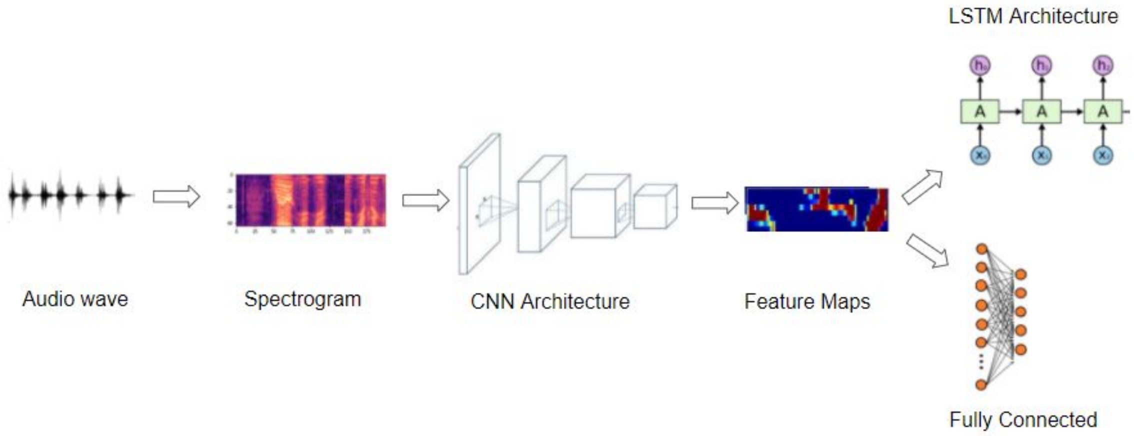
Exploring the best way to represent an audio signal for ML model

On the example of audio classification

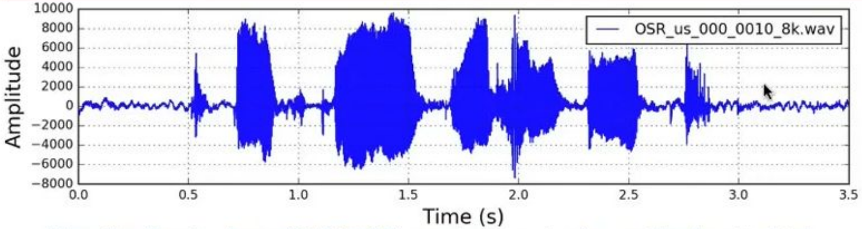
- ★ Raw signal data
- ★ Short-Time Fourier-Transform (STFT)
- ★ Mel-spaced Filterbank
- ★ Mel Spectrogram
- ★ Discrete Cosine Transform (DCT)



Looking closely at audio signal post-processing

Boulder

Mel-Spectrogram and Mel-Frequency Cepstral Coefficients (MFCCs)



https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

$X_i \in \mathbb{R}^{200} \to i\text{-th frame of signal } x \text{ (25ms frames)}$

$80 \to \text{frame step (10ms)} \implies X \in \mathbb{R}^{200\times 350} \; (350 = (28,200-200)/80)$

$\widetilde{X}_i \in \mathbb{C}^K \to \text{discrete Fourier transform of } X_i \implies \widetilde{X} \in \mathbb{C}^{K\times 350}$

$$\widetilde{X}_i(k) = \sum_{n=1}^{N} X_i(n) g(n) e^{-j2\pi kn/N}, k = 1, \ldots, K \qquad N = 200$$

$K = 257 \to \text{number of discrete Fourier transform coefficients}$

$P_i(k) = \frac{1}{N}|\widetilde{X}_i(k)|^2 \to \text{Periodogram estimate of the power spectrum}$

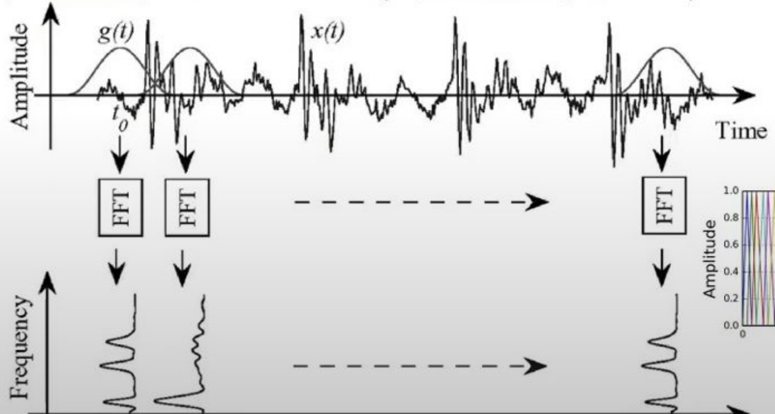
$\implies P \in \mathbb{R}^{257\times 350}$

