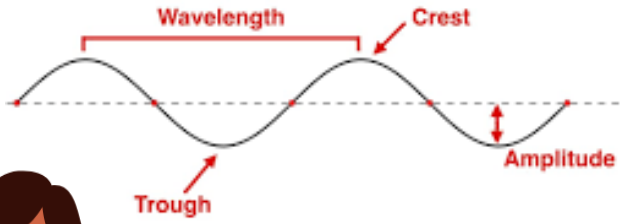
Unit 2: Sound / Audio System

2.1 Overview sound system
2.2 Producing digital audio
2.2 Music and speech
2.3 Speech Generation
2.4 Speech Analysis
2.5 Speech Transmission
2.6 Representation of audio files
2.7 Computer Music -MIDI
2.8 MIDI versus Digital Audio

#Past Questions

- **2023 Q3**: What is speech? Explain speech generation method.
- **2024 Q3**: What is sound? Explain the speech generation method.
- **2023 Q11**: What is MIDI? What features of MIDI makes it suitable for multimedia applications? Calculate the file size for 10 seconds of recording of stereo music at 44.1 kHz, 16-bit resolution.
- **2024 Q8**: Calculate the file size in bytes for a 30-second recording at 44.1 kHz, 8 bits resolution stereo sound.



Introduction to Sound

Definition:

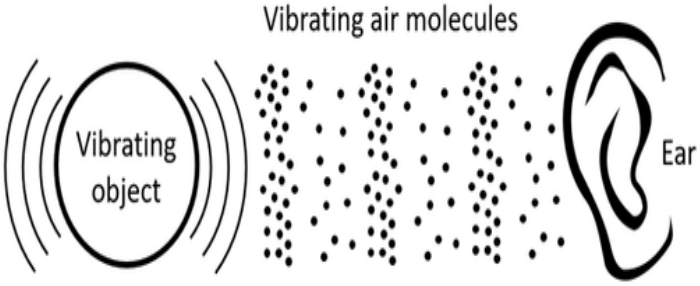
Sound is a form of energy produced by vibrations traveling through a medium, typically air. These vibrations create waves that propagate outward, carrying audible sensations to our ears.

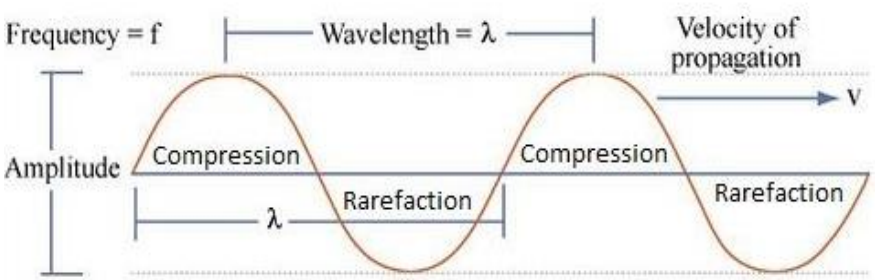
Units of Sound are

Decibels (dB)

Hertz (Hz)

Sound is a form of energy that originates from a vibration and travels as a wave through a medium, such as a solid, liquid, or gas, to our ears. When a vibrating object causes particles in the surrounding medium to move back and forth, these particles bump into others, transferring the energy and creating a sound wave. These waves continue to propagate until they reach an ear, where the vibrations cause the eardrum to vibrate, sending a signal to the brain that is interpreted as sound.

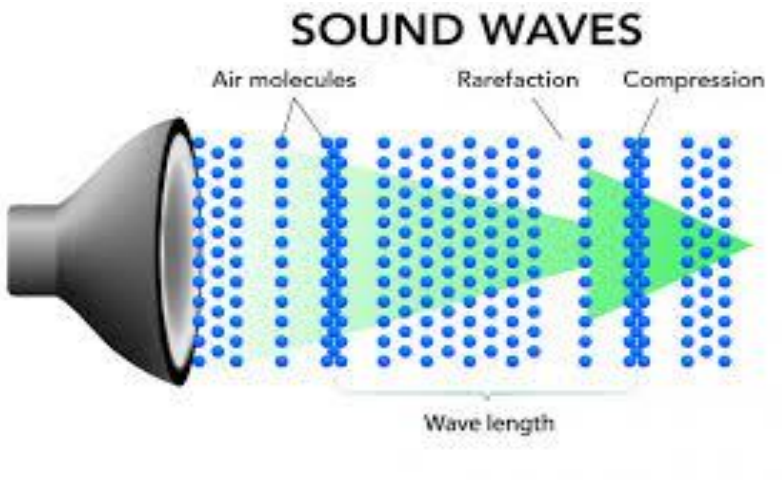




In multimedia, it includes elements like speech, music, sound effects, or ambient noise, which are captured, processed, and reproduced to convey information or evoke emotions.

Types of Sound

1. **Audible Sound**: 20 Hz – 20 kHz (human hearing range).
2. **Infrasound**: <20 Hz (earthquakes, elephant communication).
3. **Ultrasound**: >20 kHz (bats' echolocation, medical imaging).



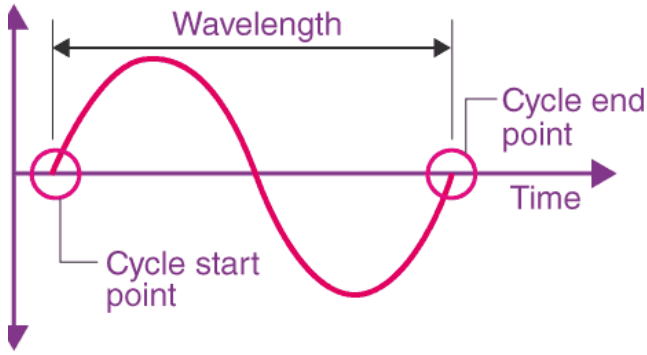
Nature of Sound Waves

- **Longitudinal Waves**: Molecules oscillate parallel to wave travel.
- **Compression**: Molecules packed closer.
- **Rarefaction**: Molecules spread apart.

