Digital Video and Audio

Agenda

- A video is a sequence of images
- A sound is characterised by its frequency (pitch) and amplitude (loudness)
- CD standard quality is 44,100Hz (sampling) and 16 bits (quantisation)
- Speech signals contain 3 types of sound, some of them are used for speech recognition
- MIDI format for music stores information such as instrument specification, beginning and end of a note, basic frequency, etc

Video

Video is the technology of electronically capturing, recording, processing, storing, transmitting, and reconstructing a sequence of still images representing scenes in motion.

Frame rate: the number of still pictures per unit of time of video.
(e.g. 30 frame/second)

Analog video: video recording method that stores continuous waves of red, green and blue intensities.

Digital video: video recording system that works by using a digital rather than an analog video signal

Refresh rate and frame rate (*)

Refresh rate: the number of times in a second that the display hardware draws the data (i.e. repeated drawing of identical frames)

Frame rate: measures how often a video source can feed an entire frame of new data to display

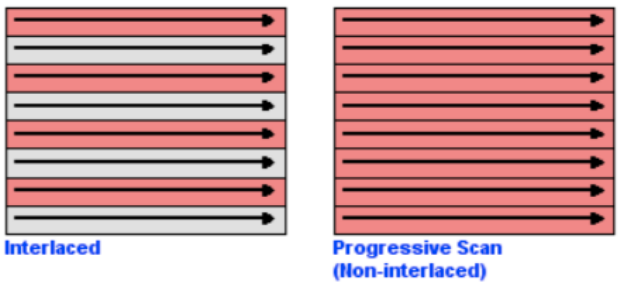
- Typical rates: 24 frames per second (framerate) and 48 or 72 Hz (refresh rate)
- If the frame rate is higher than the refresh rate, the display will not be able to display all of the frames the computer is producing

(你在游戏中能够获得超过200帧的画面，但是由于显示器刷新率只有30Hz，只能“抓取”其中的30帧进行显示，最终你所看到的画面也是30帧。)

Interlaced vs Progressive

Interlaced scanning displays alternating sets of lines. Because each field happens so quickly we are given the illusion of a whole image.

Progressive video displays the entire image.



For the interlaced scanning, it reles on the fact that your eyes can’t detect what’s happening at least you don’t perceive the partially drawn images
Half appear on the screen the other half follows an instant later of 1/60 of seconds to be processed.

Exercise
A 30fps digital video uses 352 by 255 pixels video frames with a pixel depth of 8.
i) Calculate the size of 1 second of data.
ii) What compression ratio would be needed to transmit 1 second of data in real-time over a 64 Kbps communication channel?
Solution:

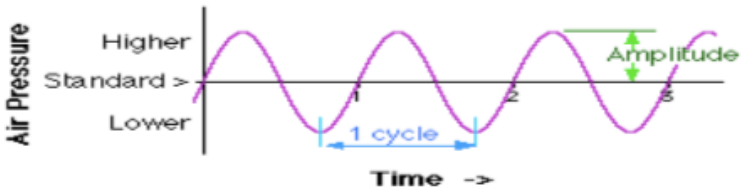
1) size of 1 second of data $= \frac{352 \cdot 255 \cdot 8}{8} \cdot 30 = 2692.8$ KB

2) $\frac{21542400}{64000} = 336.6$

Sound

Sound is a physical phenomenon produced by the vibration of matter, such as a violin string, or a block of wood.
As the matter vibrates, pressure variations are created in the air surrounding it.
This alteration of high and low pressure is propagated through the air in a wave-like motion

Characteristics of Sound Waveforms



Frequency determines the pitch (higher frequency = higher pitch)

- Infra-sound: from 0 to 20 Hz
- Human hearing frequency range: 20 Hz – 20 kHz
- Ultrasound: from 20 kHz to 1 GHz

Amplitude of the wave determines the volume or intensity (a property subjectively heard as loudness)