Short-Time Fourier-Transform (STFT)

$x \to$ signal in the time domain (e.g., $x \in \mathbb{R}^{28,200=3.525 \text{ seconds}\times 8,000 \text{ Hz}}$)

sampling frequency (number of samples per seconds)

$g \to$ sliding window function (e.g., Hamming function)



Mel-spaced Filterbank

http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

https://www.youtube.com/watch?v=WJI-17MNpdE

$300 \text{ Hz} \to$ lower frequency

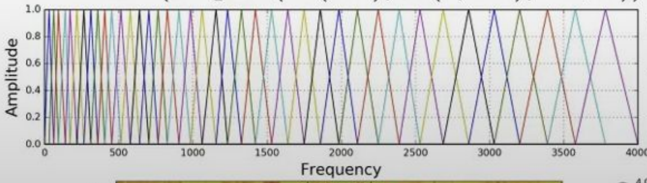
$4,000 \text{ Hz} \to$ upper frequency

$M(f) = 1,125 \ln(1 + f/700) \to$ convert frequency to Mel scale

$M^{-1}(m) = 700(\exp(m/1,125) - 1) \to$ convert Mel scale to Hz

$M^{-1}(\text{linspace}(M(300), M(4,000), 26+2))$

$26 \to$ number of triangular filters



$T \in \mathbb{R}^{26\times 257} \implies E = TP \in \mathbb{R}^{26\times 350}$

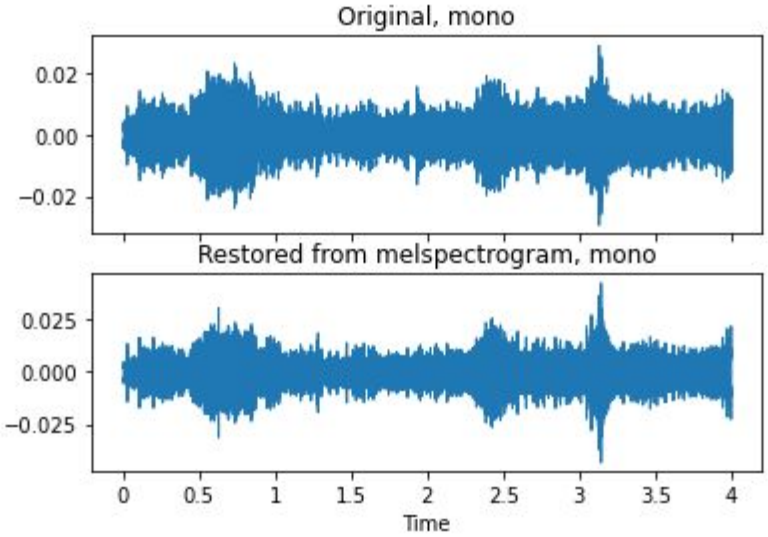
$E_i(l) \to$ amount of energy in filter bank l at frame i

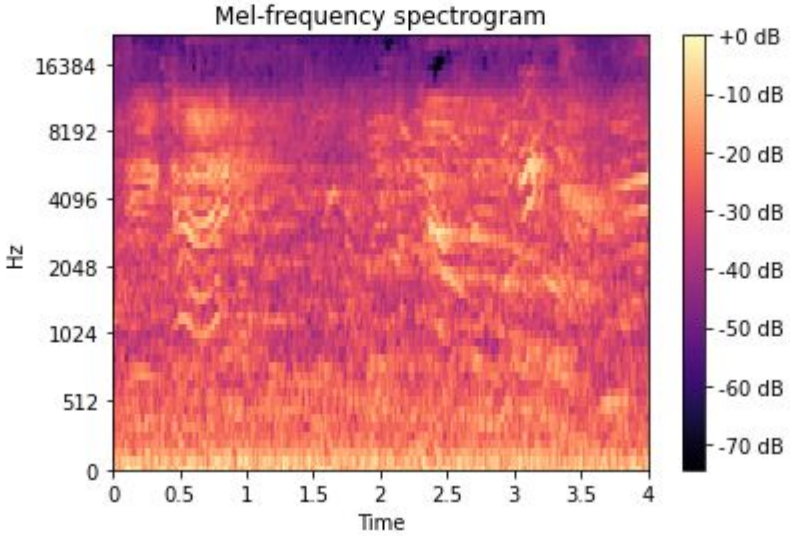
$\log(E) \in \mathbb{R}^{26\times 350} \to$ log filter bank energy