Example: Clap your hands → compress air molecules → sound travels outward as wavefronts.

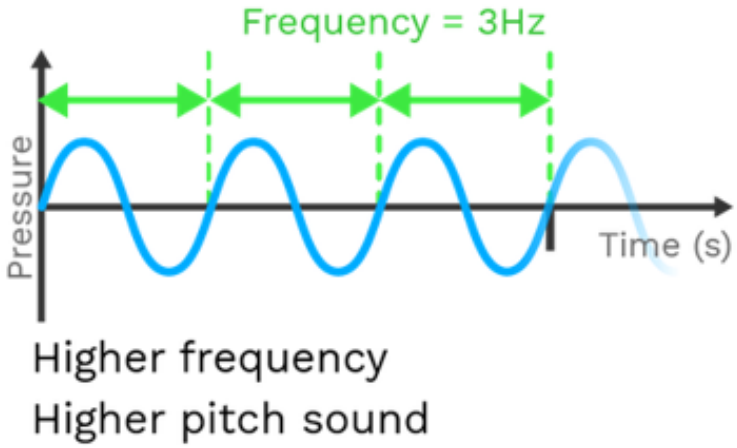
Characteristics of Sound

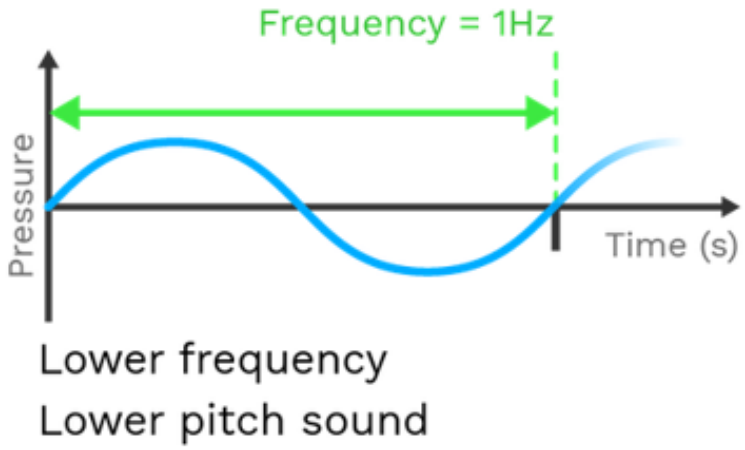
1. **Frequency (Pitch)**



   - Number of vibrations / wave cycle per second (measured in Hertz, Hz).
   - High frequency = high pitch (e.g., a whistle, female voice).
   - Low frequency = low pitch (e.g., a drum, male voice).
   - Humans can generally hear sounds with frequencies between 20 Hz and 20,000 Hz; frequencies below this range are infrasound, and those above are ultrasound.

Example: Male voice (~85–180 Hz) vs. female voice (~165–255 Hz).

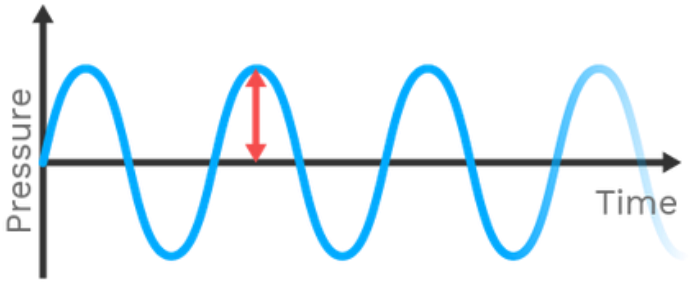




2. **Amplitude (Loudness)**

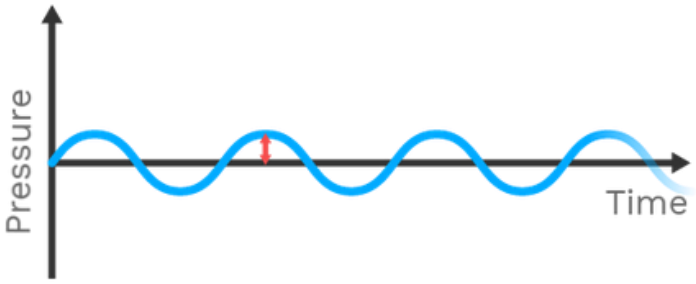
- The height or intensity of the wave (measured in decibels ,dB).
- Height of the wave = energy of vibration.
- Higher amplitude → louder sound and vice versa.

Example: Whisper vs. shouting.



Higher amplitude

Louder sound

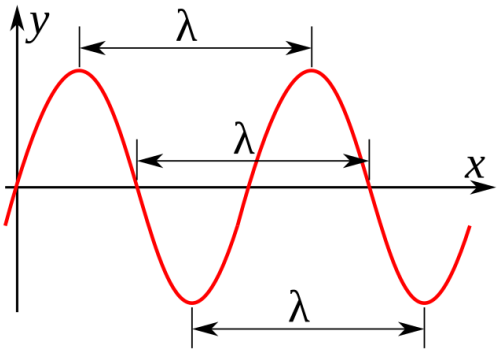


Lower amplitude

Quieter sound

3. **Wavelength**

- Distance between two successive compressions or rarefactions (crests or troughs)
- It is typically measured in units of distance, like meters (m) and denoted by (λ)
- Wavelength is inversely related to frequency i.e. Short wavelength → high frequency and vice versa.



