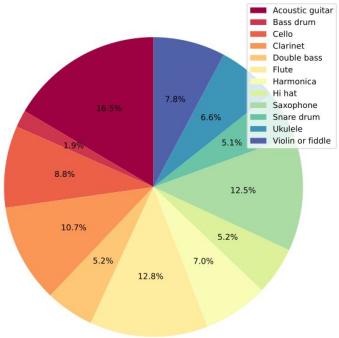
Classification of Music Based on Machine Learning

Master's Thesis no. 2273

Luka Čupić

Faculty of Electrical Engineering and Computing, University of Zagreb

Zagreb, July 8, 2020.


Acoustic guitar
Bass drum
Cello
Clarinet
Double bass
Flute
Harmonica
Hi hat
Saxophone
Snare drum
Ukulele
Violin or fiddle
16.5%
1.9%
8.8%
10.7%
5.2%
12.8%
7.0%
5.2%
12.5%
5.1%
6.6%
7.8%

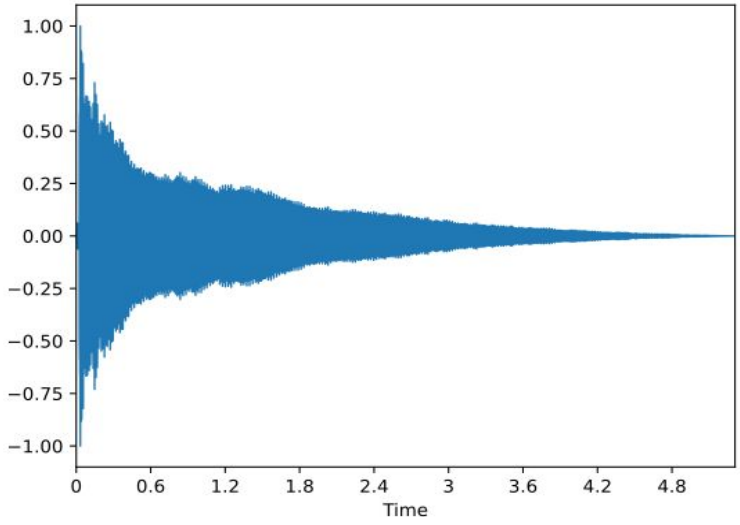
“Problems” with audio-based machine learning:

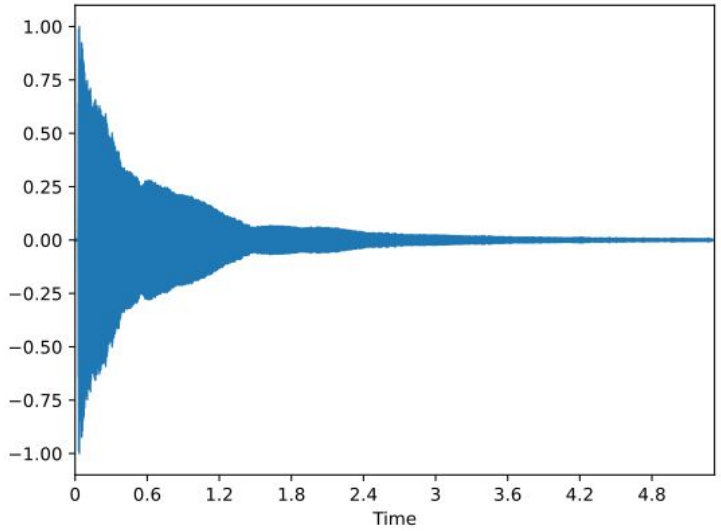
- Most ML research is **not** about audio
- Audio classification is difficult

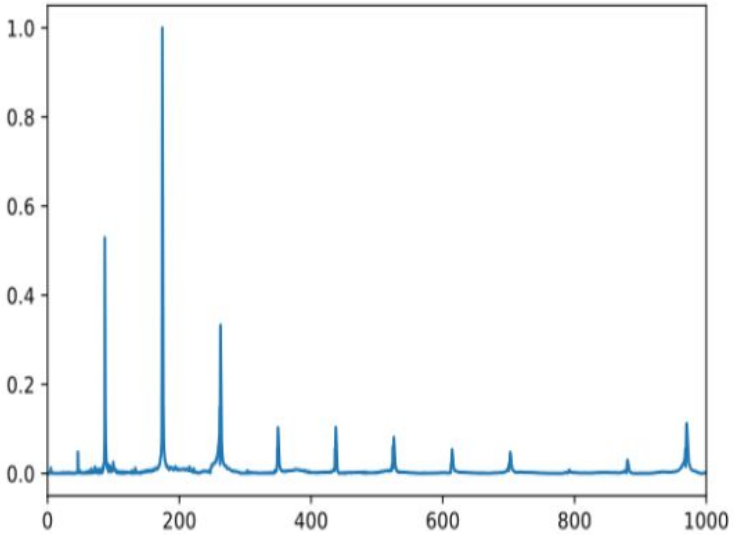
Why is audio classification difficult?

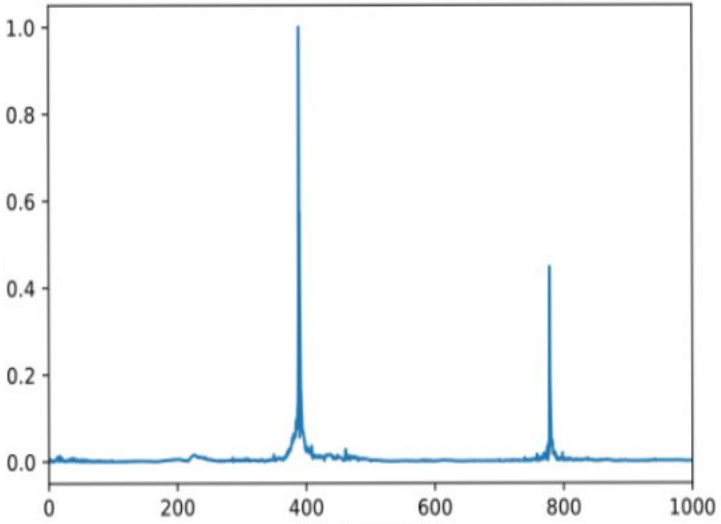
- Information contained in audio is abstract
- Audio essentially has no “meaning”

How to capture “meaning” and information from audio?


