Multimedia Systems
Sound and Audio

Er. Narayan Sapkota, M.Sc.
narayan.sapkota@eemc.edu.np

Everest Engineering College (Senior Lecturer)
Kathmandu University (Visiting Faculty)

November 24, 2024

Syllabus

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Table of Contents

1. Introduction

2. Basic Sound Concepts
   - Representation
   - Formats

3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software

4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Introduction

Audiology is the discipline focused on manipulating acoustic signals that are perceptible to humans. Key areas within this field include psychoacoustics, music, the MIDI (Musical Instrument Digital Interface) standard, as well as speech synthesis and analysis. In multimedia applications, audio is commonly used in the form of music and speech, with voice communication playing a particularly important role in distributed multimedia systems.

This chapter introduces basic audio signal technologies and the MIDI standard, along with various enabling methods such as speech synthesis, speech recognition, and speech transmission.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Basic Sound Concepts (1)

Sound is a physical phenomenon resulting from the vibration of a material, such as a violin string or a piece of wood. This vibration creates pressure fluctuations in the surrounding air, producing pressure waves that propagate outward. The oscillating pattern of these waves, known as a waveform, is illustrated in Figure 1. We perceive sound when these pressure waves reach our ears.

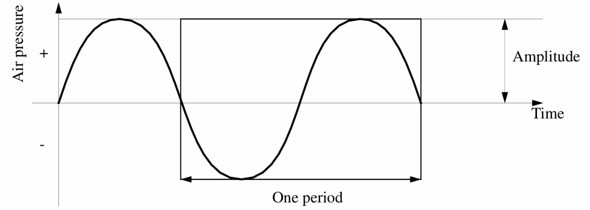

Figure 1: Pressure wave oscillation in the air

Basic Sound Concepts (2)

Sound is a physical phenomenon, while audio is an electronic representation of sound.

Sound waves typically recur at regular intervals, or periods, although their natural origins mean they are never perfectly uniform or periodic. When a sound has a noticeable periodicity, it is often classified as music, whereas sounds without this regular pattern are simply referred to as noise.

Examples of periodic sounds include those produced by musical instruments, vocalizations, wind, or birdsong. Non-periodic sounds, on the other hand, include drum beats, coughing, sneezing, or the sounds of water flowing or murmuring.

Fundamental properties/characteristics of sound are frequency and amplitude.

Basic Sound Concepts (3)

1. **Frequency:** A sound's frequency is the reciprocal of its period, representing the number of cycles per second, measured in hertz (Hz) or kilohertz (kHz). Sound frequencies are classified as:

   Infrasonic: 0–20 Hz

   Audiosonic: 20 Hz–20 kHz

   Ultrasonic: 20 kHz–1 GHz

   Hypersonic: 1 GHz–10 THz

   The audiosonic range (20 Hz–20 kHz) is essential for multimedia, encompassing signals like speech and music. Audio in this range is also

Basic Sound Concepts (4)

known as acoustic signals.

Noise, defined as sound without specific function, can be added to speech and music, though the term is flexible and can include unintelligible language as well. Wavelength is the distance travelled in one cycle. The wavelength (λ) is given by: $\lambda = \frac{v}{f}$; v: velocity and f: frequency

2. **Amplitude:** A sound has a property called amplitude, which humans perceive subjectively as loudness or volume. The amplitude of a sound is a measuring unit used to deviate the pressure wave from its mean value (idle state).

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Representation (1)

To represent the continuous curve of a sound wave on a computer, the computer must first measure the wave's amplitude at regular time intervals. It then uses these measurements to create a sequence of sampling values, commonly called samples. Figure 2 illustrates the period of a digitally sampled wave. The component that converts an audio signal into a sequence of digital samples is known as an ADC. For the reverse process, a DAC is used.

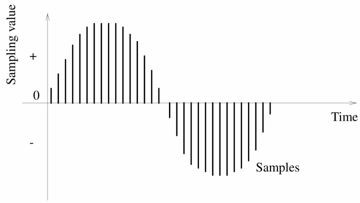

Figure 2: Sampling a wave

Representation (2)

Sampling Rate: The sampling rate is the frequency at which a continuous waveform is sampled, measured in Hertz (Hz). For instance, CDs have a sampling rate of 44,100 Hz, which slightly exceeds the human hearing range. Due to the *Nyquist theorem*, the maximum usable frequency is half the sampling rate. Therefore, a CD's sampling rate of 44,100 Hz covers frequencies up to 22,050 Hz—just within the human hearing limit.

