Multimedia Technology

Lecture 12: Sound and Music Search

Lecturer: *Dr*. Wan-Lei Zhao

Autumn Semester 2022

Email: wlzhao@xmu.edu.cn,

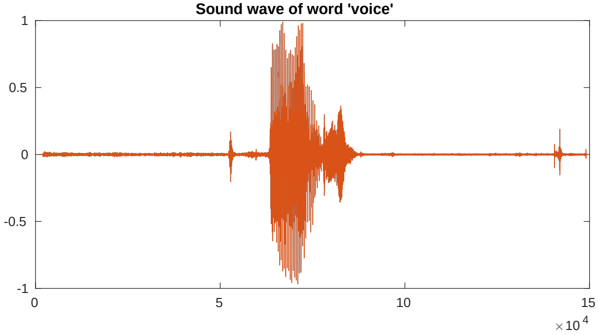
Outline

What is sound? (1)

- Sound is an accoustic wave that is caused by vibration
- It can be transmitted through gas, liquid and solid
- Properties of sound
  1. Speed: impacted by the density of the mass
  2. Frequency: of which different sounds are generated
  3. Pitch: the peaks of the sound wave
  4. Loudness: indicates the energy produced

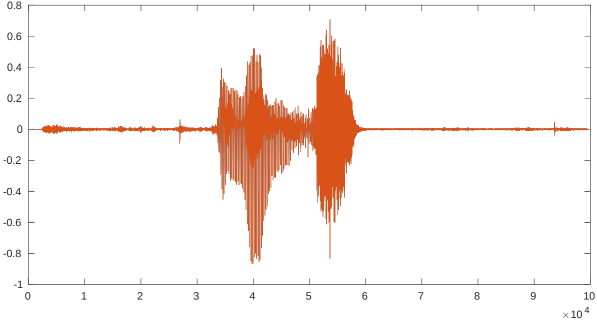
What is sound? (2)

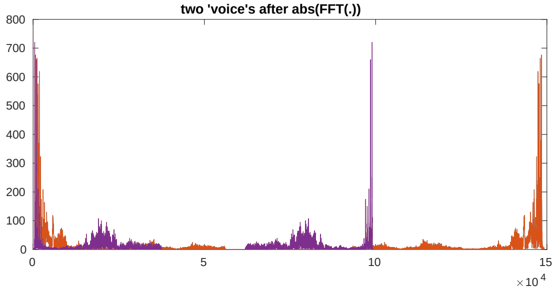
Sound of ‘voice’



What is sound? (3)

Sound of ‘voice’ from my student





My voice

My student's voice

- It is hard to analyze in time domain
- Sound in frequency domain

How Sound is Captured by Human being?

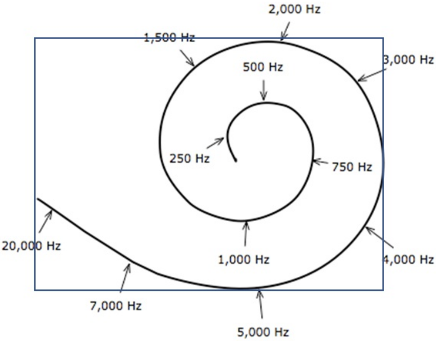
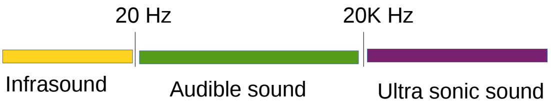


Figure: Cochlea of Human being.

- It is captured by Cochlea of our hears
- And decoded into signals of different frequencies

Categories of Sound

1. Speech
2. Music: singing and sound from instruments
3. Sounds from nature: from animals, birds, and insects, etc.
4. Noises
5. The mixture of above four categories



- We are only sensitive to sounds between 20Hz and 20K Hz

How Sound is Captured by Digital Devices

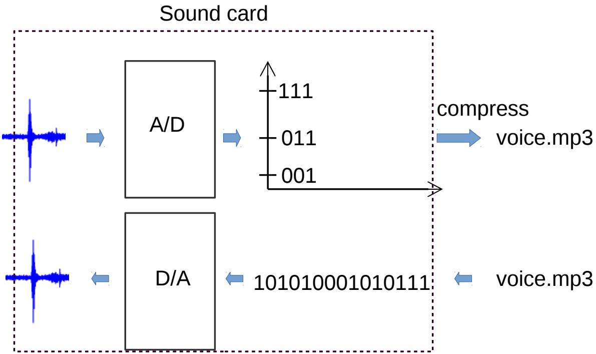


Figure: The pipelines of transforming voice into mp3 and its reverse procedure.