Example: Violin (short wavelength) vs. bass guitar (long wavelength).

4. **Velocity (Speed of Sound)**

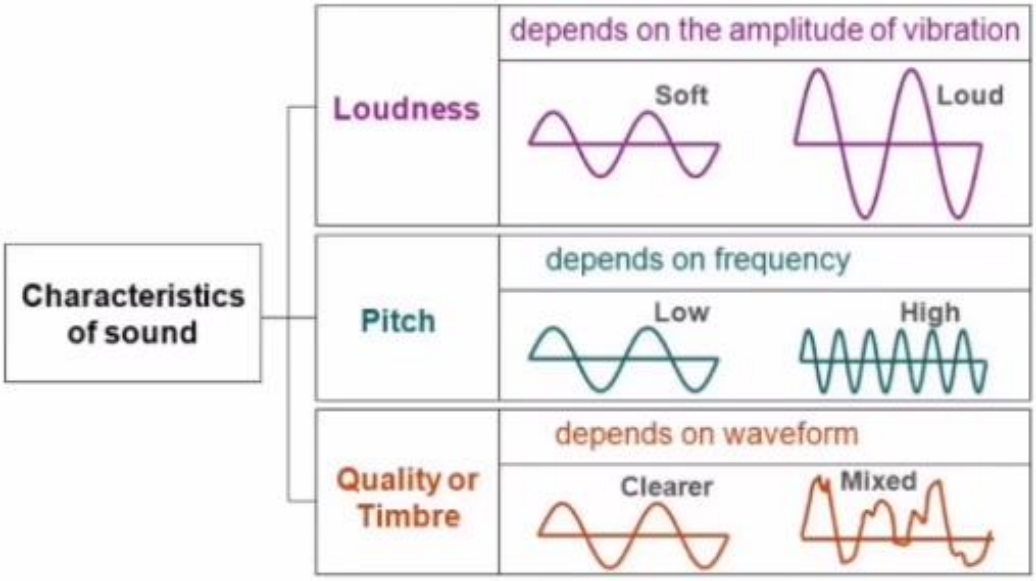
- The speed at which sound travels through a medium, which varies with the medium's properties (faster in solids, slower in gases).
- It can be calculated using the formula, $v=f\times\lambda$
- In air at room temp ≈ **343 m/s**, faster in water (~1500 m/s) and steel (~6000 m/s).

Example: Lightning is seen before thunder is heard because light travels much faster than sound.

5. **Timbre (Quality of Sound)**

- This is the quality or unique "color" of a sound that distinguishes sounds of same pitch and loudness from different sources.
- It's what allows you to distinguish a flute from a violin even when they play the same note at the same volume.
- Caused by overtones and harmonics.

Example: A note “C” played on piano vs. violin sounds different due to timbre.

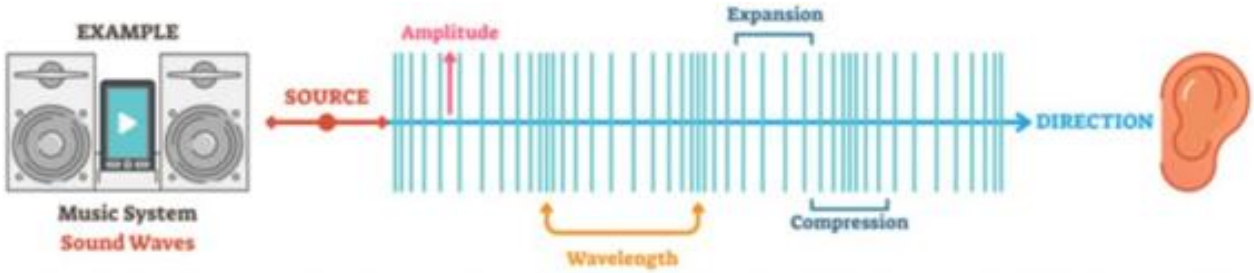


Types of Sound waves:

Sound waves can be classified into two main types based on their motion: transverse waves and longitudinal waves.

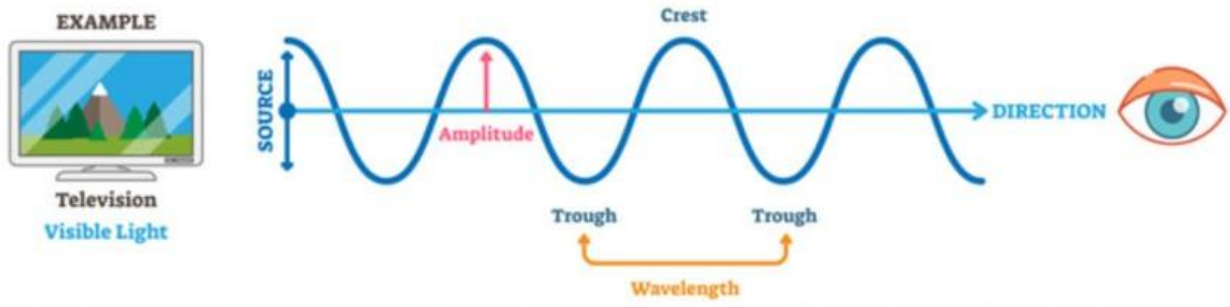
Longitudinal waves

- **Definition:** In a longitudinal wave, the particles of the medium vibrate parallel to the direction of the energy transfer.
- **Mechanism**: This creates alternating regions of high pressure (compressions) where particles are crowded together, and low pressure (rarefactions) where particles are spread apart.
- **Examples**: Sound waves traveling through air and water are primary examples of longitudinal waves.