Quantization: Digitizing an analog signal involves two steps: sampling and quantization. Sampling captures discrete values at regular intervals, while quantization converts these samples into a limited number of levels. For instance, 8-bit quantization provides 256 levels, whereas 16-bit (CD-quality) provides over 65,536 levels. Figure 3 illustrates a 3-bit quantization example.

Representation (3)

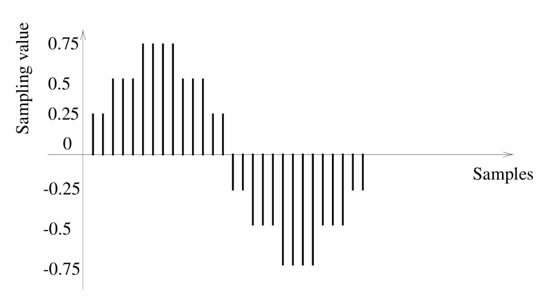

Figure 3: 3-bit quantization

With 3-bit quantization, values are limited to eight distinct levels: 0.75, 0.5, 0.25, 0, -0.25, -0.5, -0.75, and -1. Lower bit depth results in a more angular waveform and generally reduced sound quality, as fewer bits represent the original signal less precisely.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Formats (1)

Sound formats define how digital audio data is encoded, stored, and compressed. Each format has unique characteristics that make it suitable for different applications, such as streaming, archiving, or high-quality playback.

1. **Uncompressed Formats:** Uncompressed audio formats preserve the original audio quality without reducing file size, making them large but highly accurate. These formats are ideal for professional audio production or high-quality audio archiving.

   WAV (Waveform Audio File Format): A widely-used uncompressed format developed by Microsoft and IBM, WAV files store audio in raw, uncompressed PCM (Pulse Code Modulation) form. It is a standard format for CDs and is often used in professional audio editing due to its high-quality and broad compatibility.

   AIFF (Audio Interchange File Format): Created by Apple, AIFF is similar to WAV in terms of quality and file size. It is often used on Apple devices and by audio professionals requiring lossless quality.

Formats (2)

2. **Lossless Compressed Formats:** Lossless compression reduces the file size while preserving the original audio quality. These formats are beneficial for listeners and professionals who want high-quality audio without the large file size of uncompressed formats.
   FLAC (Free Lossless Audio Codec): A popular open-source format that compresses audio without any loss in quality. FLAC files are widely supported, and the format is preferred for high-resolution audio storage and streaming.
   ALAC (Apple Lossless Audio Codec): Apple's version of a lossless codec, ALAC compresses audio while maintaining quality. It is compatible with iTunes and other Apple devices, making it a common choice for Apple users.

Formats (3)

3. **Lossy Compressed Formats:** Lossy compression significantly reduces file size by discarding some audio data, typically focusing on parts less perceivable to human hearing. While this reduces quality slightly, it makes these formats ideal for online streaming and limited storage.
   MP3 (MPEG Layer-3 Audio): Perhaps the most well-known audio format, MP3 offers high compression rates, which makes it ideal for streaming and storage on devices with limited space. While MP3 does sacrifice some quality, it remains popular for its efficiency and compatibility.
   AAC (Advanced Audio Codec): A successor to MP3, AAC generally offers better sound quality at similar bit rates. It is widely used in online streaming services like YouTube and Apple Music, as well as on most smartphones and computers.

Formats (4)

4. **Specialized Formats:** These formats serve specific needs and applications, such as surround sound systems, portable devices, or interactive media.

   OGG Vorbis: An open-source lossy format, OGG Vorbis is often used in gaming and streaming. It provides good sound quality at lower bit rates and is patent-free, making it a popular choice for open-source projects.

   DSD (Direct Stream Digital): A high-resolution audio format developed by Sony and Philips, DSD is known for its exceptional sound quality and is commonly used in SACDs (Super Audio CDs). However, DSD files are large and often require specialized equipment for playback.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Basic Music Concepts

Any sound, including music, can be represented digitally as a sequence of binary-encoded samples, either uncompressed or compressed. However, this format lacks semantic detail (meaning in language or logic), meaning a computer cannot identify if the sequence represents speech or music, or distinguish musical notes and instruments without advanced processing.