- Share my story ...

Period, Frequency and Sampling rate (1)

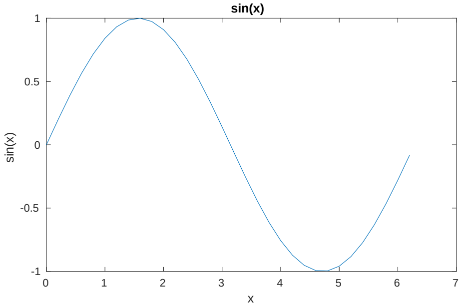


Figure: One period of sin(x), $T = 2\pi$, $F = \frac{1}{2\pi}$.

- One period of sin(x) curve/wave
- Its period is 2π and its frequency is $\frac{1}{2\pi}$
- Here, we assume this curve covers one second time

Period, Frequency and Sampling rate (2)

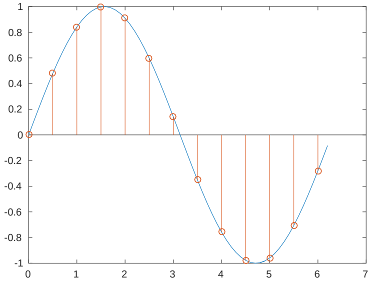


Figure: Sampling on one period of sin(x), $T = 2\pi$, $F = \frac{1}{2\pi}$.

- 13 positions are sampled. So the sampling rate is $1/13$
- Sampling frequency is $\frac{1}{13}$, $SF = 0.0769$

Period, Frequency and Sampling rate (3)

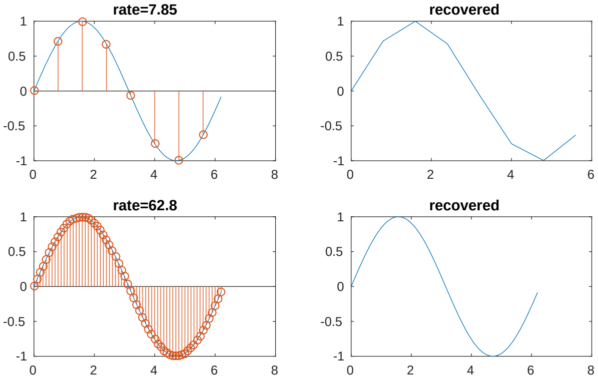


Figure: Sampling on one period of sin(x), $T = 2\pi$, $F = \frac{1}{2\pi}$.

- Sampling frequency is $\frac{1}{8}$, $SF = 0.125$
- According to *Shannon* theory, *FS* should be two times bigger than *F*

Period, Frequency and Sampling rate (4)

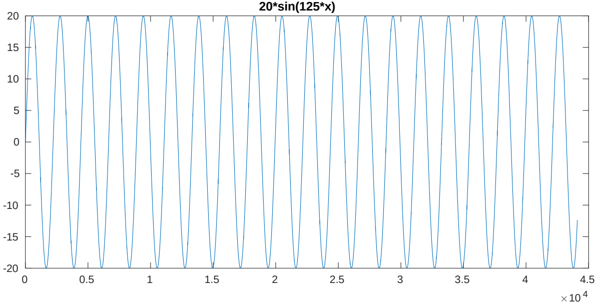


Figure: Wave of $sin(125{\cdot}x)$, $T \approx 0.05$, $F \approx 20$.

Sound of $20{\cdot}sin(125{\cdot}x)$

Period, Frequency and Sampling rate (5)

Sound of $20 \cdot sin(125 \cdot x)$, SF=44100

Sound of $20 \cdot sin(250 \cdot x)$, SF=44100

Sound of $50 \cdot sin(250 \cdot x)$, SF=44100

Sound of $50 \cdot sin(250 \cdot x)$, SF=35000

Mono Records and Sterero Records (1)

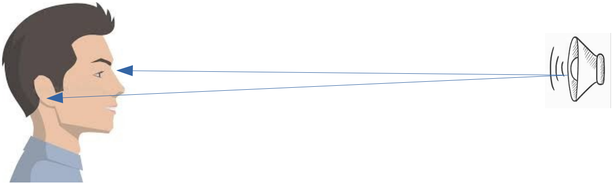


Figure: Two sound waves from the same source come into two hears.