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time

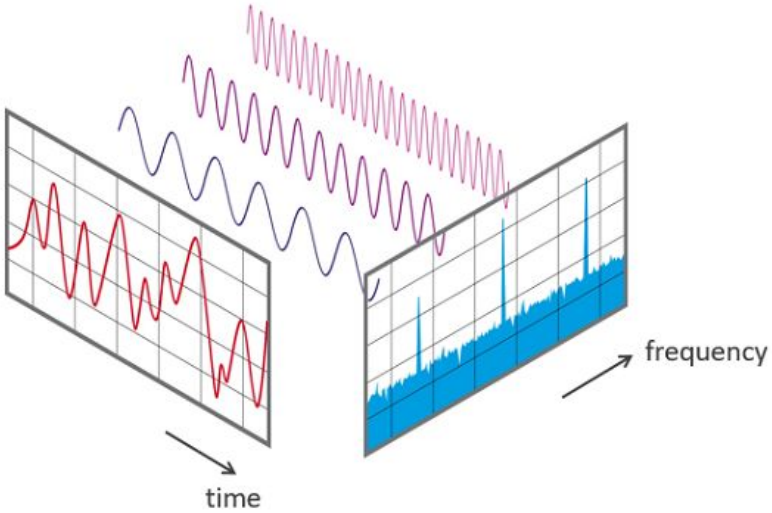

1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000


1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000

Time domain → Frequency domain:

- **Fourier Transform (FT)**: decomposition of a signal into its basic frequencies
- **Discrete Fourier Transform (DFT)**: Fourier Transform for discrete signals
- Result is a **periodogram**

Time domain → Frequency domain:



Feature extraction:

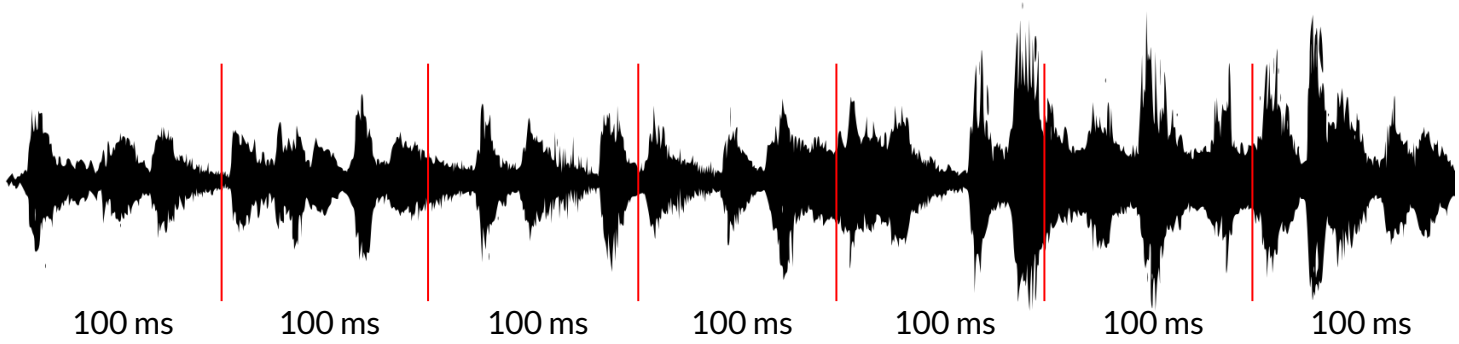
- Transforming original audio data into a more suitable form for machine learning
- Using DFT and other techniques to extract relevant information from audio
- **Mel-frequency cepstral coefficients (MFCCs)**: state-of-the-art for audio

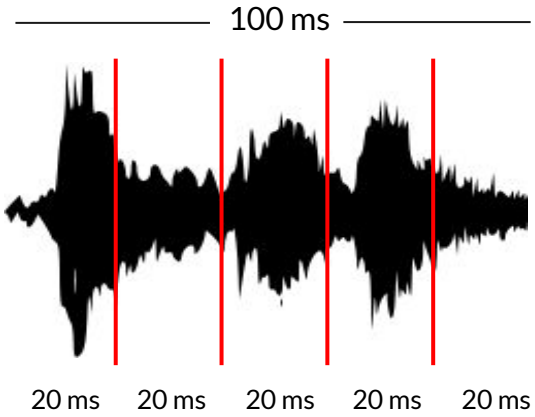
Raw audio → useful features (MFCCs):

1. Prepare raw audio
2. Perform Fourier Transform on prepared audio
3. Non-linearity of human hearing
4. Discrete Cosine Transform

Raw audio → useful features (MFCCs):

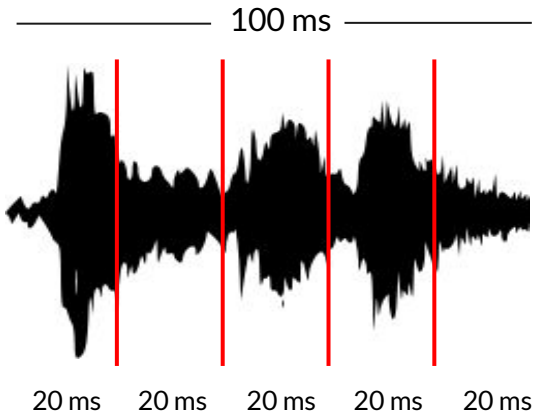
1. **Prepare raw audio**
2. Perform Fourier Transform on prepared audio
3. Non-linearity of human hearing
4. Discrete Cosine Transform