For music, a symbolic representation is possible through scores, and computers achieve this using the MIDI (Musical Instrument Digital Interface) standard, developed in the 1980s. MIDI encodes musical elements such as notes, timing, and instrument assignments, enabling more meaningful musical representation.

Music is a form of sound that is artistically arranged using elements like rhythm, melody, and harmony to create a pleasing or meaningful auditory experience.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Introduction to MIDI (1)

MIDI represents a set of specifications used in instrument development to enable easy musical information exchange between instruments from different manufacturers. The MIDI protocol works as a complete music description language in binary form, with each word describing a musical performance action assigned a specific binary code.

A MIDI interface consists of two components:

1. **Hardware to connect equipment:** MIDI hardware specifies the physical connection of musical instruments, adding a MIDI port to an instrument, defining a MIDI cable for connecting two instruments, and processing the electrical signals received over the cable.

Introduction to MIDI (2)

2. **Data format for processing information:** The MIDI data format encodes information processed by the hardware, but it does not include individual sampling values as seen in audio data formats. Instead, it encodes specific details for each instrument, such as the start and end of scores, base frequency, loudness, and the instrument used.

The MIDI data format is digital, with data grouped into MIDI messages. When a musician presses a key, the MIDI interface generates a MIDI message that indicates the start of a score and its intensity, transmitting this message to connected devices. When the key is released, another MIDI message is created and transmitted.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Devices (1)

An instrument that follows both components of the MIDI standard is called a MIDI device (e.g., a synthesizer) and can communicate with other MIDI devices over designated channels. The MIDI standard specifies 16 channels, with each MIDI device (musical instrument) mapped to a specific channel.

Musical data sent over a channel is reproduced by the synthesizer on the receiver's end. The MIDI standard also defines 128 instruments using numbers, including various noise effects (e.g., a phone ringing or an airplane taking off). For example, 0 represents a piano, 12 a marimba, 40 a violin, and 73 a flute.

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

Devices (2)

1. **Sound Generators:** Sound generators perform the actual task of synthesizing sound. The rest of the synthesizer's components are primarily used to control the sound generators.
2. **Microprocessors:** The microprocessor communicates with the keyboard to detect which notes the musician is playing and with the control panel to interpret the musician's commands. It also sends and receives MIDI messages.
3. **Keyboard:** The keyboard provides the musician with direct control over the synthesizer.
4. **Control Panel:** The control panel manages functions that are not directly related to notes and durations (which are controlled by the keyboard).

Devices (3)

5. **Auxiliary Controllers:** Auxiliary controllers offer additional control over the notes played on the keyboard. Two common controls on a synthesizer are pitch bend and modulation.
6. **Memory:** The synthesizer's memory stores patches for the sound generators and settings for the control panel.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Messages (1)

MIDI messages transmit information between MIDI devices and determine what kinds of musical events can be passed from device to device. The format of MIDI messages consists of the status byte (the first byte of any MIDI message), which describes the kind of message, and data bytes (the following bytes).

MIDI messages are divided into two different types– Channel Messages and System Messages.

1. **Channel Messages:** Channel messages go only to specified devices. There are two types of channel messages– Channel Voice Messages and Channel Mode Messages

Messages (2)

- *Channel Voice Messages* send 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. Each message has at least one and usually two data bytes that accompany the status byte to describe these sound qualities. Examples of channel voice messages are 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. They set the MIDI channel receiving modes for different MIDI devices, stop spurious notes from playing, and affect other aspects of MIDI control. Examples are 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– System Real-Time Messages, System Common Messages, and System Exclusive Messages.

Messages (3)

- *System Real-Time Messages* are vital, short, and one byte. They carry extra data with them. These messages synchronize the timing of MIDI devices in performance, therefore, it is important that they be sent at precisely the time they are required. To avoid delays, these messages are sent in the middle of other messages if necessary. Examples of such messages are System Reset, Timing Clock (MIDI clock), etc.
- *System Common Messages* prepare devices to start or synchronize a song, such as selecting a song or tuning synthesizers. Examples include Song Select and Tune Request.
- *System Exclusive Messages* allow MIDI manufacturers to create customized MIDI messages to send between their MIDI devices. This coding starts with a system-exclusive-message, where the manufacturer is specified, and ends with end-of-exclusive messages.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Standards (1)

For more information visit this link

The MIDI standard is a technical protocol that allows electronic musical instruments, computers, and other equipment to communicate, control, and synchronize with each other. Developed in the early 1980s by a coalition of major instrument manufacturers, MIDI transformed music production by enabling the integration and control of diverse musical instruments and devices from different manufacturers.