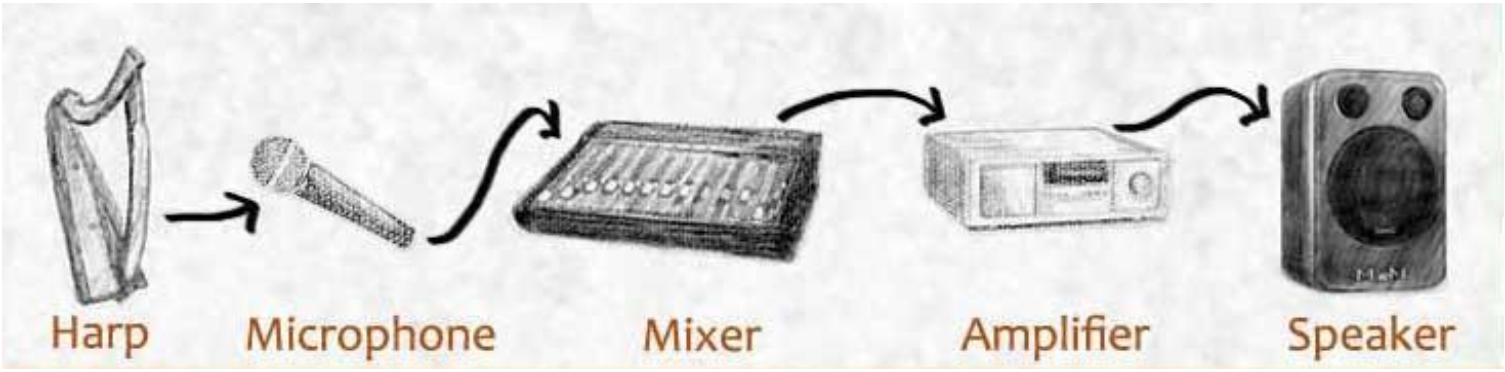
Transverse waves

- **Definition:** In a transverse wave, the particles of the medium vibrate perpendicularly (at right angles) to the direction of the wave's advance.
- **Key difference**: While sound can be transmitted as a transverse wave, it is not the primary form, and this only occurs under very specific circumstances. A key difference is that transverse waves create crests and troughs, not compressions and rarefactions.
- **Examples:** Light waves and water ripples are more common examples of transverse waves.



2.1 Overview of Sound System

A sound system is a collection of electronic audio devices(such as microphone, mixer, amplifier, speaker) that work together to capture, process, amplify, and reproduce sound.
A sound system is a complete ecosystem for capturing, processing, storing, transmitting, and reproducing sound. It bridges analog sound waves (physical vibrations) and digital data.



Core Components of a Sound System

1. **Sound Sources/ Harp:** Devices or instruments that create audio signals (e.g., A singer's voice, musical instruments, laptops, media players).
2. **Microphones:** Convert sound waves into electrical signals for processing.
3. **Audio Mixers:** Combine multiple audio signals, adjust levels, apply effects, and route audio to outputs.
4. **Amplifiers:** The signal coming from the mixer is not strong enough to power speakers. The amplifier boosts this weak signal to a level that is powerful enough to drive the speakers and produce audible sound.
5. **Speakers:** Convert amplified electrical signals back to sound waves audible to listeners.

2.2 Producing Digital Audio

Sound waves are analog(continuous) but computer only understand digital data (i.e. 0 & 1). So, sound must be converted from analog to digital form using a process called digitization.

Digital audio is created by **converting analog sound waves** (continuous vibrations) into **discrete digital signals** that computers can store, process, and play back.

Steps in Producing Digital Audio

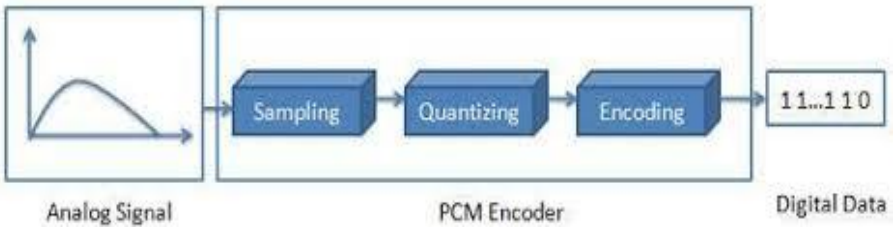
1. **Sound Capture**

- **Source**: Sound is generated by vibrations (e.g., voice, instruments) and captured using a microphone.
- **Conversion**: The microphone converts analog sound waves into electrical signals.

Example: Your voice entering a microphone is an analog signal and before computers can store it, it must be converted into digital form.

2. Analog-to-Digital Conversion (ADC)

This process has two main steps:



(a) Sampling

- The analog signal is measured (sampled) at regular intervals of time.
- The number of samples per second = **Sampling Rate** (measured in Hz).
- According to the **Nyquist theorem**, the sampling rate must be at least **twice the highest frequency** of the sound to avoid distortion (aliasing).

Common Sampling Rates:

- **44.1 kHz** → CD quality.
- **48 kHz** → Standard in video/audio production.
- **96 kHz / 192 kHz** → High-resolution audio.

Example: Human hearing is ~20 Hz – 20 kHz → CD uses 44.1 kHz to capture all frequencies.

(b) Quantization

- Each sample is assigned a numerical value representing its amplitude.
- Precision depends on **bit depth** (number of bits used per sample).

Bit Depth Examples:

- **8-bit** → 256 possible values (low quality, noisy).
- **16-bit** → 65,536 values (CD quality, clear sound).
- **24-bit** → 16.7 million values (studio quality).