Computer Representation of Sound

Sampling

Sampling rate: the rate at which a waveform is sampled.
e.g. the CD standard sampling rate of 44100 Hz means that the waveform is sampled 44100 times / second

Quantization

the resolution or quantization of a sample value depends on the number of bits used in measuring the height of the waveform (usually 8-bit or 16-bit)
8-bit — speech | 16-bit — music
Exercise 1
A high-quality (CD standard at 44.1KHz) audio signal with 2 channels of 16-bit samples is transmitted uncompressed over an ISDN 64Kbits/s communications channel.

i) Calculate the number of seconds taken to transmit a one-second burst of audio
ii) Estimate what compression ratio would be needed to transmit the audio in real-time

Solution:

1) $44100 \cdot 2 \cdot 16 = 1411200$ Kb, $\frac{1411200}{64000} = 22.05$ (s)

2) Compression ratio $= \frac{1411200}{64000} = 22.05$

Exercise 2

The bandwidth of a music signal is between 15 Hz and 20 KHz, assuming the Nyquist sampling rate is used, with 16 bits per sample:

- Derive the bit rate that is generated by the digitisation procedure
- What is the memory in Mbytes required to store a 10 minute passage of stereophonic music?

Solution:

1) $fs \geq 2 \cdot fm, fs \geq 2 \cdot 20K = 40\text{K}, \; RB = Fs \cdot Nb = 40000 \cdot 16 = 640$ Kbps

2) Stereophonic — 32 bits, Memory $= \frac{40000 \cdot 32}{8} \cdot 10 \cdot 60 = 96$ Mbytes

Aliasing (sampling error)

The reason a too-low sampling rate results in aliasing is that there aren't enough sample points from which to accurately interpolate the sinusoidal form of the original wave.

If we take more than two samples per cycle on an analog wave, the wave can be precisely reconstructed from the samples.

Measuring Sound Amplitude in Decibels

For sound, the reference point is the air pressure amplitude for the threshold of hearing

A decibel in the context of sound pressure level is called decibels-sound-pressure-level (dB_SPL)

Let E be the pressure amplitude of the sound being measured and E0 be the sound pressure level of the threshold of hearing. Then decibels-sound-pressure-level, (dB_SPL) is defined as

$$dB\_SPL = 20 \log_{10}\left(\frac{E}{E_0}\right)$$

E0 = 0.00002 Pa

Exercise

- What would be the amplitude (in decibels) of the audio threshold of pain, given as 30 Pa?
- what would be the pressure amplitude of normal conversation, given as 60 dB?

Solution:

1) $20 \log\left(\frac{30}{0.00002}\right) = 123.5$ dB

2) $60 = 20\log\left(\frac{E}{0.00002}\right)$, E = 0.02 Pa

dB_SPL is an appropriate unit for measuring sound because the values increase logarithmically rather than linearly, which is a better match for the way humans perceive sound

- if you increase the amplitude of an audio recording by 10 dB, it will sound about twice as loud
- 3 dB change in amplitude is the smallest perceptible change

Signal to Quantisation Noise Ratio (SQNR)

$$SQNR = 20\log_{10}\left(\frac{\max(quantization\ value)}{\max(quantization\ error)}\right)$$

Let n be the bit depth of a digitised media file (e.g. digital audio). Then the signal-to-quantisation noise ratio **SQNR (or dynamic range)** is:

$$SQNR = 20\log10(2n) = 20n\log10(2) \sim \mathbf{6n}$$

Dynamic Range

an n-bit digital audio file has a dynamic range (or, equivalently, a signal-to-noise-ratio) of 6n dB.

Dynamic range is a relative measurement — the relative difference between the loudest and softest parts representable in a digital audio file, as a function of the bit depth.

What is the dynamic range (SQNR) of a 16 bit digital audio file?

- How about a 8 bit digital audio file?