- It is a typical stereo sound
- Similar as stereo image view, we are able to locate the source of sound
- In music recording, it could be single channel (mono), or multiple channels

Mono Records and Sterero Records (2)

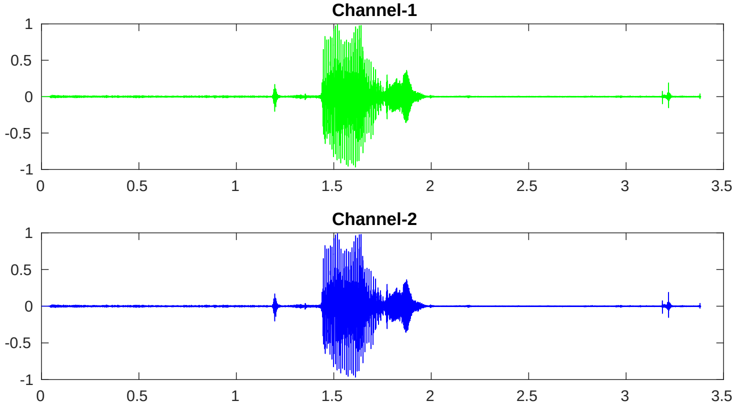


Figure: Sound from two channels.

- When convert the sounds from two channels into one
- We take the average

Size of a sound file (1)

- Given the sampling frequency is 44100 Hz
- Each sampled sound is encoded with 16 bits
- For a two minutes sound wave, what is the size of the sound wave kept in computer?
- Please work it out ....

Size of a sound file (2)

- Given the sampling frequency is 44100 Hz
- Each sampled sound is encoded with 16 bits
- For a two minutes sound wave, what is the size of the sound wave kept in computer?
- Please work it out ....

The answer is 21.17M bytes.

Music Formats inside Computer

- There are many music formats, the popular ones
  1. wav, comparable to BMP in image, it is also compressed
  2. mp3, mp4
  3. CD, for music lover
  4. WMA, from YAMAHA, well supported by Windows
  5. MIDI, keeps data allows sound card to reproduce music from various digital instruments
  6. RM, standard format from RealPlayer
- Music Player and Editor
  1. RealPlayer
  2. Windows Media Player
  3. ffplayer, from ffmpeg pakage
  4. Audacity: player and editor, it is free

Major Research Subjects Related to Sound and Music

- Low level
  1. Music/Sound Search
  2. Music Search by Humming
  3. Genre classification
- High Level
  1. Speech Recognition
  2. Voice Reognition
  3. Machine Translation
  4. Speech Synthesize

Music Genre Classification

Speech Recognition

Machine Translation

Speech Synthesize

Voice Recognition

Outline

Content based Music/Sound Search

- Problem Statement
  1. Given a piece of sound/music wave
  2. find out similar/relevant sound/music from a repository
- Solution
  1. Work out appropriate representation for sound/music
  2. Perform fast comparison between query and the features in the repository
- The focus will be on feature representation

Overview of Spectrogram and Cepstrum

- Sound wave in time domain is hard to analyze
- It is easier for us to perceive it in frequency domain
- Spectrogram shows the energy distribution in a spectrum of frequencies
- Cesptrum is the inverse of Spectrum, which is more similar to our hearing system

Spectrogram of a Sound Wave

- It shows the energy distribution of Sound Wave in its frequency domain
- Two major steps to produce the view
  1. Cut sound wave into fixed size frames with overlapping
  2. Appy Short Time Fourier Transform on each frame

"Sound of Silience" from my student, SF=44100

Spectrogram of a Sound Wave: wave to frames (1)

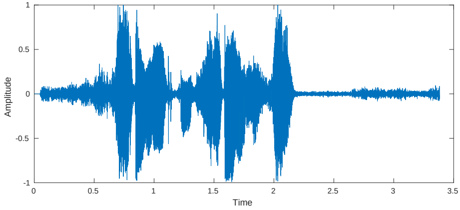


Figure: Wave of "Sound of silience" (one channel).

```
[signal, FS] = audioread('./silience.wav');
ts = size(signal);
sig_len = ts(1);
t = ts(1)/FS; %duration of the wave
tm = [0:t/ts(1):t-t/ts(1)];
plot(tm, signal(:,1));
```

Spectrogram of a Sound Wave: wave to frames (2)

```
framesz    = 0.02; %duration of one frame
stride     = 0.01; %overlapping between two frames
framelen   = floor(framesz*sig_len);
fstep      = stride*sig_len;
nframes    = ceil(abs(sig_len - framelen) / fstep);
pad_len    = nframes * fstep + framelen;
pad        = zeros(1, round(pad_len - sig_len));
pad_signal = [signal, pad];
```