1. **MIDI Message Structure:** MIDI messages are small packets of data that are transmitted in real-time. They contain instructions about which notes to play, how long to play them, and various other parameters. These messages can be categorized into several types:
    - Note On/Off: Indicates when a note is pressed or released.
    - Control Change (CC): Used to control various instrument parameters (e.g., modulation, sustain pedal).

Standards (2)

- Program Change: Used to change the instrument sound or preset (e.g., switching from piano to strings).
- Pitch Bend: Used to bend the pitch up or down smoothly.
- Channel Pressure and Polyphonic Key Pressure: Used to apply after-touch effects on a per-note or per-channel basis.
- MIDI messages are compact, typically only 1 to 3 bytes long, which allows them to be transmitted quickly. Messages are split into status bytes (indicating message type and channel) and data bytes (indicating note values, velocities, or other parameters).

2. **MIDI Channels:** The MIDI standard provides 16 channels per MIDI connection. This allows multiple instruments or sounds to be controlled independently over a single MIDI cable. Each instrument or sound is assigned to a specific channel (e.g., drums on channel 10, piano on channel 1). This setup allows a single MIDI stream to control multiple instruments simultaneously.

Standards (3)

3. **MIDI Connectors and Hardware:** The original MIDI standard specifies the use of 5-pin DIN connectors and MIDI cables for connecting devices. There are three types of MIDI ports:
   MIDI In: Receives MIDI messages.
   MIDI Out: Sends MIDI messages generated by the instrument.
   MIDI Thru: Passes incoming MIDI messages to another device.
4. **General MIDI (GM) Standard:** This standard was introduced to ensure compatibility between MIDI files on different devices. It specifies a standardized set of 128 instruments and sounds (e.g., pianos, strings, drums) and assigns them to specific program numbers. This way, a MIDI file will produce similar sounds regardless of the device or software playing it.

Standards (4)

5. **MIDI Clock and Sync:** This includes a system for synchronizing devices, known as MIDI Clock. This ensures that devices can stay in sync when playing back or recording. MIDI Clock provides a timing pulse that allows devices like drum machines, sequencers, and synthesizers to operate at the same tempo.

   There is also MIDI Time Code (MTC), which is used for synchronization with video and other time-based media, allowing MIDI devices to sync with external timecode formats for film and broadcast.

6. **System Exclusive (SysEx) Messages:** These messages are unique to each manufacturer. They allow specific instruments to send and receive proprietary data, such as patch settings or updates, without affecting other MIDI devices. SysEx messages are often used for device-specific data exchange, like bulk parameter dumps or firmware updates.

Standards (5)

7. **MIDI File Format:** The MIDI standard file format is Standard MIDI File (SMF), which allows MIDI data to be saved, edited, and replayed. SMF files are platform-independent and often use the .mid file extension. They are commonly used for storing compositions, arrangements, and musical ideas in a way that can be shared across software and hardware.
8. **MIDI 2.0:** In 2020, MIDI 2.0 was introduced as an update to the original standard. MIDI 2.0 is backward-compatible with MIDI 1.0 but offers expanded functionality, including:

   Higher Resolution: MIDI 2.0 offers 32-bit resolution for parameters, providing much finer control.

Standards (6)

Bidirectional Communication: MIDI 2.0 allows devices to communicate in both directions, so devices can exchange information about their capabilities and settings.

Profiles and Property Exchange: Profiles allow instruments to conform to specific sets of functionality, like “Piano” or “MIDI Polyphonic Expression (MPE)”. Property exchange allows instruments to exchange information, like patch names, in real time.

Improved Timing and Velocity Precision: Timing and expression data are far more precise, benefiting dynamics and articulations.

