Feature extraction for audio signals

Fundamentals of Acoustics and Sound
Topic 2

Christian Sejer Pedersen

AALBORG UNIVERSITY
DENMARK

Agenda

Quick discussion of exercises from last week

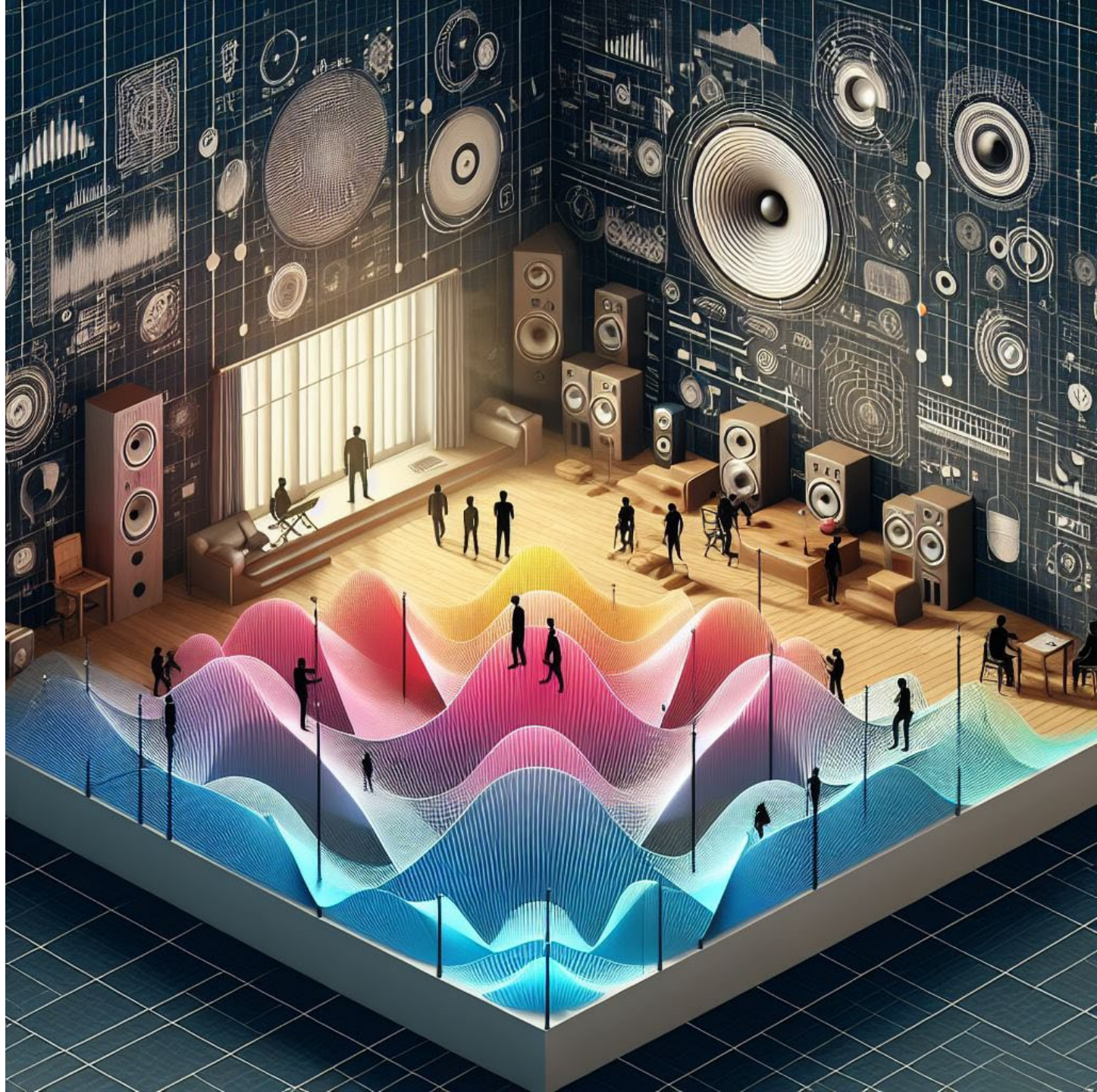
Common features for AI for audio signals

