 2: Generation of Speech

The sound script is used to drive the time - or frequency -dependent sound concatenation process.

3.2.3 Problem of speech synthesis:

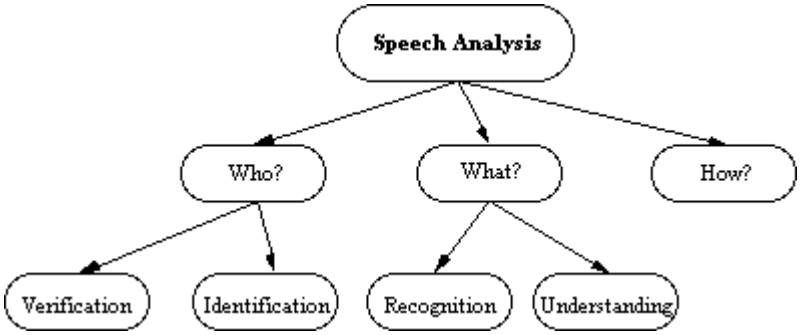
- Ambiguous pronunciation. In many languages, the pronunciation of certain words depends on the context.
- Example: 'lead'
- This is not so much of a problem for the German language
- It is a problem for the English language
- Anecdote by G. B. Shaw:
  - if we pronounce "gh" as "f" (example: "laugh")
  - if we pronounce "o" as "i" (example: "women")
  - if we pronounce "ti" as "sh" (example: "nation"), then why don't we write "ghoti" instead of fish?

4 SPEECH ANALYSIS

Purpose of Speech Analysis:

- Who is speaking: speaker identification for security purposes
- What is being said: automatic transcription of speech into text
- How was a statement said: understanding psychological factors of a speech pattern (was the speaker angry or calm, is he lying, etc)

The primary goal of speech analysis in multimedia systems is to correctly determine individual words (speech recognition).



Speech analysis is of strong interest for multimedia systems. Together with speech synthesis, different media transformations can be implemented. The primary goal of speech analysis is to correctly determine individual words with probability <=1. A word is recognized only with a certain probability. Here, environmental noise, room acoustics and a speaker's physical and psychological conditions play an important role.

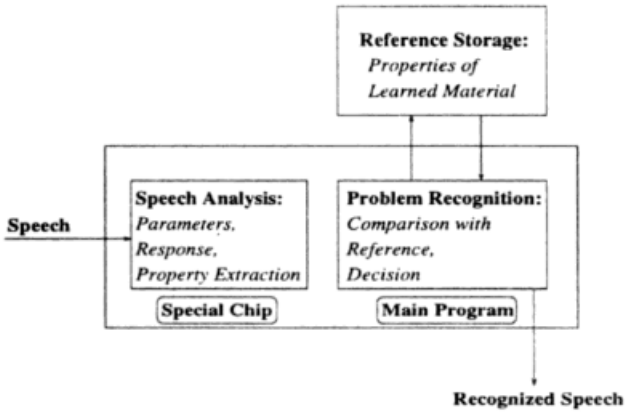


Figure 2: speech recognition system

The system which provides recognition and understanding of a speech signal applies this principle several times as follows:

- In the first step, the principle is applied to a sound pattern and/or word model. An acoustical and phonetical analysis is performed.
- In the second step, certain speech units go through syntactical analysis; thereby, the errors of the previous step can be recognized. Very often during t 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. Even today, this step is non-trivial to implement with current methods known in artificial intelligence and neural nets research. The result of this step is an understood speech.

These steps work mostly under the consideration of time and/or frequency-dependent sounds. The same criteria and speech units (formants, phones, etc.) are considered as in speech generation/output.

There are still many problems into which speech recognition research is being conducted:

- A specific problem is presented by room acoustics with existent environmental noise. The frequency-dependent reflections of a sound wave from walls and objects can overlap with the primary sound wave.
- Further, word boundaries must be determined. Very often neighboring words flow into one another.
- For the comparison of a speech element to the existing pattern, time normalization is necessary. The same word can be spoken quickly or slowly. However, the time axis cannot be modified because the extension factors are not proportional to the global time interval. There are long and short voiceless sounds (e.g., s, sh). Individual sounds are extended differently and need a minimal time duration for their recognition.