Listing 1: frame size, step size and padding

- Calculate frame length
- Calculate step size
- Calculate size of padding

Spectrogram of a Sound Wave: wave to frames (3)

```
fmat = pad_signal(1:framelen);
p  = fstep;

for i = 2:nframes
    frame = pad_signal(p:(p+framelen-1));
    fmat  = [fmat;frame];
    p     = p + fstep;
end
```

Listing 2: extract frames

- Cut the signals into frames
- With 'fstep', the overlapping between two consecutive frames is allowed

Spectrogram of a Sound Wave: short time FT

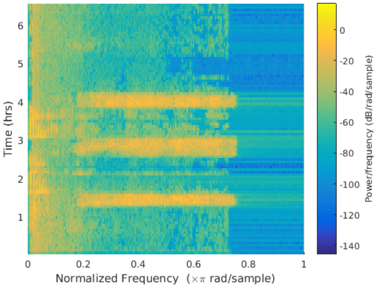
```
N   = 1024;
ffmat = []
for m = 1:nframes
    fframe = abs(rfft(fmat(m,:), N));
    fframe = 20*log10(fframe);
    ffmat  = [ffmat; fframe];
end
```

Listing 3: perform STFT

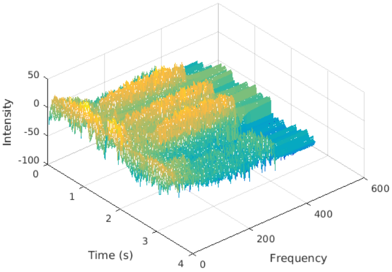
$$x(k) = \sum_{k=1}^{N} x(n) \cdot e^{\frac{-j \cdot 2 \cdot \pi \cdot (k-1) \cdot (n-1)}{N}}, 1 \leq k \leq N. \quad (1)$$

- The STFT signals are further taken $|\cdot|$ and $log10(\cdot)$
- Now each frame are transformed into a spectrum of frequencies

Spectrogram of a Sound Wave: visualize



(a) Heatmap of spectrogram



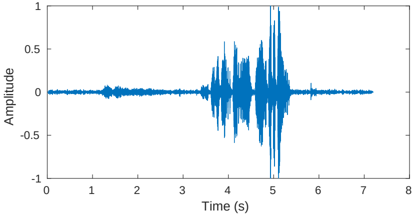
(b) 3D view of spectrogram

Figure: Spectrogram of "sound of silience".

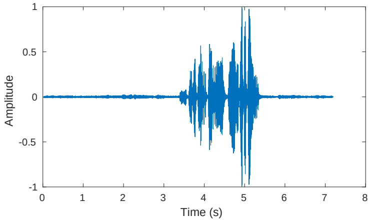
```
hmap = HeatMap(ffmat');
```

Listing 4: visualize the spectrogram

A Study Case (1)



(a) Sound from two species



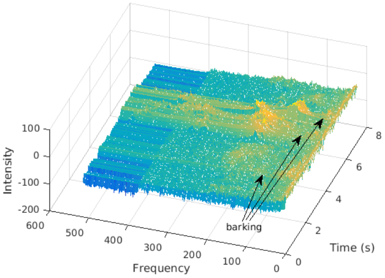
(b) Sound from one species

Figure: Sound wave from species.

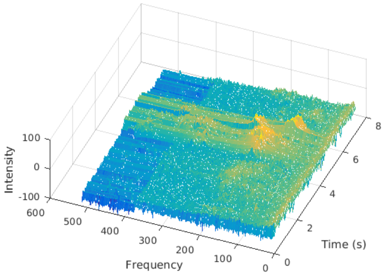
Sound-1

Sound-2

A Study Case (2)



(a) Spectrogram from two species



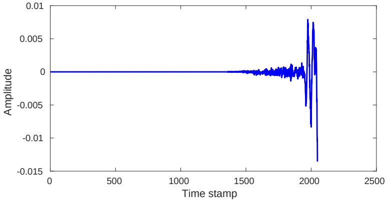
(b) Spectrogram from one species

Figure: Spectrogram of sound from species.