Example: 16-bit audio captures soft and loud sounds more accurately than 8-bit.

3. Encoding and File Storage

- Formats like **PCM (Pulse Code Modulation)** store samples as binary.

Once digitized, audio can be stored in various file formats:

- **WAV / AIFF** → Uncompressed (high quality, large size).
- **MP3 / AAC** → Compressed with some quality loss.
- **FLAC / ALAC** → Lossless compression.

Example: A 1-second, 44.1 kHz/16-bit stereo clip = 44,100 samples × 2 channels × 2 bytes = 176.4 KB.

A 3-minute CD-quality (WAV) song ~30 MB, but as MP3 ~3–5 MB.

4. Digital Signal Processing (DSP)

After conversion, audio can be modified using DSP:

- Noise reduction
- Equalization (bass/treble adjustments)

- Reverb, echo, pitch correction

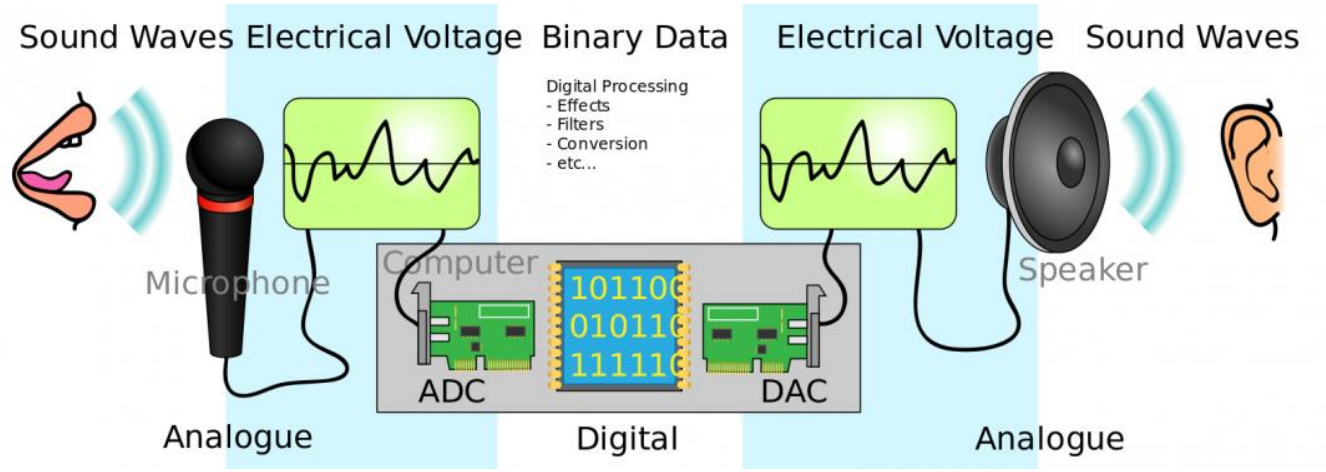
Example: Auto-Tune in music uses DSP to correct singers' pitch.

5. Digital-to-Analog Conversion (DAC)

For playback:

- The digital samples are **converted back** into continuous analog signals using a **DAC**.
- These signals are amplified and sent to **speakers/headphones**, producing audible sound.

Example: When you listen to Spotify, your phone's DAC converts the digital MP3 stream into analog sound.



Key Technical Aspects

- **Sampling Rate**: Higher rates (e.g., 96 kHz) capture more detail but increase file size.
- **Bit Depth**: Higher depths (e.g., 24-bit) offer greater dynamic range.
- **Compression**: Lossy (e.g., MP3) reduces size with some quality loss; lossless (e.g., FLAC) preserves original data.

Tools and Equipment

- **Hardware**: Microphones, audio interfaces, mixers, and speakers.
- **Software**: DAWs (Digital Audio Workstations) for production and editing.

2.3 Music and Speech

Music and speech are two fundamental forms of audio content processed and created in multimedia systems.

Music

- **Definition**: Organized sound with rhythm, melody, harmony, and dynamics intended to evoke aesthetic or emotional responses.
- **Characteristics**: Composed of notes, scales, beats, and instrumental or vocal elements arranged in time.

- **Digital Processing**: Involves recording instruments/voice, synthesizing sounds, mixing multiple tracks, and applying effects like reverb and equalization.
- **Applications**: Entertainment, background scores, games, therapy, and artistic expression.
- **Tools**: Digital Audio Workstations (DAWs) such as Ableton, FL Studio, Pro Tools allow composing, arranging, and producing digital music.

Speech

- **Definition:** Spoken language used for communication, consisting of phonetic sounds, words, and sentences conveying meaning.
- **Characteristics**: Contains variations in pitch, tone, volume, and speed, with specific linguistic patterns.
- **Digital Processing**: Includes speech recognition, synthesis, enhancement, noise reduction, and coding for transmission.
- **Applications**: Voice assistants, telecommunication, podcasts, language learning, and accessibility tools.
- **Techniques**: Speech synthesis (text-to-speech), speech recognition, audio compression for clarity and efficiency.

2.4 Speech Generation

Speech generation, also known as speech synthesis, is the process of creating artificial human speech that can be heard through a device or computer.

Speech generation is also known as text to speech (TTS), which means that it converts written or text input into spoken or audible output. TTS technology uses various algorithms and techniques to generate human-like speech from written text.