Standards (7)

9. **MIDI Polyphonic Expression (MPE):** MPE is a standard that extends MIDI's expressiveness by allowing each note played on a device to have its own channel of control data. This is especially useful for expressive instruments like the ROLI Seaboard, where each finger's pressure, slide, and position can be detected independently.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Software (1)

Once a computer is connected to a MIDI system, a wide variety of MIDI applications can be run on it. Digital computers offer composers and sound designers unparalleled control over the evolution and combination of sonic events.

The software applications generally fall into four major categories:

1. **Music Recording and Performance Applications:** These applications allow the recording of MIDI messages as they are received from other MIDI devices, and also provide options for editing and playback of messages during performances.
2. **Musical Notation and Printing Applications:** These applications enable users to write music using traditional musical notation. The music can then be played back through a performance program or printed on paper for live performance or publication.

Software (2)

3. **Synthesizer Patch Editors and Librarians:** These programs allow storage of various synthesizer patches in the computer's memory and disk drives, and also facilitate patch editing within the computer.

   Music Education Applications: These applications use the computer monitor, keyboard, and other controllers of attached MIDI instruments to teach different aspects of music.

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:

1. **The Sensing Stage:** In this stage, data is collected from controllers that read gesture information from human performers on stage.
2. **The Processing Stage:** Here, the computer reads and interprets the information from the sensors and prepares the data for the response stage.

Software (3)

3. **The Response Stage:** In this final stage, the computer and various sound-producing devices work together to create a musical output.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Speech (1)

Speech can be perceived, understood, and generated by both humans and machines. Humans adapt efficiently to different speakers and their unique speech habits, allowing them to understand speech despite variations in dialects and pronunciation. The human brain can differentiate between speech and noise, a task facilitated by using both ears, as filtering with only one ear makes this process significantly harder. The human speech signal includes a subjective lowest spectral component known as pitch, which is not directly proportional to frequency.

The human ear is most sensitive within the range of 600 Hz to 6000 Hz. Research by Fletcher and Munson has demonstrated that the ear is substantially less sensitive to low and very high frequencies than to frequencies around 1 kHz.

Speech (2)

Speech signals have two properties that can be utilized in speech processing:

1. *Voiced Speech Signals:* These signals display an almost periodic behavior during certain time intervals, making them quasi-stationary for about 30 milliseconds.
2. *Spectral Characteristics:* The spectrum of audio signals shows characteristic peaks in 3-5 frequency bands, called formants, which occur due to resonances in the vocal tract.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Speech Generation/Synthesis (1)

Speech generation, also known as speech synthesis, is the process of creating artificial human speech that can be heard through a device or computer. In other words, speech generation is a process that converts text into an acoustic signal, making written language audible.

With computers, one can synthetically generate speech, where the generated signals do not sound quite natural but can be easily understood. An example of such an artificial sounding voice can be heard at the airport. On the other hand, a voice can sound natural but may be very difficult to understand. Speech recognition often uses matching rules or statistically based methods. There are, and will continue to be in the near future, considerable differences between the speech generation and recognition efficiencies/capabilities of the human brain and a high-performance computer.

Speech Generation/Synthesis (2)

Speech generation research has a long history. By the middle of the 19th century, Helmholtz had already built a mechanical vocal tract coupling several mechanical resonators, with which sound could be generated. In 1940, Dudley produced the first speech synthesizer through imitation of mechanical vibration using electrical oscillation.

An important requirement for speech generation are:

- *Real-time signal generation:* Speech generation should be real-time signal generation
- *No lengthy pre-processing:* When above requirement is met, a speech output system can automatically transform text into speech without lengthy pre-processing. Some applications only require a limited vocabulary, such as the spoken time announcement of a telephone answering service. However, most applications need a large vocabulary, if not an unlimited one.

Speech Generation/Synthesis (3)

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

A computer system used for speech generation is called a speech synthesizer. Figure 4 shows the components of such a system.