Time domain features

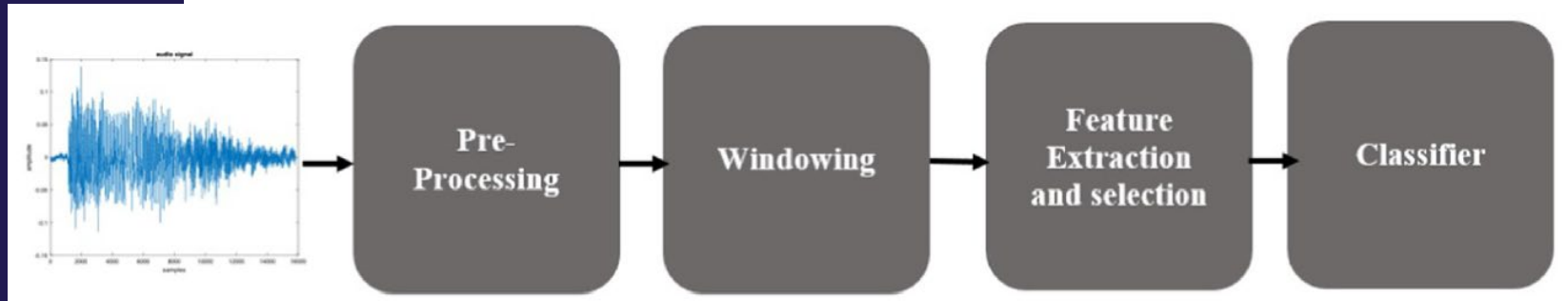
Frequency domain features

Time-frequency domain features

Deep features



Common audio features for AI



- Audio signals take up much space
- Extracting “useful” features reduces computational complexity and processing power needed for machine learning and deep learning
- The features must be compact, **but highlight the important characteristics of the signal!**
- Therefore, the choice of features and their chosen parameters are very important for any successful machine learning/deep learning!

Many features exist!

Garima Sharma et al. “Trends in audio signal feature extraction methods”, Applied Acoustics, Volume 158, 2020

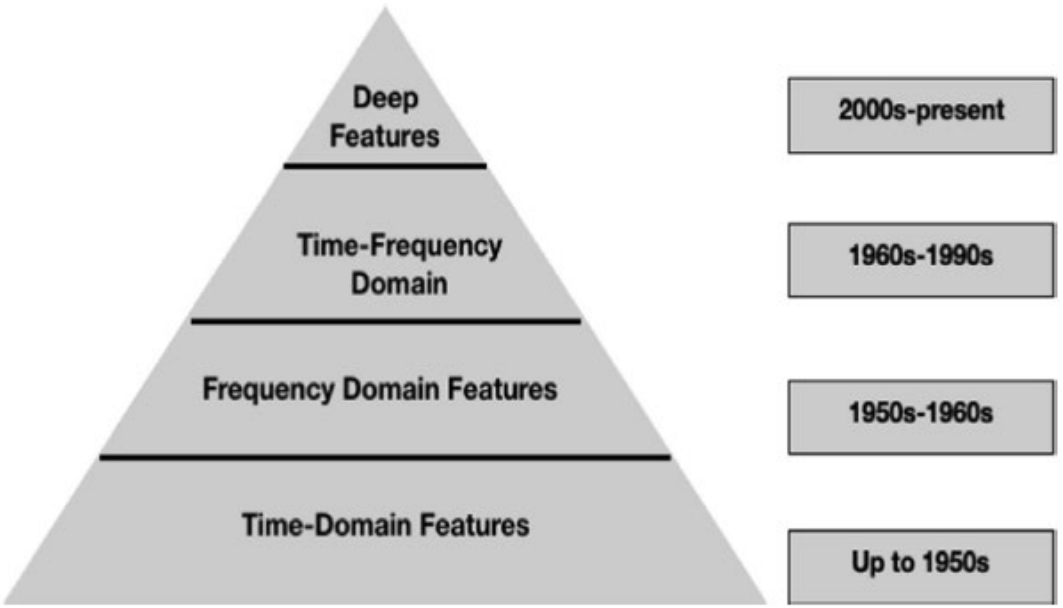


Fig. 6. Evolution of Audio Feature Extraction.

Time domain Features
-Zero crossing rate, Modified ZCR
-Amplitude based features: AD, ADSR, log attack time, shimmer
-Energy based features: Short time energy, Volume, Temporal Centroid
-Auto-correlation based features
- Rhythm based features

Cepstral Domain Features
- Mel Frequency Cepstral Coefficients (MFCCs)
- Linear Prediction Cepstral Coefficients
- Perceptual linear prediction (PLP) cepstral coefficients
- RASTA-PLP
- GFCCs
- GTCCs

Image based features
- Local Binary Patterns
- Local Ternary Patterns
- Histogram of gradients (HOG) feature
- Scale invariant feature transform (SIFT)

Frequency Domain Features
-Auto-regression based features: LPC, CELP, LSF
- Peak Frequency
- MSAF-MULTIEXPANDED
- SMOFS-MULTIEXPANDED
- STFT/Time frequency based: time-frequency matrix, sub-band energy ratio, spectrum envelope, SPSF, GDF
- EMS based
- LTAS based
- Chroma related features
- Tonality based features: FF, pitch histogram, pitch profile, harmonicity, harmonic-to-noise ratio, jitter
- Spectrum shape based features: Spectral Centroid, Spectral centre, spectral roll-off, spectral spread, skewness, kurtosis, Spectral slope, spectral decrease, bandwidth, spectral flatness, spectral crest factor, entropy, flux, OBSC.