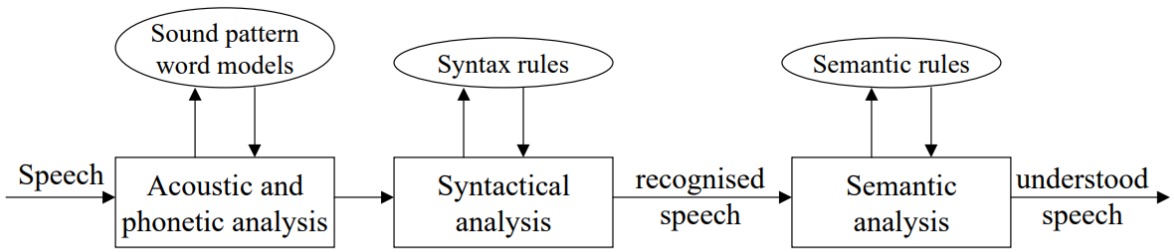


Figure 3: Components of speech recognition systems

Speech recognition systems are divided into speaker-independent recognition systems and speaker-dependent recognition systems. A speaker-independent system can recognize with the same reliability essentially fewer words than a speaker-dependent system because the latter is trained in advance. Training in advance means that there exists a training phase for the speech recognition system, which takes a half an hour. Speaker-dependent systems can recognize around 25,000 words; speaker independent systems recognize a maximum of about 500 words, but with a worse recognition rate. These values should be understood as gross guidelines. In a concrete situation, the marginal conditions must be known. (e.g., Was the measurement taken in a sound deadening room?, Does the speaker have to adapt to the system to simplify the time normalization?, etc.)

5 SPEECH TRANSMISSION

The area of speech transmission deals with efficient coding of the speech signal to allow speech/sound transmission at low transmission rates over networks. The goal is to provide the receiver with the same speech/sound quality as was generated at the sender side. This section includes some principles that are connected to speech generation and recognition.

- **Signal Form Coding:**

  This kind of coding considers no speech-specific properties and parameters. Here, the goal is to achieve the most efficient coding of the audio signal. The data rate of a PCM-coded stereo-audio signal with CD-quality requirements is: 1,411,200 bits/s Telephone quality, in comparison to CD-quality, needs only 64 Kbit/s. Using Difference Pulse Code Modulation (DPCM), the data rate can be lowered to 56 Kbits/s without loss of

quality. Adaptive Pulse Code Modulation (ADPCM) allows a further rate reduction to 32 Kbits/s.

- **Source Coding:**
  Parameterized systems work with source coding algorithms. Here, the specific speech characteristics are used for data rate reduction. Channel vo-coder is an example of such a parameterized system. The channel vo-coder is an extension of a sub-channel coding.