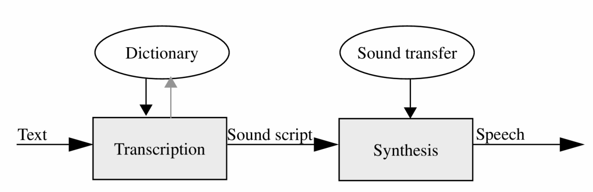

Figure 4: Components of a speech synthesis system, using sound concatenation in the time range

Speech Generation/Synthesis (4)

1. **Phonetic Transcription:** The first step involves a transcription, or translation, of the text into the corresponding phonetic transcription. Most methods use a lexicon containing a large quantity of words, syllables, or tone groups. The creation of such a library is extremely complex and can be an individual implementation or a common lexicon used by several people. The quality can be continually improved by interactive user intervention. This means that users recognize defects in the transcription formula, improve their pronunciation manually, and gradually integrate their findings into the lexicon.
2. **Acoustic Signal Conversion:** The second step converts the phonetic transcription into an acoustic speech signal, where concatenation can occur in the time or frequency range. While the first step normally has a software solution, the second step often requires signal processors or dedicated hardware processors.

Speech Generation/Synthesis (5)

Addressing Pronunciation Ambiguities: In addition to the challenges posed by co-articulation and prosody, speech synthesis systems must address pronunciation ambiguities. These ambiguities can lead to misinterpretations, such as “the grass is full” or “the glass is fool,” instead of the intended phrase “the glass is full.” The only way to solve this problem effectively is to provide additional contextual information.

Basic Notions* (1)

- 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.

Basic Notions* (2)

- 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.

1. **Time-dependent Sound Concatenation:** see page 44 and 45 of book [1]
2. **Frequency-dependent Sound Concatenation:**see page 45 and 46 of book [1]

Table of Contents

Speech Analysis (1)

Speech analysis/input deals with the research areas shown in Figure 5.

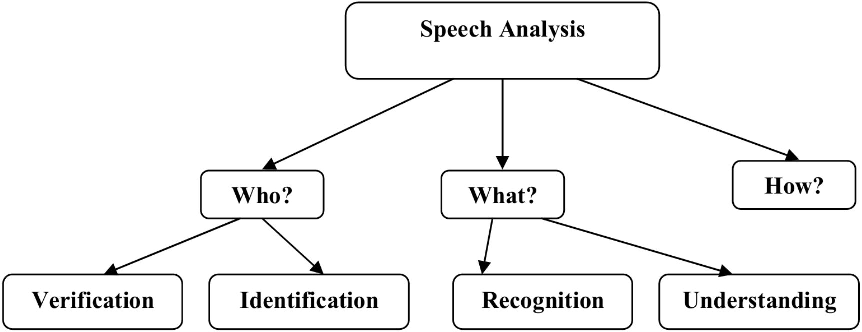

Figure 5: Research areas of speech analysis

Speech Analysis (2)

Human speech exhibits distinct characteristics that are unique to each speaker. As a result, speech analysis can be used to identify and verify who is speaking. This process allows for the recognition of a speaker based on their individual traits. The computer identifies and authenticates the speaker by analyzing their acoustic fingerprint. An acoustic fingerprint refers to a digitally stored speech sample (e.g., a specific phrase) from an individual.

Another application of speech analysis is to understand the content of what has been said, i.e., to interpret and transcribe the speech signal. Based on the sequence of speech, the corresponding text can be generated. This technology can lead to systems like speech-controlled typewriters, translation tools, or assistive devices for individuals with disabilities.

Furthermore, speech analysis is also used to study speech patterns related to how something is said. For instance, the tone of a spoken sentence can

Speech Analysis (3)

vary depending on whether the speaker is calm or angry.

The primary goal of speech analysis is correctly determine individual words with probability ≤ 1. A word is only recognized only with a certain probability. Therefore, environmental noise, room acoustics and speaker's physical and psychological conditions play important role.

Second, the speech elements are compared with existent references to determine the mapping to one of the existent speech elements. The identified speech can be stored, transmitted or processed as a parameterized sequence of speech elements.

Table of Contents

Speech Recognition* (1)

Speech recognition is the process of transforming an acoustic signal, picked up by a microphone or telephone, into a sequence of words or a similar format. Speech recognition systems are classified into speaker-independent and speaker-dependent systems. Speaker-dependent systems, which are pre-trained, can recognize with higher accuracy, while speaker-independent systems can recognize with lower accuracy.

The principle of speech recognition (Figure 6) involves comparing distinctive features of individual utterances (any stream of speech between two periods of silence) with a set of previously extracted speech elements. Typically, these features are quantized, allowing for a structured analysis of the speech sequence. The result is then matched against known references to classify it within one of the existing speech elements. Once identified, utterances can be stored, transmitted, or processed as a parameterized sequence of speech components.