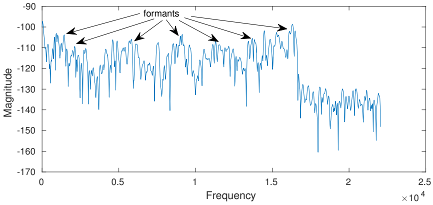
- Phones could be easier to be observed in spectrogram

Outline

An Overview of Mel Freuqency Cepstrum Co-efficients (1)



(a) Sound of 'voice'



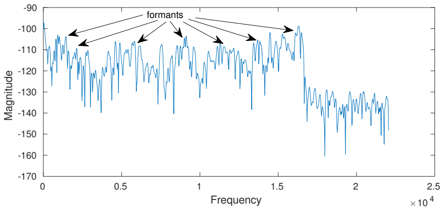
(b) Periodogram of sound of 'voice'

Figure: Periodogram and the formants.

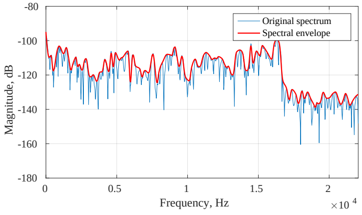
- **Paul Mermelstein** is believed to be the inventor[1]
- **Paul Mermelstein** attributed the invention to **Bridle** and **Brown**.
- MFCC aims to capture the formants (frequency peak) in sound wave

[1] Distance measures for speech recognition, psychological and instrumental, in Pattern Recognition and Artificial Intelligence, 1976.

An Overview of Mel Freuqency Cepstrum Co-efficients (2)



(a) Periodogram of sound of 'voice'

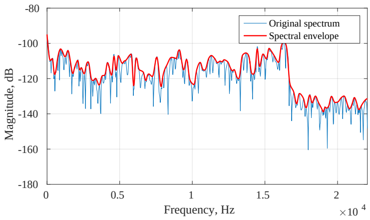


(b) Envelope of the wave

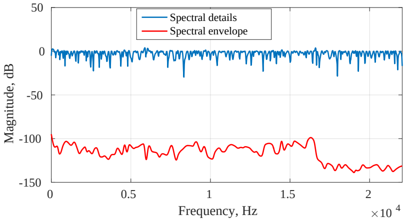
Figure: Periodogram and the formants.

- MFCC aims to capture the formants (frequency peak) in sound wave
- The characteristics of voice from a certain source, e.g. man or animal
- Before deep learning, it is the most popular feature of sound

Mel Freuqency Cepstrum Co-efficients: the theory (1)



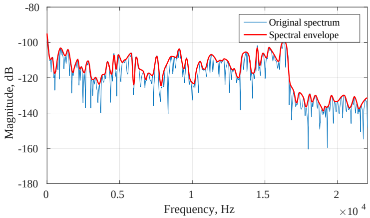
(a) Envelope of the wave



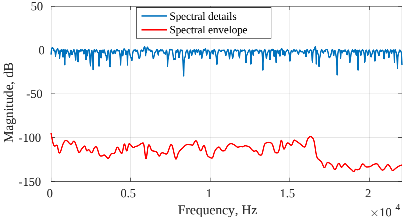
(b) Envelope and details

- Given $x(k)$ is the signal after Fourier transform
- We want have $x(k) = h(k) \cdot e(k)$
- $h(k)$ is the envelope signal
- $e(k)$ is the details
- Problem: how to distill $h(k)$ out

Mel Freuqency Cepstrum Co-efficients: the theory (2)



(c) Envelope of the wave



(d) Envelope and details

- $h(k)$ corresponds to low frequency signal in the curve
- $e(k)$ corresponds to high frequency signal
- Apply *log* on $x(k) = h(k) \cdot e(k)$, we have

$$log(x(k)) = log(h(k)) + log(e(k)) \tag{2}$$

Mel Freuqency Cepstrum Co-efficients: the theory (3)

- $h(k)$ corresponds to low frequency signal in the curve
- $e(k)$ corresponds to high frequency signal
- Apply *log* on $x(k) = h(k) \cdot e(k)$, we have

$$log(x(k)) = log(h(k)) + log(e(k)) \tag{3}$$

- Since we care only about the magnitude, we have

$$|x(k)| = |h(k)| \cdot |e(k)| \tag{4}$$

$$log(|x(k)|) = log(|h(k)|) + log(|e(k)|) \tag{5}$$

- As we take the low frequency part of $log(|x(k)|)$, we get $log(|h(k)|)$

Mel Freuqency Cepstrum Co-efficients: implementation (1)

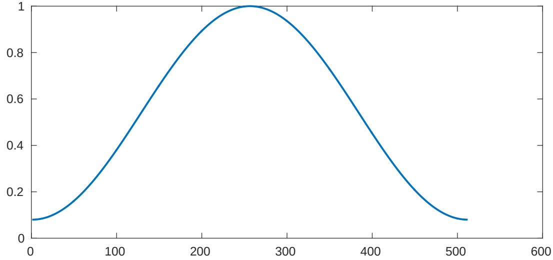


Figure: Hamming Window.