Solution:

1) SQNR = 6n = 64 dB

2) SQNR = 6n = 48 dB

Quantisation Error

Two ways to deal with quantisation error: **Audio dithering + Noise shaping**

Audio dithering: add small random values to samples in order to mask quantisation error

Noise shaping: it redistributes the quantisation error so that the noise is concentrated in the higher frequencies, where human hearing is less sensitive

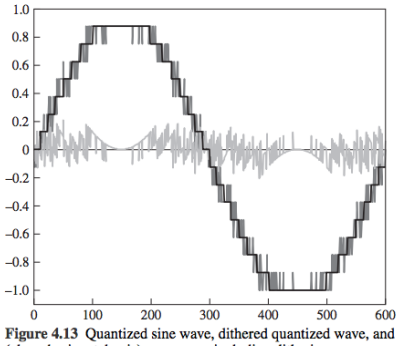
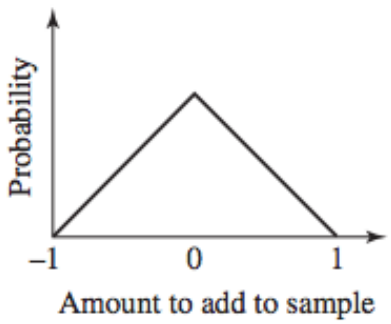


Figure 4.13 Quantized sine wave, dithered quantized wave, and (along horizontal axis) error wave including dithering



Speech signals

Types of Speech Sounds

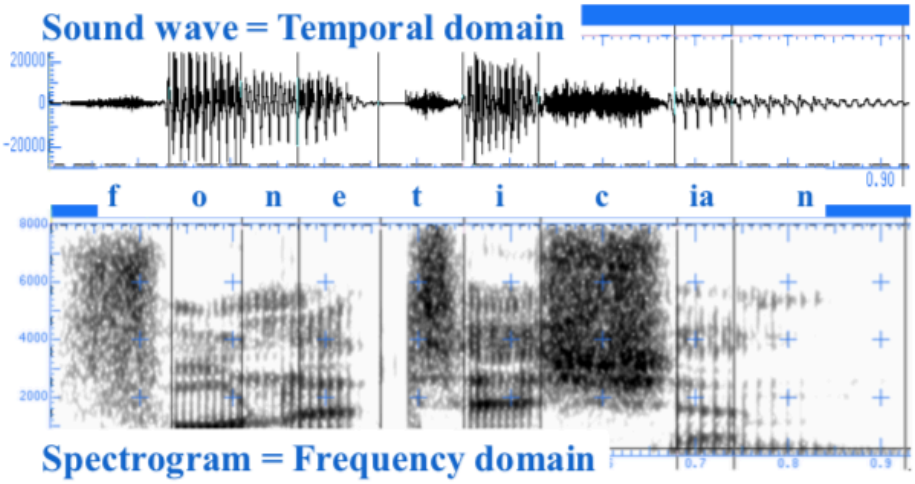
Voiced sounds : the vocal chords are vibrated, which can be felt in the throat. All vowels are voiced.
Fricatives (unvoiced sounds) : a consonant, such as f or s in English, produced by the forcing of air through a constricted passage.
Plosives (also unvoiced sounds) : a speech sound produced by complete closure of the oral passage and subsequent release accompanied by a burst of air, as in the sound (d) in dog.

Properties of three types of sounds

Voiced:

1. periodic behaviours (quasi-stationary signals for around 30 ms)
2. formants — The spectrum of speech signals (voiced sounds) show characteristic maxima (occur because of resonances of the vocal tract)

Fricatives: noise in the signal
Plosives: clearly starting (with some silence)