Speech Recognition* (2)

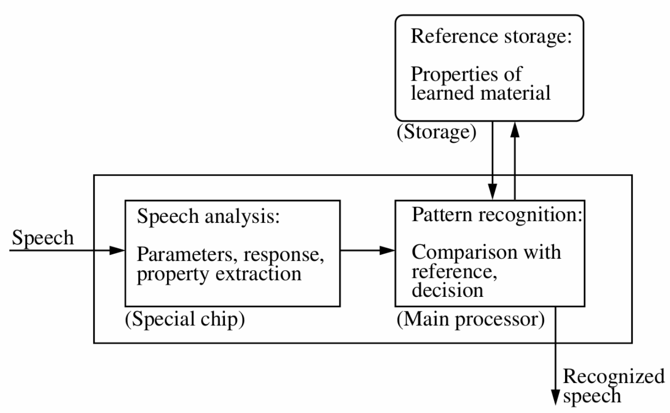

Figure 6: The speech recognition principle: the tasks are distributed to system components by the basic principle "extract characteristics to reduce data."

Speech Recognition* (3)

Most practical methods vary in how they define characteristic properties. The principle illustrated in Figure 6 can be applied multiple times, each iteration focusing on different characteristics. The application of this principle, as depicted in Figure 6, can be broken down into the steps shown in Figure 7.

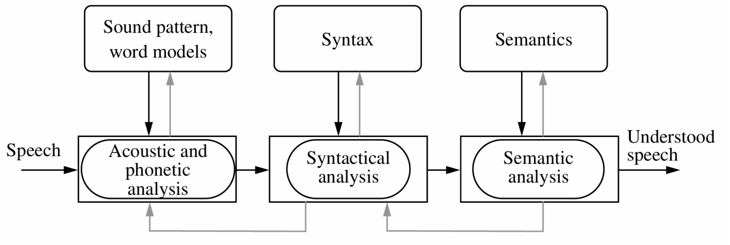

Figure 7: Speech recognition components

Speech Recognition* (4)

1. **Acoustic and Phonetic Analysis:** Referring to the characteristic properties of the chosen method, the first step applies the principle shown in Figure 6 to sound patterns and/or word models.
2. **Syntactic Analysis:** The second step uses the speech units determined in the first step to run a syntactic analysis on them. This process can detect errors in the first run and serves as an additional decision tool because the first step does not normally provide a final decision.
3. **Semantic Analysis:** The third step analyzes the semantics of the speech sequence recognized to this point. This step can detect errors from the previous decision process and remove them by using another interplay with other analytical methods. Note that even with current artificial intelligence and neural network technologies, the implementation of this step is extremely difficult.

Table of Contents

1. Introduction
2. Basic Sound Concepts
   - Representation
   - Formats
3. Basic Music Concepts
   - Introduction to MIDI
   - Devices
   - Messages
   - Standards
   - Software
4. Speech
   - Speech Generation/Synthesis
   - Speech Analysis
   - Speech Recognition*
   - Transmission

Transmission (1)

The area of speech transmission deals with efficient coding of the speech signal to allow speech/sound transmission at low transmission rates over networks, aiming to reduce perceptible quality degradation. 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.

1. **Pulse Code Modulation:** Signal form encoding does not consider speech-specific properties and/or parameters. Here, the goal is to achieve the most efficient coding of the audio signal. A straightforward technique for digitizing an analog signal (waveform) is Pulse Code Modulation (PCM). This method is simple, but it still meets the high quality demands stereo-audio signals in the data rate used for CDs:

$$\text{rate} = 2 \times \frac{44,100}{s} \times \frac{16\text{bits}}{8\text{bits/byte}} = 176,400 \text{ bytes/s} = 1,411,200 \text{ bytes/s}$$

Transmission (2)

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 Kbit/s without loss of quality. Adaptive Pulse Code Modulation (ADPCM) allows a further rate reduction to 32 Kbit/s.

2. **Source Encoding:** Source encoding is a technique used in speech transmission to reduce the amount of data required for encoding signals by exploiting the inherent characteristics of the signal. In the case of speech, this involves transformations that depend on its specific features, such as suppressing silence during pauses between words. This reduction is achieved by recognizing and eliminating parts of the signal that are not essential for conveying the message, thereby improving efficiency.