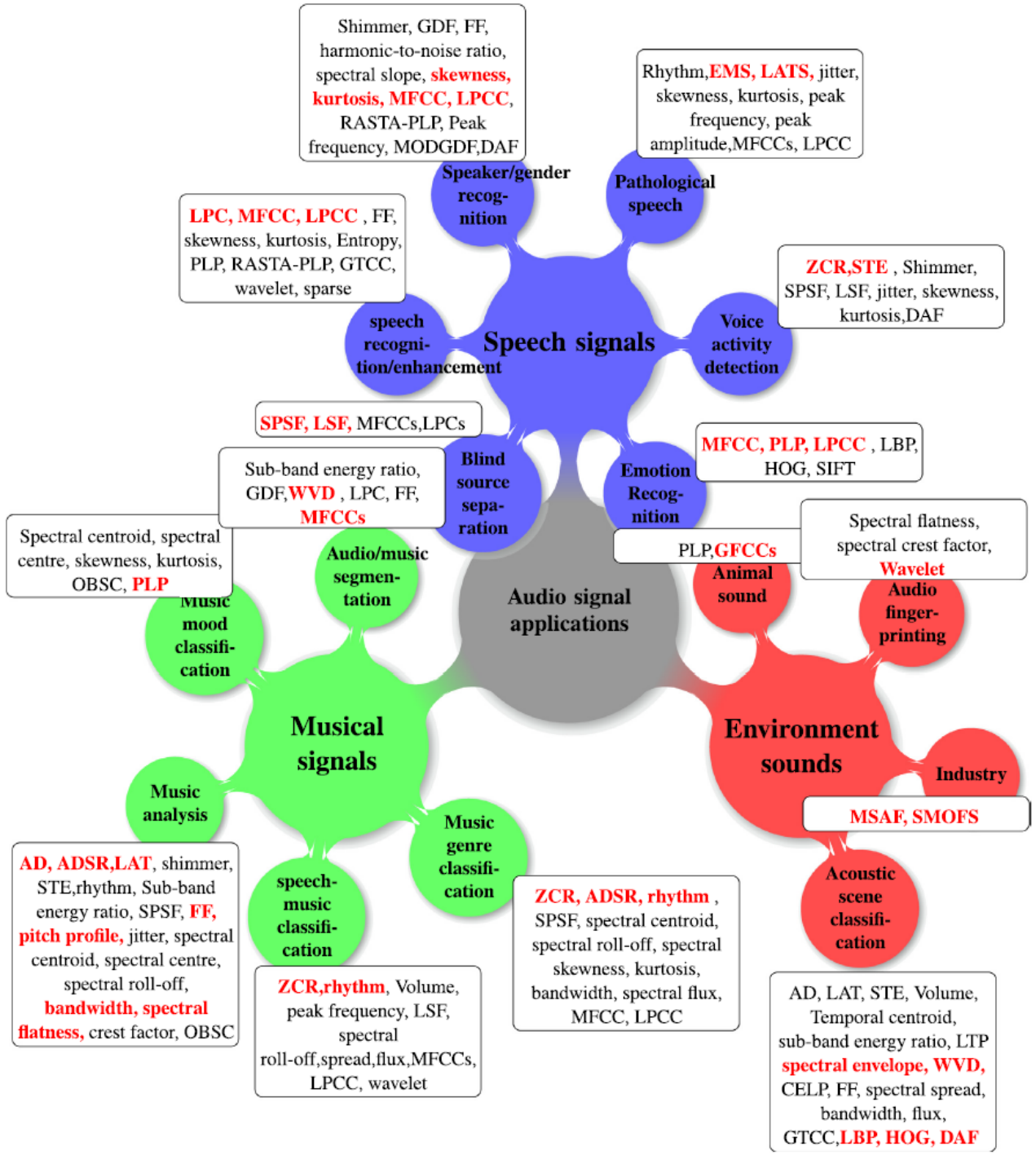
Deep Features
- CNN features
- deep neural networks (DNNs)
- recurrent neural networks (RNNs)
- Deep stacked auto-encoder (SAE)
- unidirectional long short term memory network (LSTM)
- bi-directional long short term memory (BLSTM)

Sparsity based features
- Matching Pursuit (MP)
- Orthogonal matching pursuit
- Stage wise greedy method
- Basic pursuit
- Coordinate descent etc.