Speech Generation Methods

There are three main ways to generate speech:

1. **Reproduced speech output**
    - This method uses prerecorded speech (stored as PCM samples) and simply plays it back.
    - It is the simplest approach, but requires a lot of storage if many words/sentences are needed.
    - Data compression methods can be applied to save space.
2. **Time dependent sound concatenation**
    - Speech is produced by joining smaller units of sound together in proper sequence and timing.
    - There are four main levels of concatenation:

    a) Phone sound concatenation

        - Uses the smallest speech sound units (phonemes).
        - Difficult because transitions between phones are not smooth.

    b) Diphone sound concatenation

- Uses pairs of phones (the transition between one phone and the next).
- Reduces transition problems, producing smoother speech.

c) Syllable sound concatenation

- Instead of single sounds, whole syllables are stored and joined.
- Makes speech generation easier and more natural.

d) Word sound concatenation

- Entire words are stored and retrieved.
- Produces the best pronunciation, but needs a large storage space for all words.

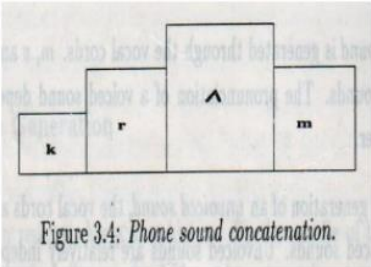


Figure 3.4: *Phone sound concatenation.*

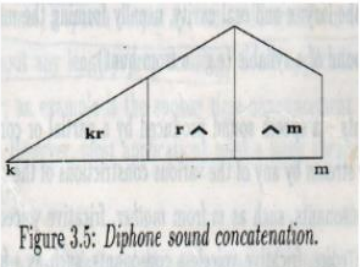


Figure 3.5: *Diphone sound concatenation.*

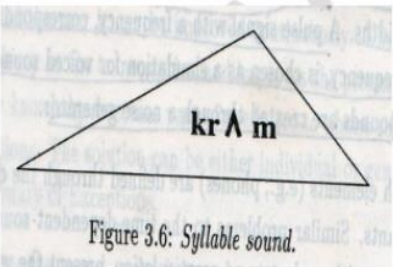


Figure 3.6: *Syllable sound.*

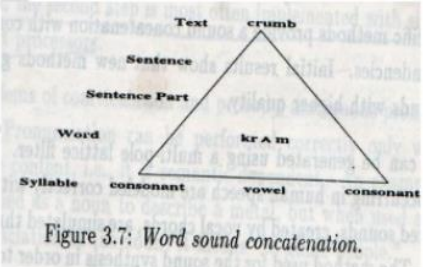


Figure 3.7: *Word sound concatenation.*

This figure show how different sound units can be combined to form meaningful speech.

3. **Frequency dependent sound concatenation (Formant synthesis)**

- Speech is generated by analyzing and synthesizing the frequency components of speech.
- Uses **formants** (frequency peaks in the speech spectrum) to simulate the vocal tract.
- A formant synthesis model passes sound through filters to generate natural-like speech.
- Advantage: Requires less storage compared to storing full words.

2.5 Speech Analysis

Speech analysis is the process of examining and interpreting a speech signal to extract meaningful information. It goes beyond simply recognizing words and can be used to understand a wide range of characteristics about the speaker and the content of their message.

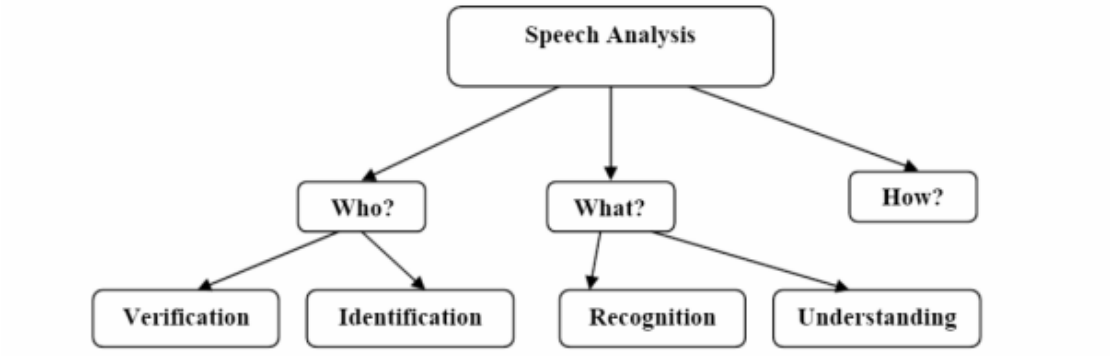


Figure 2.10: Research areas of speech analysis.

Human speech has certain characteristics determined by a speaker. Hence, speech analysis/input deals with the following three areas. Who, what and How?

(1) Who? : 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.

(2) What? : Another main task of speech analysis is to analyze what has been said i.e. to recognize and understand the speech signal itself.

(3) How? : Another area of speech analysis tries to research speech patterns with respect to how a certain statement was said.

2.6 Speech Transmission

Speech transmission is the **process of encoding, sending, and reproducing speech signals** over a communication network (like phone calls, VoIP, video conferencing).
The main goal is to deliver **clear, natural speech at low transmission rates** so that it uses less bandwidth but still sounds good.

Techniques of Speech Transmission

1. Pulse Code Modulation (PCM)

- A direct method of digitizing speech.
- The analog signal is sampled, quantized, and encoded into digital bits.
- Produces high-quality audio (used in CDs).
- **Data rate**: ~176,400 bytes/s for stereo.
- **Pros**: High fidelity.
- **Cons**: High bandwidth required.

2. Source Encoding

- Uses **compression** to reduce data rate.
- Exploits redundancies in the speech signal.
- Example: Removing unnecessary frequencies humans can't hear.
- **Result**: Smaller file size, faster transmission.
- Common in VoIP codecs (e.g., G.711, AMR, Opus).