   Parametric systems, like the channel vocoder, use such speech-specific characteristics to encode only the most relevant features of the signal,

Transmission (3)

such as pitch, tone, and frequency, rather than transmitting the entire signal. This results in a more compressed and efficient representation of speech, which is particularly useful in applications that require low-bandwidth transmission.

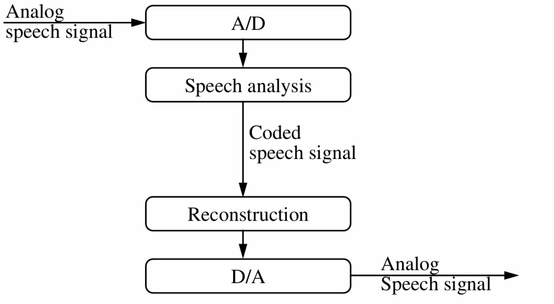

Figure 8: Components of a speech transmission system using source encoding

Transmission (4)

3. **Recognition-Synthesis methods:** Recognition-synthesis methods are a cutting-edge approach in speech transmission aimed at significantly reducing data volumes while maintaining high quality. Current research seeks to cut the data rate to around 6 Kbit/s, with the goal of maintaining the quality equivalent to an uncompressed 64 Kbit/s signal.

   This method involves analyzing the speech and then synthesizing it during reconstruction, allowing the transmission of only essential speech characteristics, such as formants, which represent the center frequencies and bandwidths for use by digital filters. The result is a substantial reduction in data usage, potentially lowering the rate to approximately 50 bit/s. However, further work is needed to enhance the quality of the synthesized speech and improve its recognition accuracy to ensure the method becomes viable for widespread use.

Transmission (5)

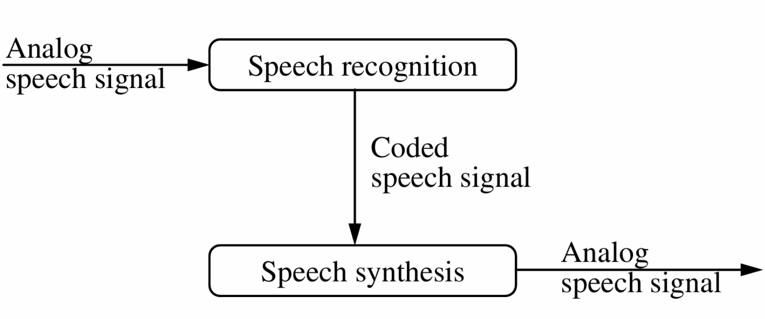

Figure 9: Components of a recognition-synthesis system for speech transmission

Transmission (6)

4. **Achievable Quality:** Achievable quality in speech and audio transmission is a critical factor in multimedia systems, as it determines the minimum data rate required to maintain a defined level of audio fidelity. A key point of discussion is that telephone-quality audio can be achieved at a data rate of less than 8 Kbit/s. Figure 10 demonstrates the relationship between audio quality and the number of bits per sampling value, showing how reducing the number of bits can still maintain good quality.

   For example, by reducing the bits per sample from 16 to 2, the data rate can be reduced to just one-eighth of its original value, yet still deliver audio quality comparable to that of a standard CD. This illustrates how significant data compression can be achieved while preserving adequate audio quality.

Transmission (7)

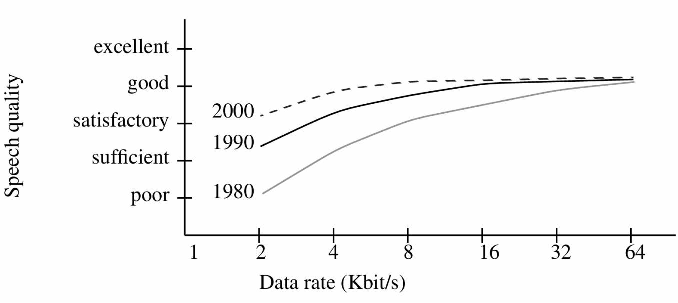

Figure 10: Quality of compressed speech in relation to the compressed signal's data rate

R. Steinmetz and K. Nahrstedt.
Multimedia: Computing, Communications and Applications.
Pearson, 2012.