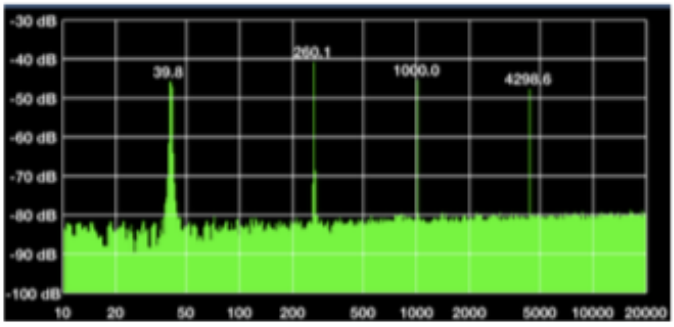
Temporal and Frequency Domains

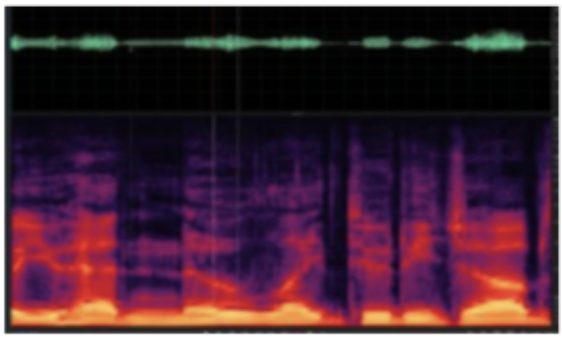
Sound can be represented either over the time domain or the frequency domain
In the frequency domain, data is stored as the amplitudes of frequency components
The time domain and frequency domain are equivalent. They both fully capture the waveform. They just store the information about the waveform differently

- A complex waveform is equal to an infinite sum of simple sinusoidal waves, beginning with a **fundamental frequency** and going through frequencies that are integer multiples of the fundamental frequency (**harmonic frequencies**)

Audio Histogram

Audio processing programs sometimes offer a statistical analysis of your audio files, which analyse sample values in the time domain
An audio histogram shows how many samples there are at each amplitude level in the audio selection

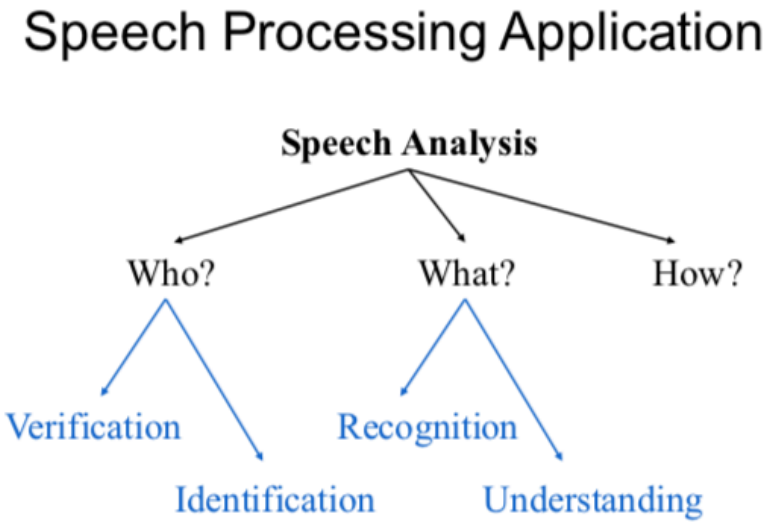




Power Spectrum **Spectrogram**

The power spectrum S(f) of time series x(t) describes the distribution of power into frequency components composing that signal
A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time.

Speech Processing Applications



Verification: Prerecorded signal (voice print), matching
Identification: Database of prerecorded signal
Recognition: Recognize the speeches (convert speeches to sentences)
Understanding: Semantic, progmatic knowledge

Music

Digital audio (Two ways to store): sampled & quantised digital audio and MIDI
MIDI (Musical Instrument Digital Interface)
MIDI stores "sounds events" or "human performances of sound"
The difference between sampled digital audio and MIDI is analogous to the difference between bitmapped graphics and vector graphics