100 ms
100 ms
100 ms
100 ms
100 ms
100 ms
100 ms

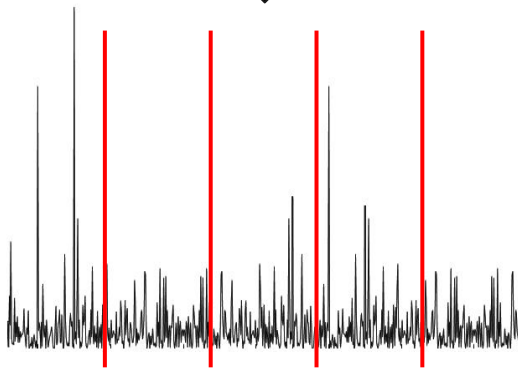

100 ms
20 ms
20 ms
20 ms
20 ms
20 ms

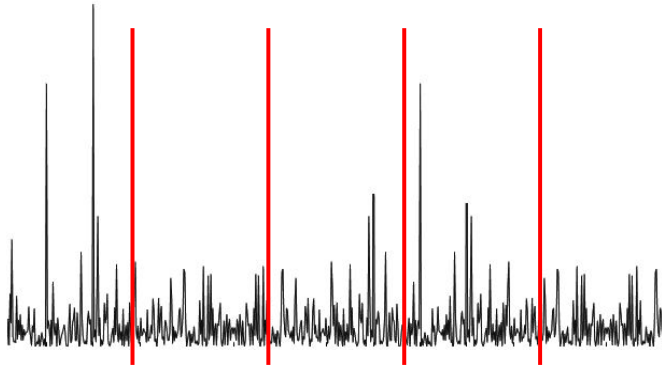
Raw audio → useful features (MFCCs):

1. Prepare raw audio
2. **Perform Fourier Transform on prepared audio**
3. Non-linearity of human hearing
4. Discrete Cosine Transform

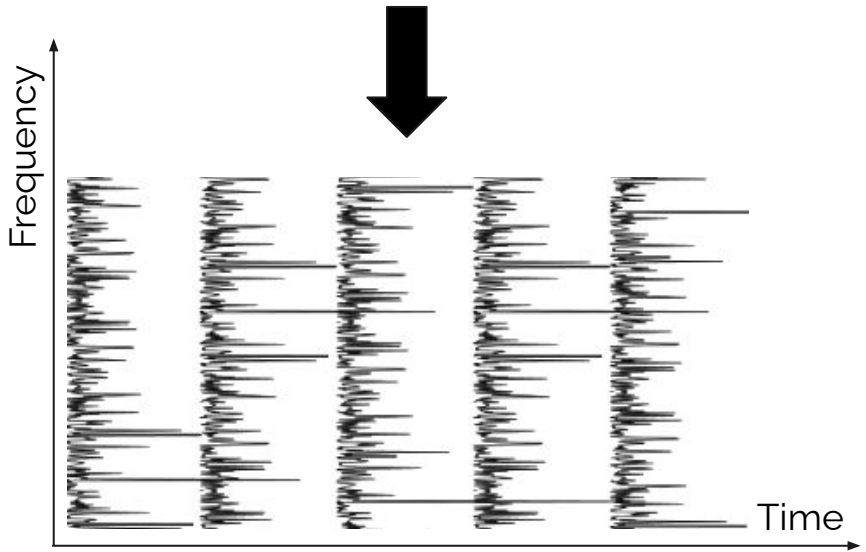


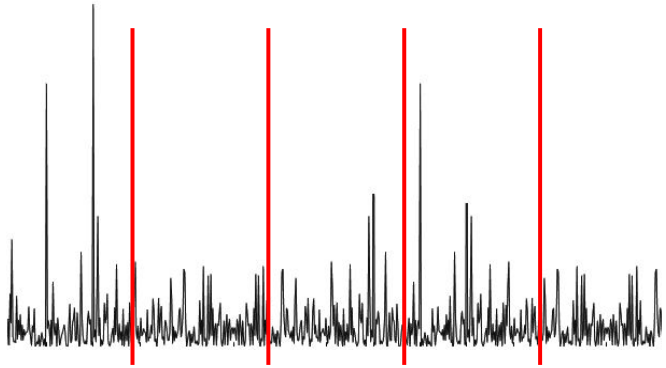
Discrete Fourier Transform



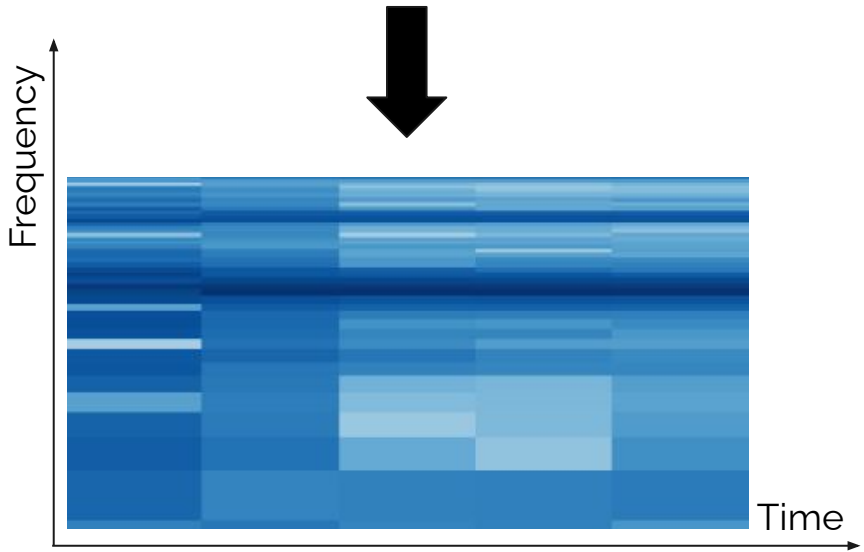


Short-time Fourier Transform





Short-time Fourier Transform



STFT:

- Produces a **spectrogram**
- Gives information about how frequencies change through time for each 100 ms block of audio

Raw audio → useful features (MFCCs):

1. Prepare raw audio
2. Perform Fourier Transform on prepared audio
3. **Non-linearity of human hearing**
4. Discrete Cosine Transform

Hearing Experiment:

 300 Hz

 5300 Hz

 400 Hz

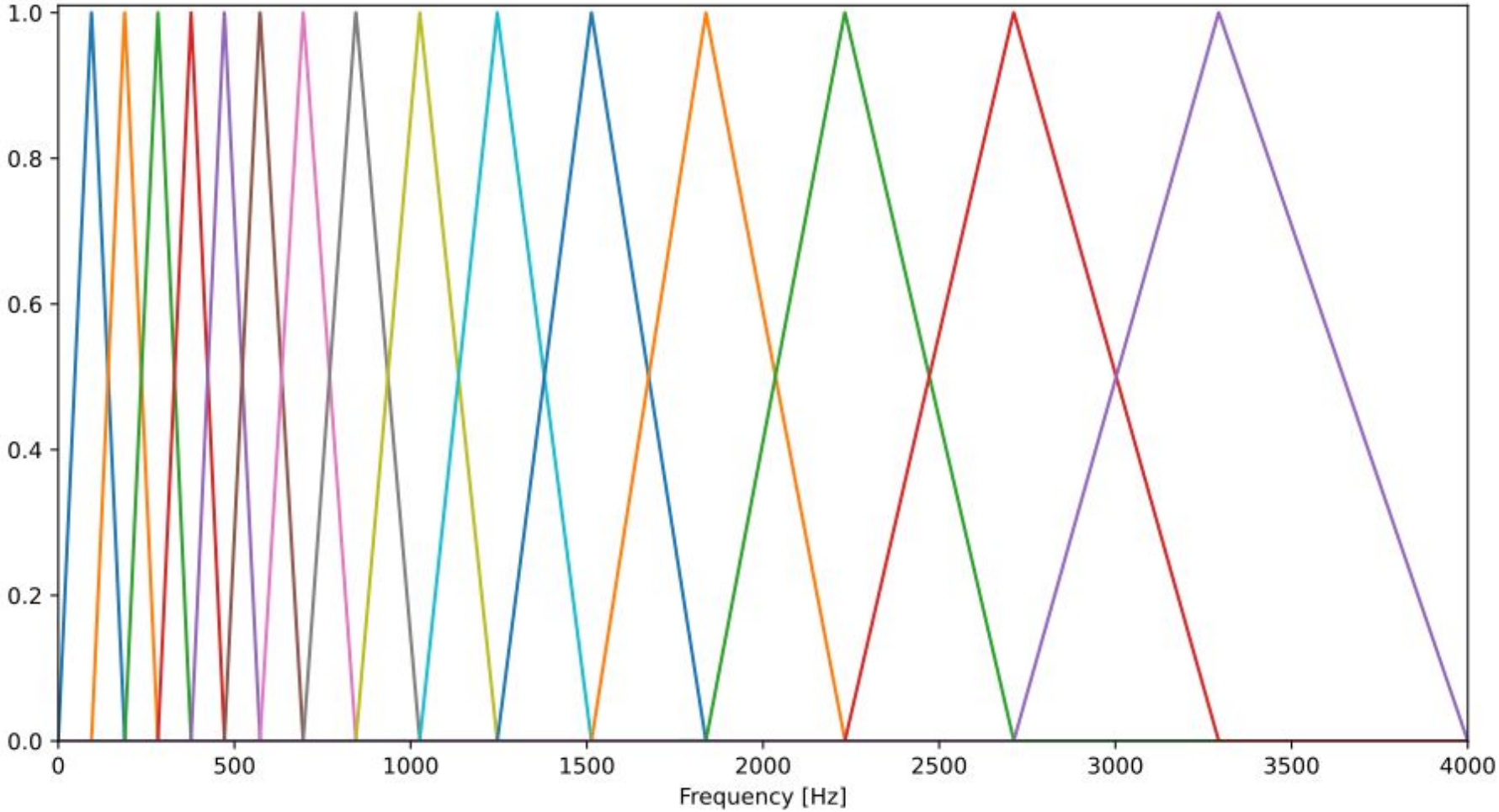
 5400 Hz

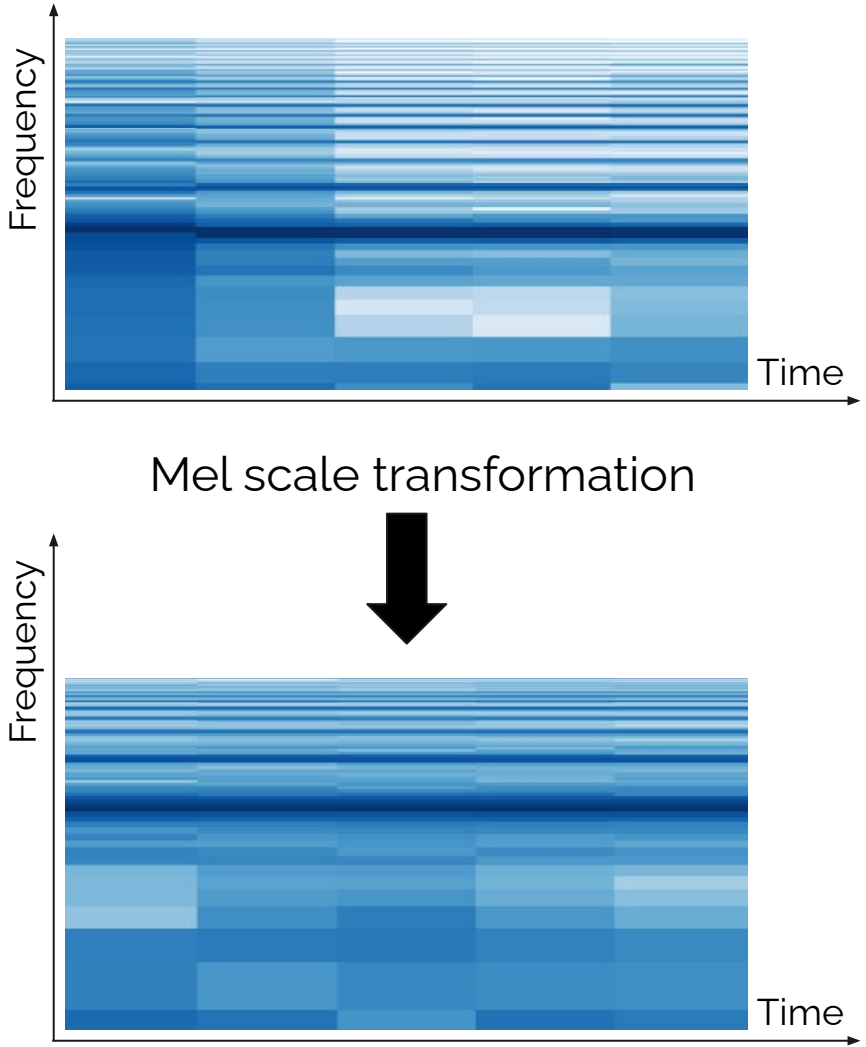
Conclusion:

- Humans are bad at perceiving high frequencies

Solution:

- **Mel scale**: rescales the audio to mimic non-linear human perception of sound
- More discriminative at lower frequencies and less discriminative at higher frequencies


1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
3000
3500
4000
Frequency [Hz]

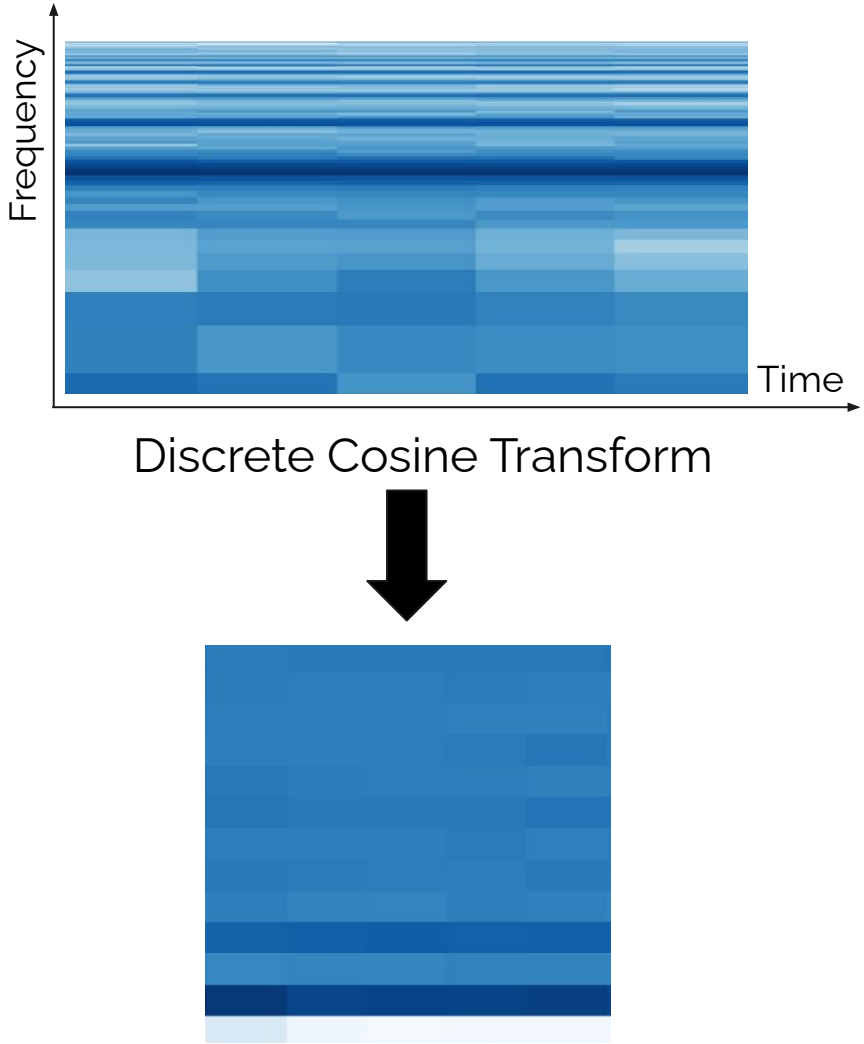

Frequency
Time
Mel scale transformation
Frequency
Time

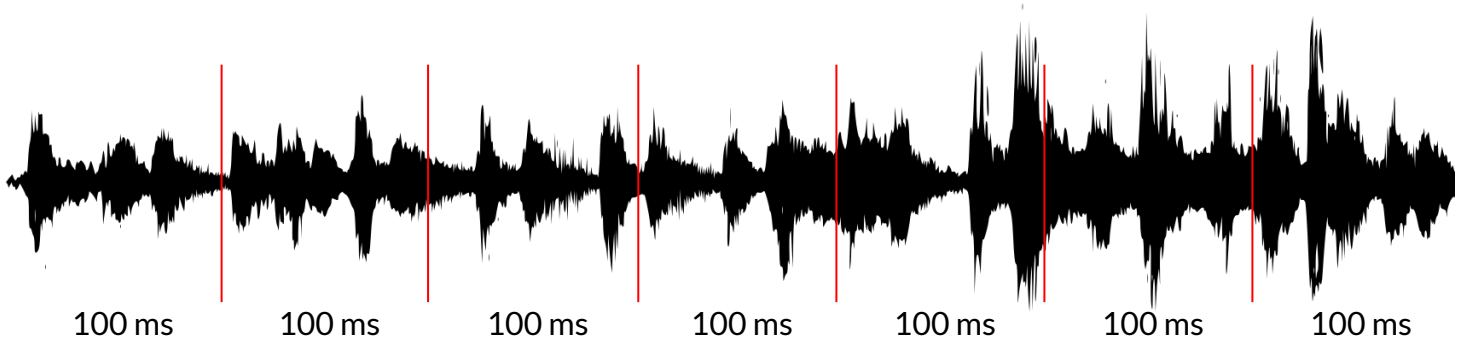
Raw audio → useful features (MFCCs):