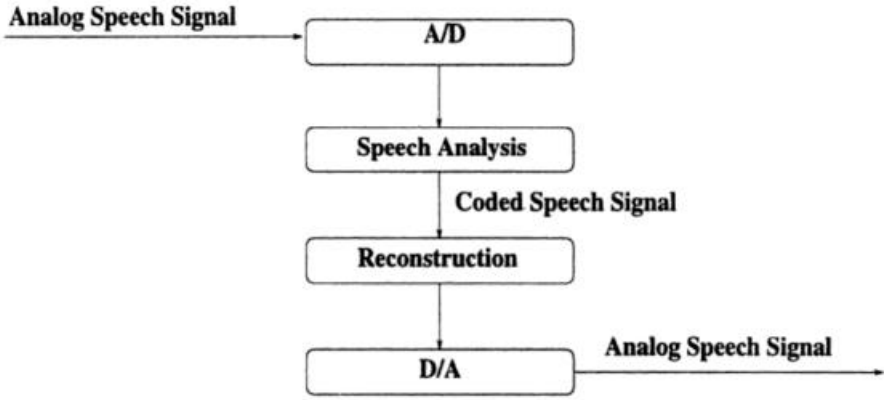


Figure 4: source coding

The signal is divided into a set of frequency channels during speech analysis because only certain frequency maxima are relevant to speech. Additionally, the differences between voiced and unvoiced sounds are taken into account. Voiceless sounds are simulated by the noise generator. For generation of voiced sounds, the simulation comes from a sequence of pulses. The rate of the pulses is equivalent to the a priori measured basic speech frequency. The data rate of about 3 Kbits/s can be generated with a channel vo-coder; however the quality is not always satisfactory.
Major effort and work on further data rate reduction from 64 Kbits/s to 6 Kbits/s is being conducted, where the compressed signal quality should correspond, after a decompression, to the quality of an uncompressed 64 Kbits/s signal.

- **Recognition/Synthesis Methods:**
  There have been attempts to reduce the transmission rate using pure recognition/synthesis methods. Speech analysis (recognition) follows on the sender side of a speech transmission system and speech synthesis (generation) follows on the receiver side.

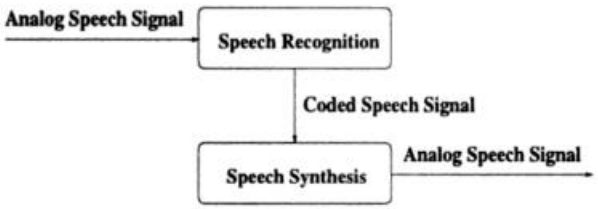


Figure 5: Recognition/synthesis system

Only the characteristics of the speech elements are transmitted. For example, the speech elements with their characteristics are the formants with their middle frequency

bandwidths. The frequency bandwidths are used in the corresponding digital filter. This reduction brings the data rate down to 50 bits/s. The quality of the reproduced speech and its recognition rate are not acceptable by today's standards.

- **Achieved Quality:**

The essential question regarding speech and audio transmission with respect to multimedia systems is how to achieve the minimal data rate for a given quality. The published function from Flanagan shows the dependence of the achieved quality of compressed speech on the data rate. One can assume that for telephone quality, a data rate of 8 Kbits/s is sufficient. Figure below shows the dependence of audio quality on the number of bits per sample value. For example, excellent CD-quality can be achieved with a reduction from 16 bits per sample value to 2 bits per sample value. This means that only 1/8 of the actual data needs to be transmitted.

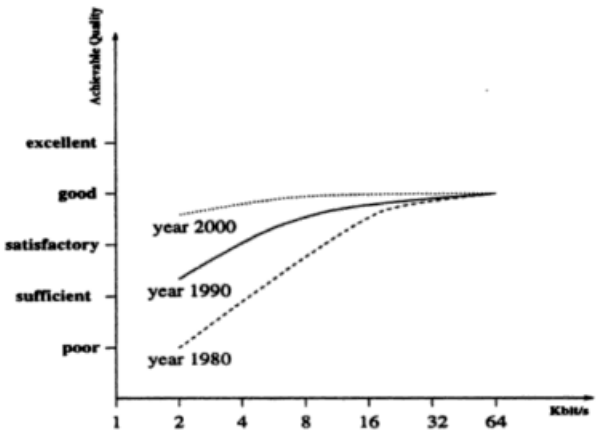
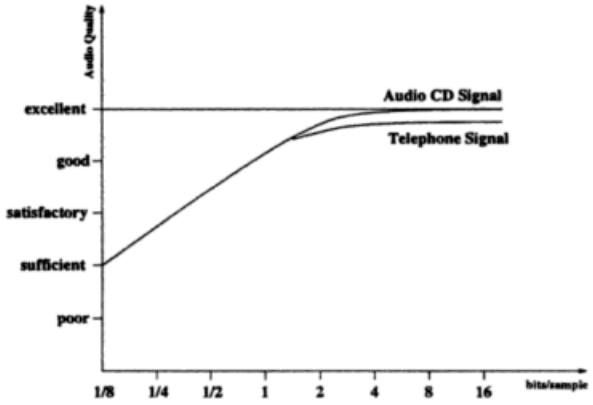


Figure 3.14: *Dependence of the achieved speech quality on the data rate.*



3 15: *Dependence of audio quality on the number of bits per sample value.*