MIDI Data Format

A MIDI file contains messages indicating when a note begins (Note On), when a note ends (Note Off), what the note is, how hard it is pressed (Velocity), how hard it is held down (Aftertouch), what instrument is played, and so forth

- Each MIDI message communicate one musical event
- The MIDI standard identifies 128 instruments (including noise effects) with unique numbers (e.g. 41 for the violin)

MIDI Hardware and Software

MIDI controllers: Hardware devices that generate MIDI messages
MIDI synthesizers: Devices that read MIDI messages and turn them into audio signals (frequency modulation synthesis and wavetable synthesis) — two methods
MIDI sequencer: hardware device or software application program that allows you to receive, store, and edit MIDI data

Questions

- Why are MIDI encoded music signals very small?
- What other advantage to MIDI audio is there compared to sampled digital audio?
- Is there any disadvantage to MIDI audio compared to sampled digital audio?

Solution:

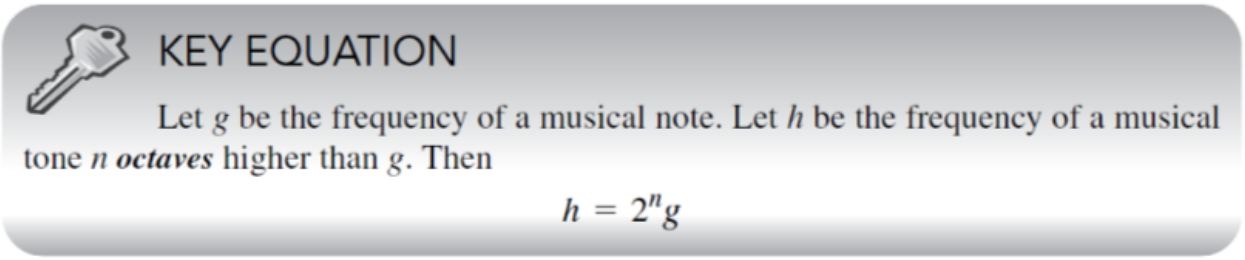
1. Digital audio files contain thousands of samples of sound, MIDI files only stores the "sounds" events, which are just strings of data. Hence, MIDI files are much smaller.
2. Advantages:
    1. Small size (much more compact)
    2. Completely editable (can cahnge the particular instrument)
    3. Sounds better
3. Disadvantages:
    1. MIDI playback will be accurate only if the MIDI playback device is identical to the device used for production
    2. MIDI cannot easily be used to play back spoken dialogue
    3. MIDI data is device dependent
    4. it can sound more artificial or mechanical than sampled digital audio

Musical Acoustics and Notation

Tones (music notation, musical sounds): characterized by **pitch, timbre, and loudness**

note: With the addition of onset and duration, a musical sound is called a note
pitch: how high or low it sounds to the human ear
timbre: tone color

The lowest frequency of a given sound produced by a particular instrument is its **fundamental frequency**. Then there are other frequencies combined in the sound, which are integer multiples of the fundamental frequency, referred to as **harmonics**.



h is the frequency of a musical tone n octaves higher than h

Exercise 1

- Given a note from a musical instrument, which contains only the following frequency components: 100Hz, 200Hz, 300Hz, and 400Hz, at what rate would you need to sample this sound to ensure that the sampled audio was of the same **fidelity** as the original note?
- Assuming that the amplitude of each harmonic is half the amplitude of the previous harmonic, sketch the signal in the frequency domain for the above note.

Solution:

1. 800 Hz (fidelity:保真度 — 再现输入信号的相似程度)

Exercise 2

If the frequency of a note A is about 440 Hz, what is the frequency of an A two octaves below the 440 Hz A?

Solution:

h = 440 Hz, g = 440 Hz / 4 = 110 Hz

Question in test

a) This question is about audio.

[5 marks]

i) Briefly explain what is shown in an audio histogram such as the one in Figure 1. In particular, comment the units used on the X and Y axes.