3. Recognition–Synthesis Method

- Advanced method for very low data transmission.
- Steps:
    1. Analyze speech into smaller elements (phonemes, formants).
    2. Transmit only **codes** representing these elements.
    3. Receiver **re-synthesizes** the speech using a synthesizer.
- **Pros**: Very low data rate.
- **Cons**: Speech may sound less natural.

Components of a Speech Transmission System

1. **Speech Input (Microphone)** – captures analog speech.
2. **ADC (Digitization)** – converts speech to digital.
3. **Encoding/Compression** – reduces bit rate.
4. **Transmission Medium** – network (wired/wireless, internet, satellite).
5. **Decoding/Decompression** – reconstructs digital speech.

6. **DAC (Output)** – converts back to analog sound for playback.

2.7 Representation of Audio Files

Audio files are the **digital representations of sound**, stored on a computer in various formats.
The way audio is represented depends on **how it is captured, compressed, and stored**.

Types of Audio File Representation:

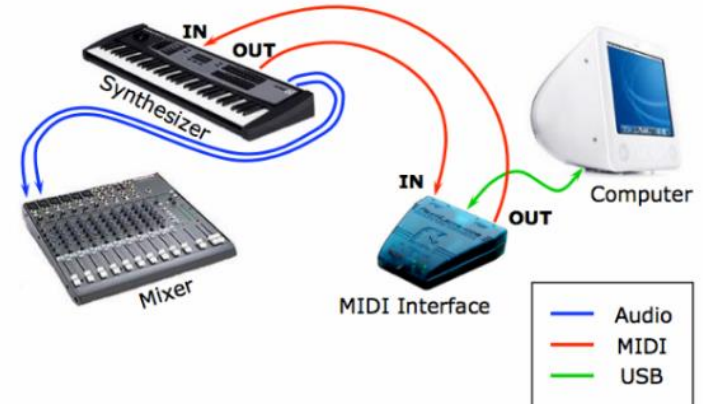
1. **Uncompressed:** These formats store the raw, digitized data directly without any compression. They offer the highest quality but result in very large file sizes. Examples include **WAV** and **AIFF**.
2. **Compressed:** These formats use algorithms to reduce file size.
    - **Lossless Compression** (e.g., **FLAC** and **ALAC**) reduces file size without discarding any data, so the original audio can be perfectly reconstructed.
    - **Lossy Compression** (e.g., **MP3** and **AAC**) removes some audio data that is considered less noticeable to the human ear to achieve a much smaller file size. This data is permanently lost.

2.8 Computer Music – MIDI(Musical Instrument Digital Interface):

MIDI is a **standard protocol** that allows **musical instruments, computers, and other devices** to communicate with each other.
Unlike digital audio (which stores actual sound waves), MIDI only stores **instructions** about music (what note, duration, instrument, etc.). →The receiving device (synthesizer, sound card, or software) generates the actual sound.

Why MIDI in Computer Music?

- Digital sound recording (WAV, MP3) can be **large** and **hard to edit**.
- MIDI is **compact** and **highly editable**.
- Widely used in **music production, multimedia, gaming, and live performance**.



Example:

- A **WAV file** of a piano note stores thousands of samples.

A **MIDI file** simply says: *"Play Piano Note C4 at 100 velocity for 2 sec."* → The synthesizer generates sound accordingly.

MIDI Hardware

MIDI devices communicate using **MIDI ports**:

- **IN** → Receives MIDI data.
- **OUT** → Sends MIDI data.

- **THRU** → Passes data to the next device.

Components of a Typical MIDI Synthesizer:

- **Sound Generator** → Produces audio signals.
- **Microprocessor** → Processes input and controls sound generator.
- **Keyboard** → Allows musician to play notes.
- **Control Panel** → Adjusts settings like tone, effects.
- **Auxiliary Controllers** → Pedals, sliders, wheels for modulation.
- **Memory** → Stores sounds (patches) and performance data.

MIDI Messages

MIDI communicates through **messages**:

(a) Channel Messages (to specific instruments):

- **Channel Voice Messages**: Sent actual performance data between MIDI devices, describing keyboard action, controller action and control panel changes. They describe music by defining pitch, amplitude, timbre, duration and other sound qualities. Examples of channel voice messages are note on, note off, channel pressure, control change etc.
- **Channel Mode Messages**: It determine the way that a receiving MIDI device responds to channel voice messages. They set the MIDI channel receiving modes for different MIDI devices. Examples of such messages are local control, all notes, omni mode off etc.

(b) System Messages (to all devices):

- **System Real-Time Messages**: They are very short and simple, consisting of only one byte. They carry extra data with them. Example, System Reset, Timing clock (MIDI clock) etc.
- **System Common Messages**: They are commands that prepare sequencers and synthesizers to play a song. Examples, Song select, Tune Request etc.
- **System Exclusive Messages**: They allow MIDI manufactures to create customized MIDI messages to send between MIDI devices.

Example: Pressing a key on a MIDI keyboard sends a **Note On message** with pitch (C4), velocity (80), and duration until key release.

MIDI Devices in Multimedia

- **Synthesizers**: Generate sound from MIDI commands.
- **Sequencers**: Record, edit, and play back MIDI data (computer software acts as sequencer).
- **Sound Cards**: Built-in MIDI synthesizers in PCs.
- **Drum Machines / Samplers**: Trigger drum sounds or samples.

Advantages of MIDI

- Very **small file size** (a song in MIDI = few KB).

- **Easily editable** (change instrument, pitch, tempo instantly).
- Supports **multi-instrumental performance** (up to 16 channels).
- Compatible across devices (keyboard, computer, DAW software).