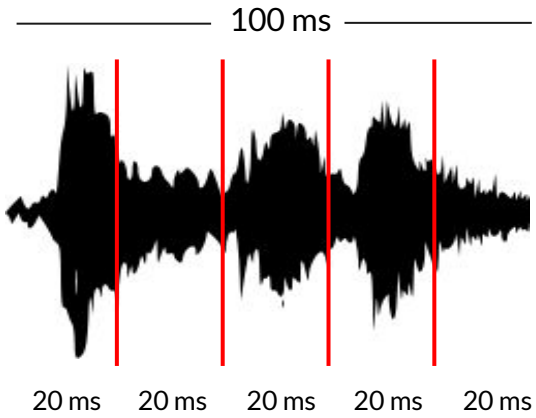
1. Prepare raw audio
2. Perform Fourier Transform on prepared audio
3. Non-linearity of human hearing
4. **Discrete Cosine Transform**

DCT:

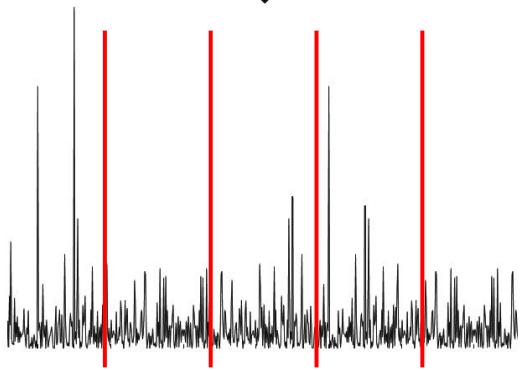
- Final step in feature extraction
- Better fits the shape of the resulting spectrum
- Keeps only lower-order coefficients because higher-order coefficients contain noise