(3 marks)

ii) Do audio histograms and audio spectrograms represent signals in the same domain? Justify your answer.

(2 marks)

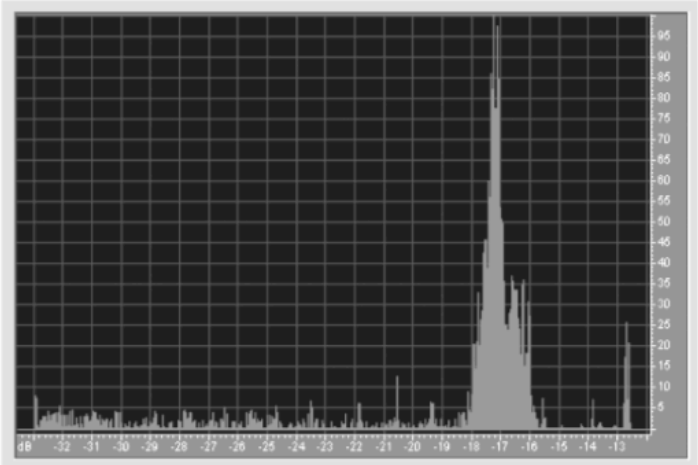


Figure 1: Audio histogram

1) This audio historgram shows how many samples are at a particular amplitude (in dB). Most of the audio samples have the amplitude between -16dB to -18dB, and -13dB. Besides, approximately with the amplitude at -17 dB, the numbers of samples are the largest, which is approximately 97~98
2) No. Audio histograms represent signals in time domain, it shows how many samples are at a particular amplitude. Spectrograms represent signals in frequency domain, it shows how the frequency spectrums change with the time varies.

a) This question is about sound encoding.

[9 marks]

i) Why is good quality sound usually encoded using a sampling rate of 44.1 kHz?

(4 marks)

ii) What is the typical sampling rate of a sound file encoded using MIDI?

(2 marks)

iii) What are speech formants? Refer to what you see in Figure 1 in your answer.

(3 marks)

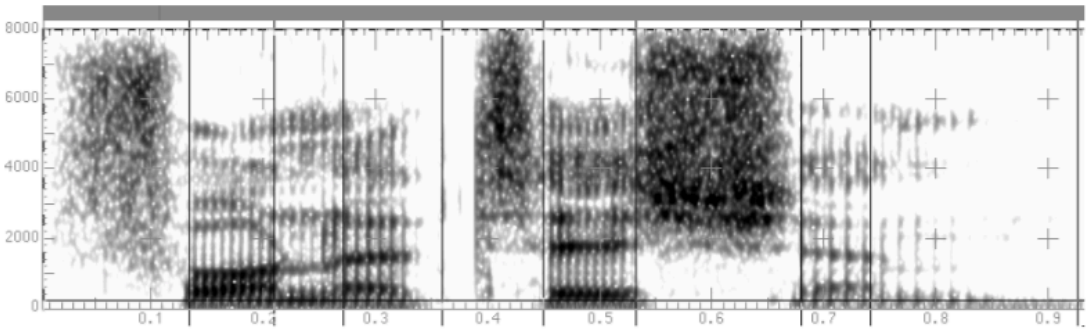


Figure 1: Speech spectrogram

1) Since the human ear can hear the sound with the frequency ranging from 20 Hz to 20 kHz, according to Nyquist theorem, the maximum freqeuency we need here is 20 kHz, to avoid aliasing, the sampling rate need to be greater than twice of the maximum frequency which is 40 kHz. Hence we choose 44.1 kHZ which is larger than the Nyquist sampling rate.
2) What is the typical sampling rate of a sound file encoded using MIDI? (44.1 kHZ)
3) Speech formants is a property of voiced sounds. It shows the characteristic maxima in the spectrogram due to the resonance. In figure 1, for example in the second block, we can see their are horizontal black lines that have greater value than others. These lines represent the speech formants of that voiced sound.