- Hamming window is used to emphasize the signals in each frame
- It is applied on each frame
- After this step, Fourier transfrom is applied on each frame

Mel Freuqency Cepstrum Co-efficients: implementation (2)

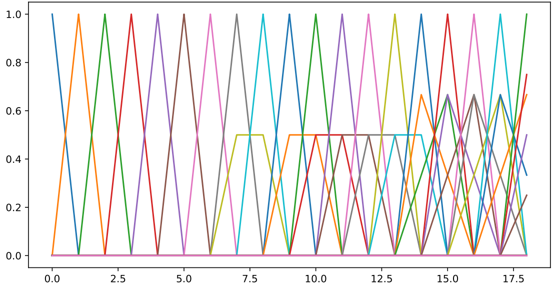


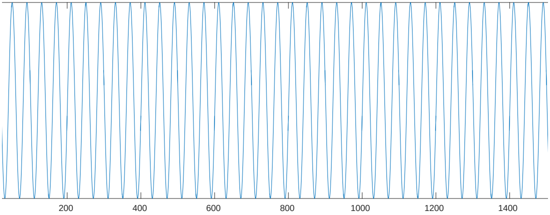
Figure: 20 Triangle fitlers together.

- Only keeps the signals within a frequency range
- Filters are denser on low frequencies
- Triangle filters are applied on aboslute signal after Fourier transform
- After triangle filtering, DCT is applied

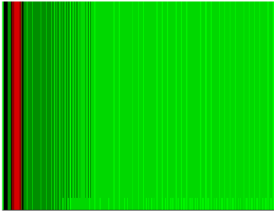
Mel Freuqency Cepstrum Co-efficients: the major steps

1. Cut sound wave into frames
2. Apply Hamming window
3. Perform short time Fourier Transform
4. Apply triangle filtering
5. Perform $log(x(k))$
6. Perform DCT on $log(x(k))$

Mel Freuqency Cepstrum Co-efficients: case study (1)



(a) $sin(400 \cdot \pi)$

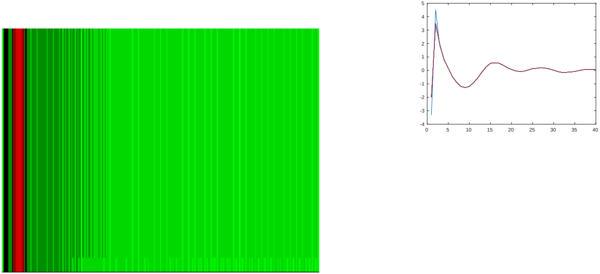


(b) after STFT

Figure: The first two steps of MFCC.

1. Cut sine wave into frames with 40% overlaps
2. Apply Hamming window on each frame
3. Apply Fourier transform, $NFFT = 512$
4. abs(.) and log10(.) on transformed signal

Mel Freuqency Cepstrum Co-efficients: case study (2)

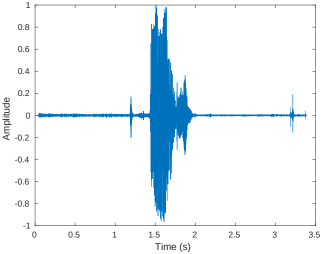


(a) $sin(400 \cdot \pi)$ (b) after STFT

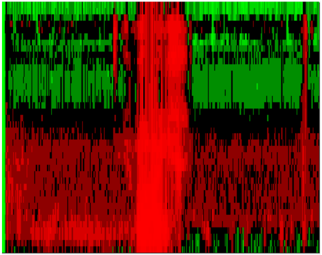
Figure: From SFT signal to MFCC.

1. Build triangle filters bank

Mel Freuqency Cepstrum Co-efficients: case study (3)



(a) Sound of “voice”



(b) after MFCC

Figure: Sound “voice” and its MFCC.

Toolkits for Audio Processing

- librosa: Python
- MIR Toolbox: Matlab
- YAAFE: Python and C++
- Timbre Toolbox: matlab

Q & A