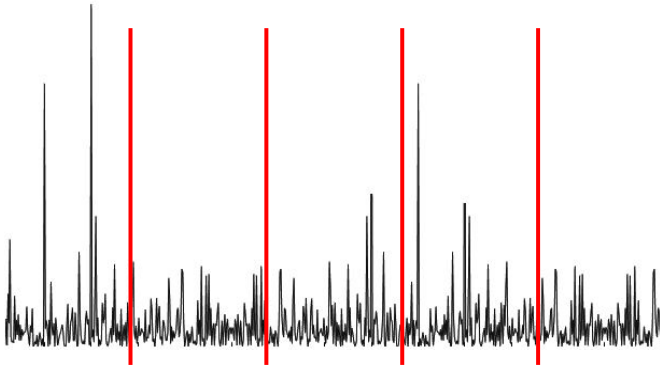

Frequency
Time
Discrete Cosine Transform


100 ms
100 ms
100 ms
100 ms
100 ms
100 ms
100 ms

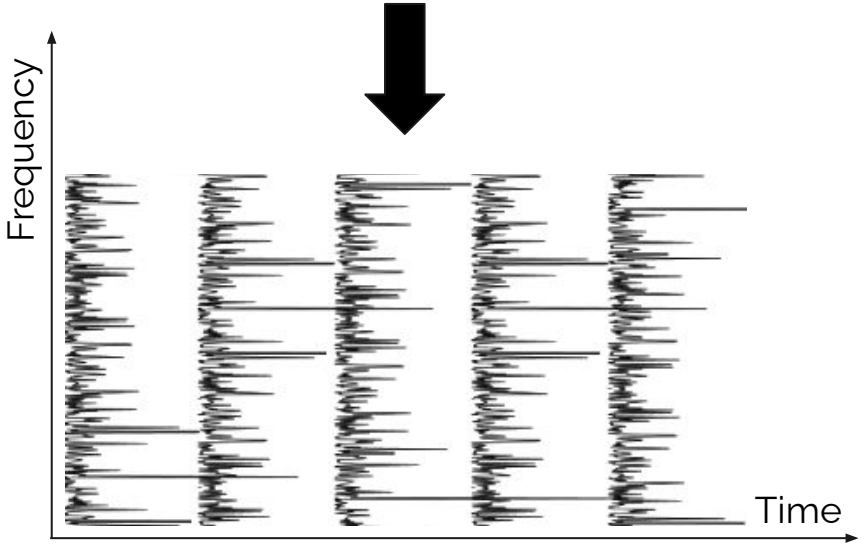


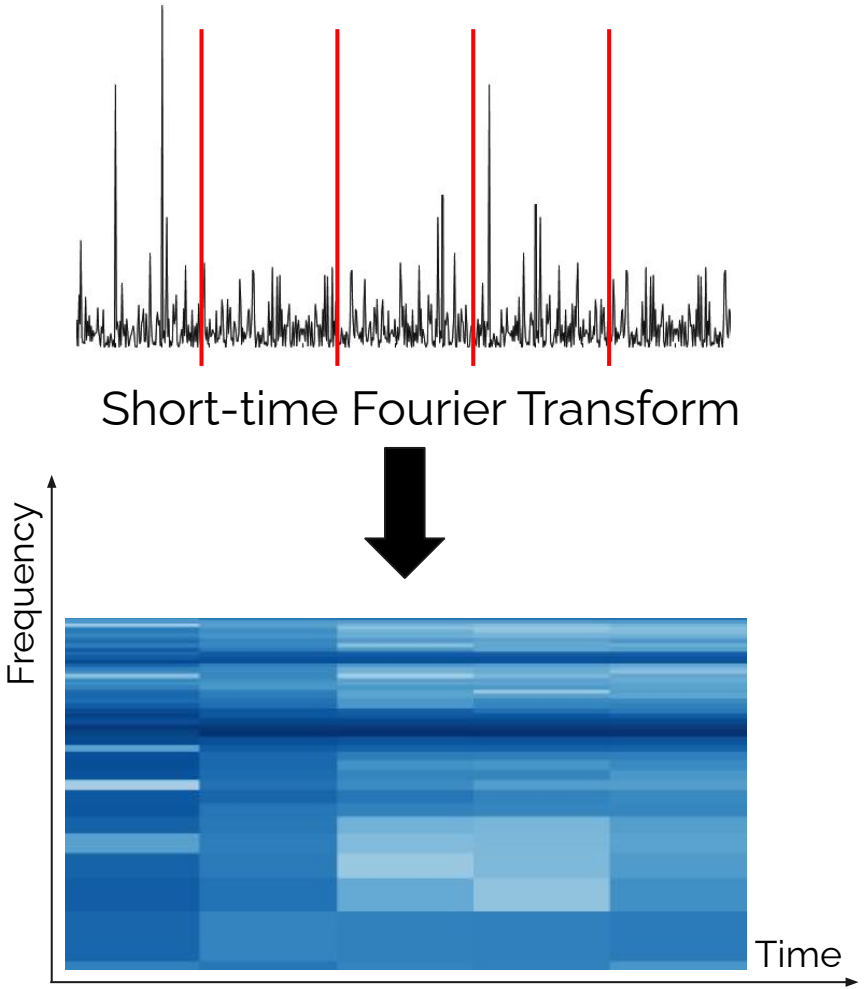
Discrete Fourier Transform

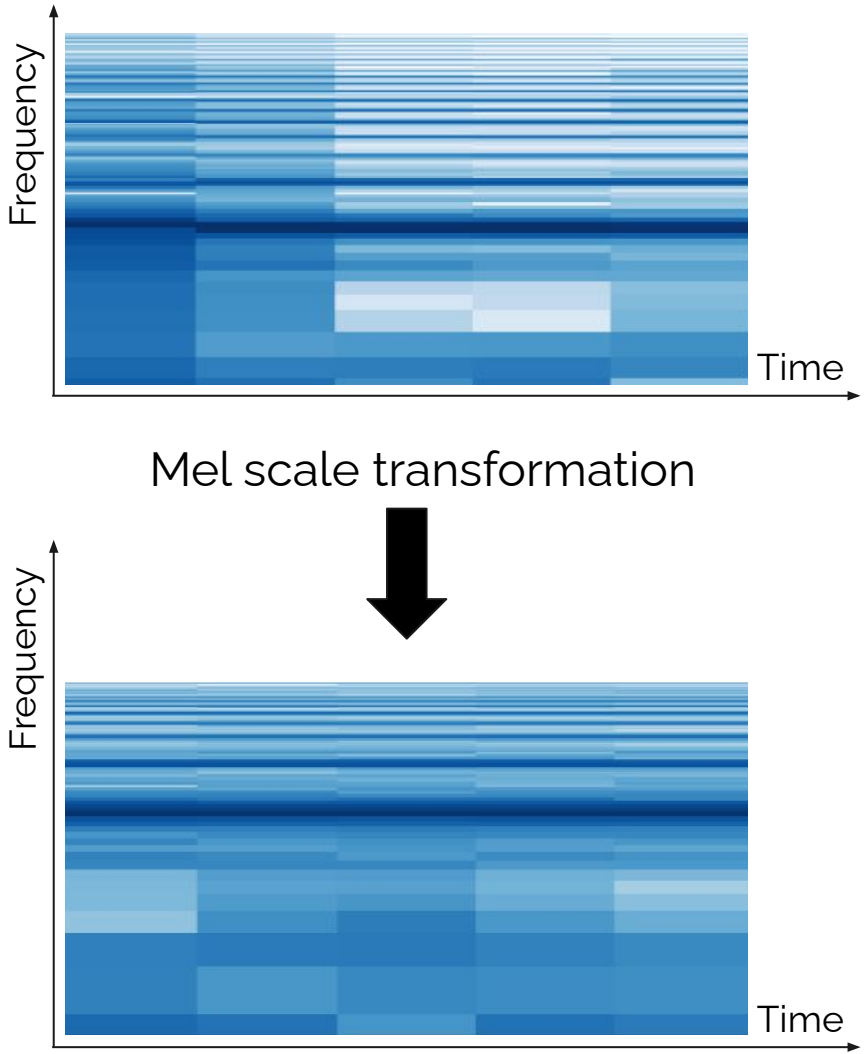


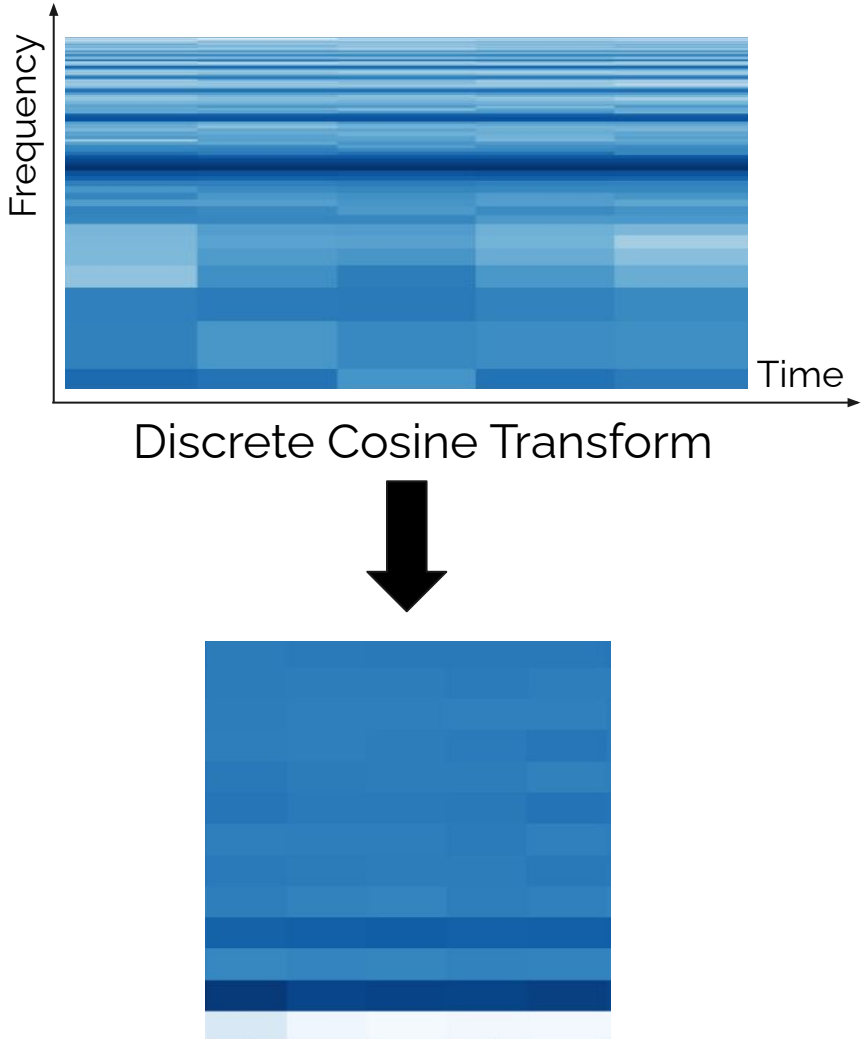


Short-time Fourier Transform

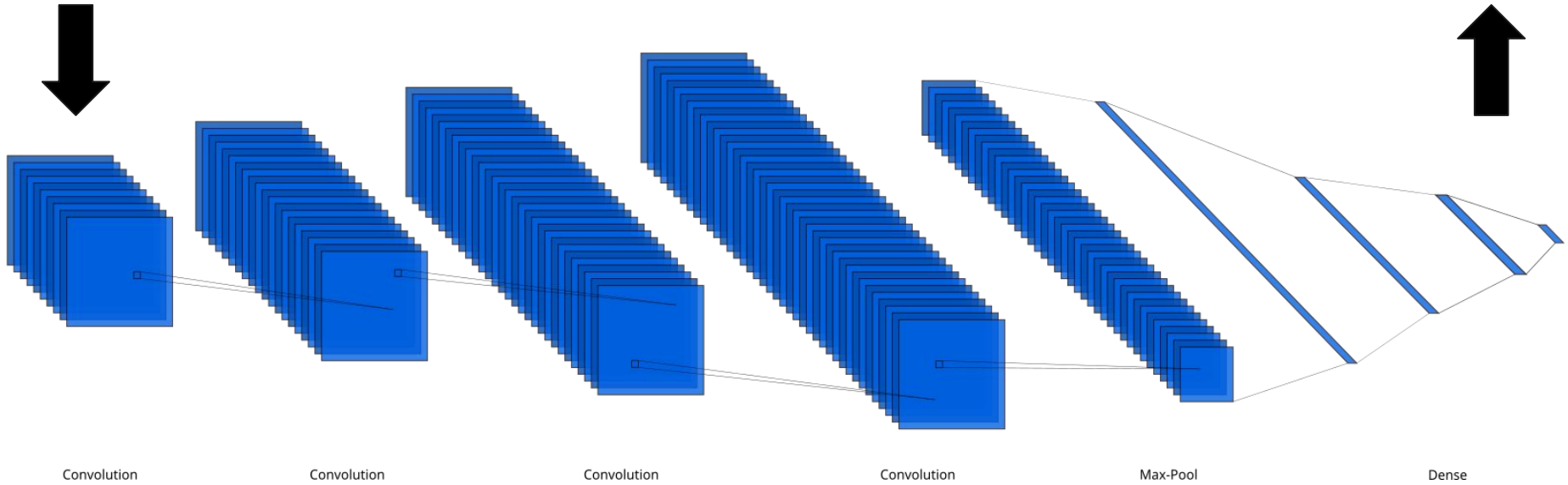


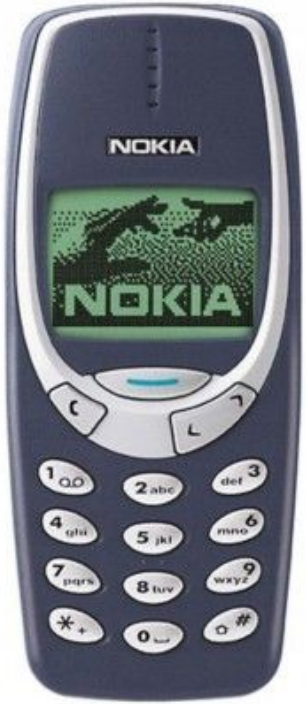

Short-time Fourier Transform
Frequency
Time


Frequency
Time
Mel scale transformation
Frequency
Time


Frequency
Time
Discrete Cosine Transform

[0.54, 0.02, 0.37, ⋯, 0.01]





97.3%

2.6%

0.1%

DEMO TIME

Results and observations:

- Validation accuracy **97.46%**
- Theoretical inference is different from practical

Conclusion:

- Audio classification is an interesting area of research with plenty of potential
- Music is one of the more interesting applications, but this can be used for **any type of audio**

Thank you. Questions?