Mel Spectrogram

Mel-Spectrogram analysis

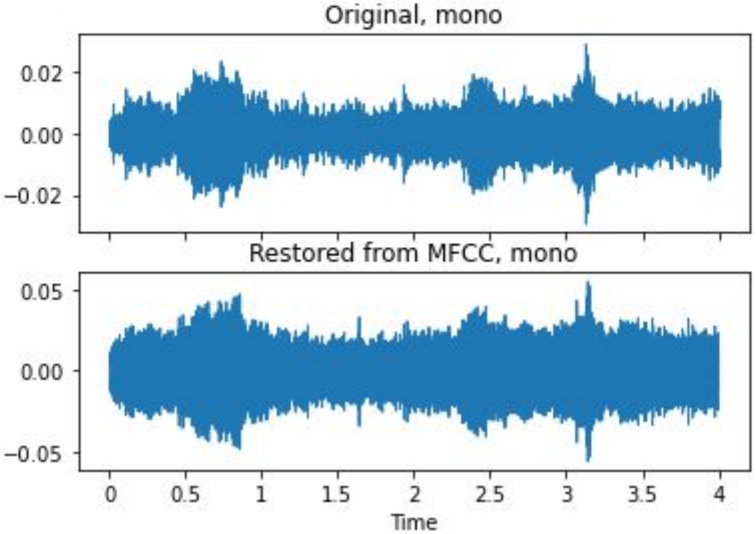
original memory use = 176400, melspectrogram = 22080

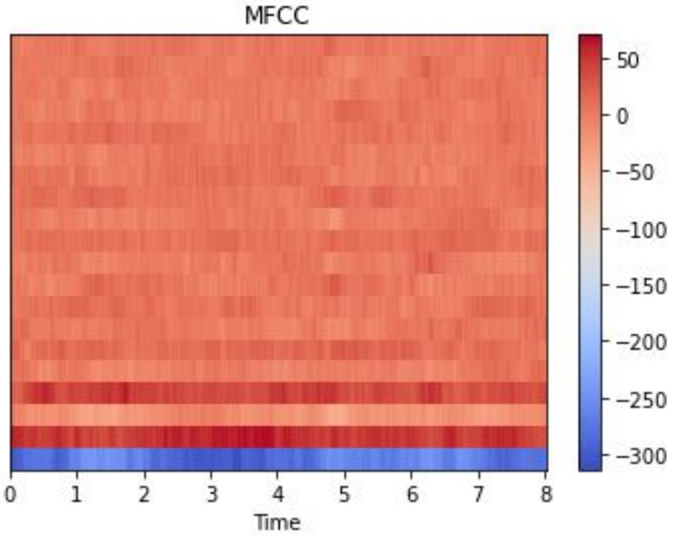




MFCC analysis

```
original memory use = 176400, MFCC = 6900
```





Mini-Conclusion, based on applying methods to signal

- If one wants to use MFCC as a input for neural network against Mel-Spectrogram, s/he wins approximately **4x** in space (usually it is an input for the first convolutional layer)
- Nobody uses raw data for the input in the latest researches, as it seems hard for a neural network extract features from such an input
- each of methods have information lost after its application

Evaluating feature extraction of two methods: MFCC vs Mel-Spectrogram – Results

For a paticular Urban8k classification task MFCC performs better (0.77 > 0.65) (accuracy scores)

It is useful to use MFCC instead of Mel-Spectrogram for a Neural Network to be able to increase the batch size if one wants a fast training

Research conclusion

- Mel-Spectrogram is commonly used for NLP, Transformers
- MFCC is commonly used for Reinforcement Learning, Hidden Markov Processes (when we care about memory)
- It is better to try both methods of pre-processing data for Neural Network, and then take the better.

Possible ways of future research:

a. As images can be represented, as signals, try to use Mel-Spectrogram or MFCC for its feature extraction

b. Train more specific architectures to explore more applications of these methods