Limitations of MIDI

- Does **not contain actual sound** → quality depends on playback device.
- Cannot capture subtle nuances of live audio (e.g., natural voice).
- Not ideal for **final music distribution** (better for editing/production).

Examples of MIDI in Use

- 🎶 Music Production → DAWs like FL Studio, Logic Pro, Ableton use MIDI for sequencing.
- 🎮 Gaming → Early games (Mario, Doom) used MIDI for background music due to small size.
- 🎤 Live Performances → Musicians use MIDI keyboards to control synthesizers.
- 🎬 Multimedia → Background scoring in animation and presentations.

How it Works:

1. A **MIDI Keyboard** or **MIDI Controller** sends performance data (note pitch, velocity, modulation, etc.).
2. The data is sent to a **computer or Digital Audio Workstation (DAW)** for recording, editing, or playback.
3. The **MIDI Interface** processes and routes MIDI data to the appropriate device.
4. A **Sound Module or Synthesizer** generates actual audio based on the MIDI instructions.
5. The final sound is played through **speakers or headphones**.

2.9 MIDI vs. Digital Audio

Digital Audio	MIDI
• Digital representation of physical sound waves	• Abstract representation of musical sounds and sound effects
• File size is large if without compression	• MIDI files are much more compact
• Quality is in proportion to file size	• File size is independent to the quality
• More software available	• Much better sound if the sound source is of high quality
• Play back quality less dependent on the sound sources	• Need some music theory
• Can record and play back any sound including speech	• Cannot generate speech

Feature	MIDI 🎹	Digital Audio 🎧
Contains actual sound	❌ No	✅ Yes
File size	Very Small (KBs)	Large (MBs)
Flexibility	Highly editable	Limited
Quality depends on	Synthesizer/sound card	Sampling rate & bit depth
Example	MIDI Keyboard file	WAV/MP3 recording

File Size calculation for Audio Recording

Audio files can be either mono or stereo.

- **Mono files** contain a single audio channel, meaning the same sound is heard from all speakers.
- **Stereo files** contain two separate audio channels—left and right—which provide a sense of direction and depth in the sound.

Formula for Audio File Size

$$\text{File Size (bytes)} = \text{Sampling Rate} \times \text{Bit Depth} \times \text{Number of Channels} \times \text{Duration (seconds)} \div 8$$

- **Sampling Rate (Hz)** → how many samples per second (e.g., 44,100 Hz for CD).
- **Bit Depth (bits per sample)** → resolution of each sample (e.g., 16-bit, 24-bit).
- **Number of Channels** → 1 = mono, 2 = stereo.
- **Duration** → length of the recording in seconds.
- Divide by 8 to convert **bits → bytes**. To convert bytes to kilobytes (KB), divide by 1,024. To convert to megabytes (MB), divide by 1,024 again

Example 1: CD Quality Audio

- Sampling rate = **44,100 Hz**
- Bit depth = **16 bits**
- Channels = **2 (stereo)**
- Duration = **60 seconds (1 min)**

➲ File size = $44,100 \times 16 \times 2 \times 60 \div 8 = 10,584,000 \text{ bytes} \approx 10.1 \text{ MB}$

So, **1 minute of CD-quality stereo audio ≈ 10 MB** (uncompressed WAV).

Example 2: Mono Voice Recording

- Sampling rate = **8,000 Hz** (telephone quality)
- Bit depth = **8 bits**
- Channels = **1 (mono)**
- Duration = **60 seconds**

$$8,000 \times 8 \times 1 \times 60 \div 8 = 480,000 \text{ bytes} \approx 0.46 \text{ MB}$$

So, **1 minute of mono 8 kHz speech ≈ 0.5 MB**.

Summary: Higher sampling rate + higher bit depth + more channels → larger file size. That's why WAV/AIFF files are big, while MP3 compresses them.

Q. Calculate the file size for 10 seconds of recording of stereo music at 44.1 kHz, 16-bit resolution.
Soln:
Given here,

- Sample rate = $44,100$ Hz
- Bit depth = 16 bits
- Channels = 2
- Duration = 10 sec

$$\text{File size} = \frac{44,100 \times 16 \times 2 \times 10}{8} = 1,764,000 \text{ bytes}$$

Convert to MB

$$\frac{1,764,000}{1024 \times 1024} \approx 1.68 \text{ MB}$$

File size for some common sampling rates and resolutions

Sampling Rate	Resolution	Stereo /Mono	Size for for 1 Min.	Comments
44.1KHz	16-bit	Stereo	10.5MB	CD-quality recording
44.1KHz	16-bit	Mono	5.25MB	A good trade-off for high-quality recordings of mono sources such as voice-overs
44.1KHz	8-bit	Stereo	5.25MB	Achieves highest playback quality on low-end devices such as most of the sound cards
44.1KHz	8-bit	Mono	2.6MB	An appropriate trade-off for recording a mono source
22.05KHz	16-bit	Stereo	5.25MB	Darker sounding than CD-quality recording because of the lower sampling rate
22.05KHz	16-bit	Mono	2.5MB	Not a bad choice for speech, but better to trade some fidelity for a lot of disk space by dropping down to 8-bit
22.05KHz	8-bit	Stereo	2.6MB	A very popular choice for reasonable stereo recording where full bandwidth playback is not possible
22.05KHz	8-bit	Mono	1.3MB	A thinner sound than the choice just above, but very usable
11KHz	8-bit	Stereo	1.3MB	At this low a sampling rate, there are few advantages to using stereo
11KHz	8-bit	Mono	650K	In practice, probably as low as you can go and still get usable results
5.5KHz	8-bit	Stereo	650K	Stereo not effective
5.5KHz	8-bit	Mono	325K	About as good as a bad telephone connection