Wavelet based features
- DWT based features
- CWT based features
- Wavelet transform based features
- Wavelet packet decomposition based

And some features are better for specific tasks

- For example, MFCC is typically very useful for speech signals and speech recognition/enhancement etc.
- Simpler features lige ZCR is very useful for voice activity detection and music genre classification etc.



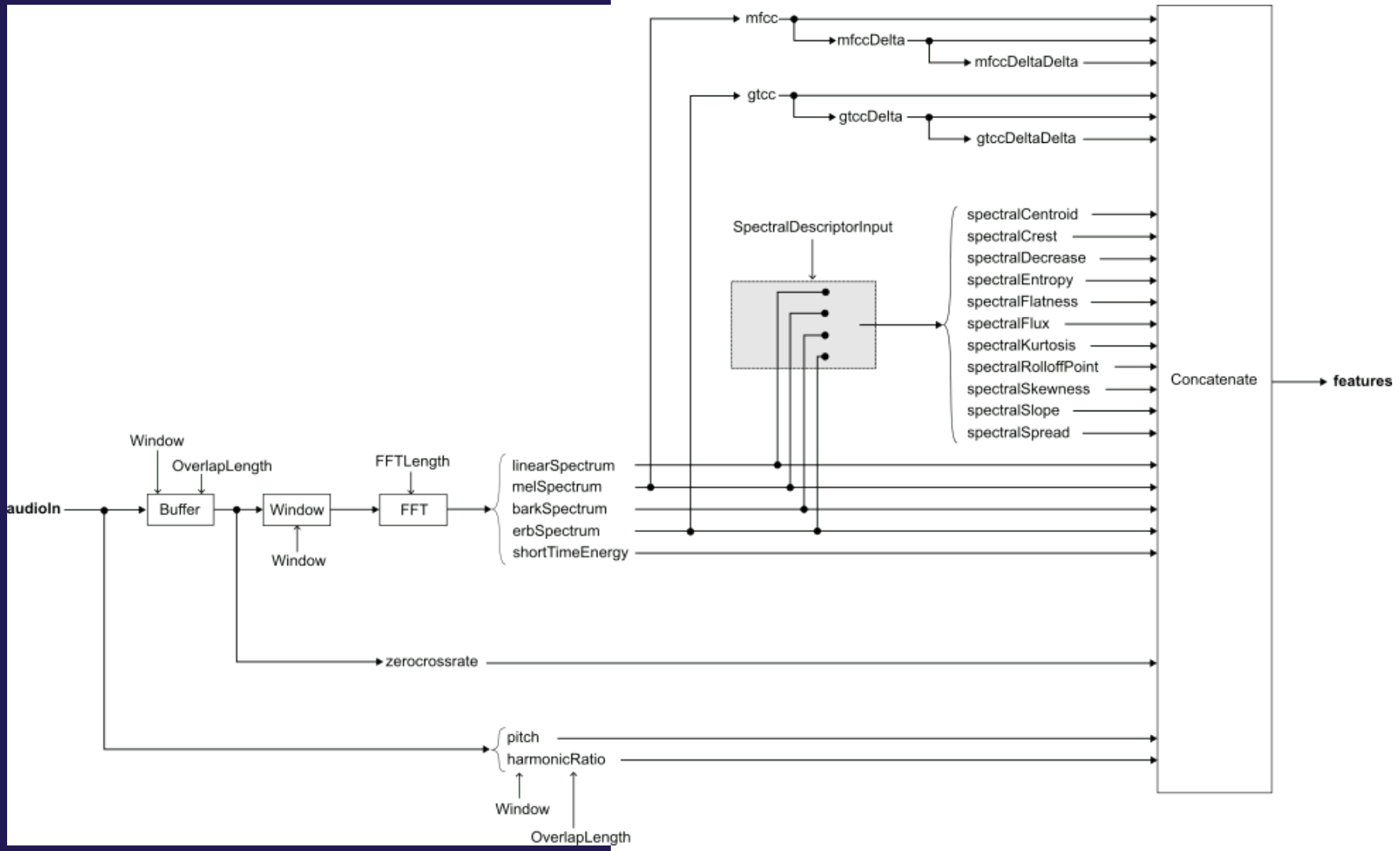
Common features for AI for audio signals

Other features exist!

- Focus on Matlab function [audioFeatureExtractor]

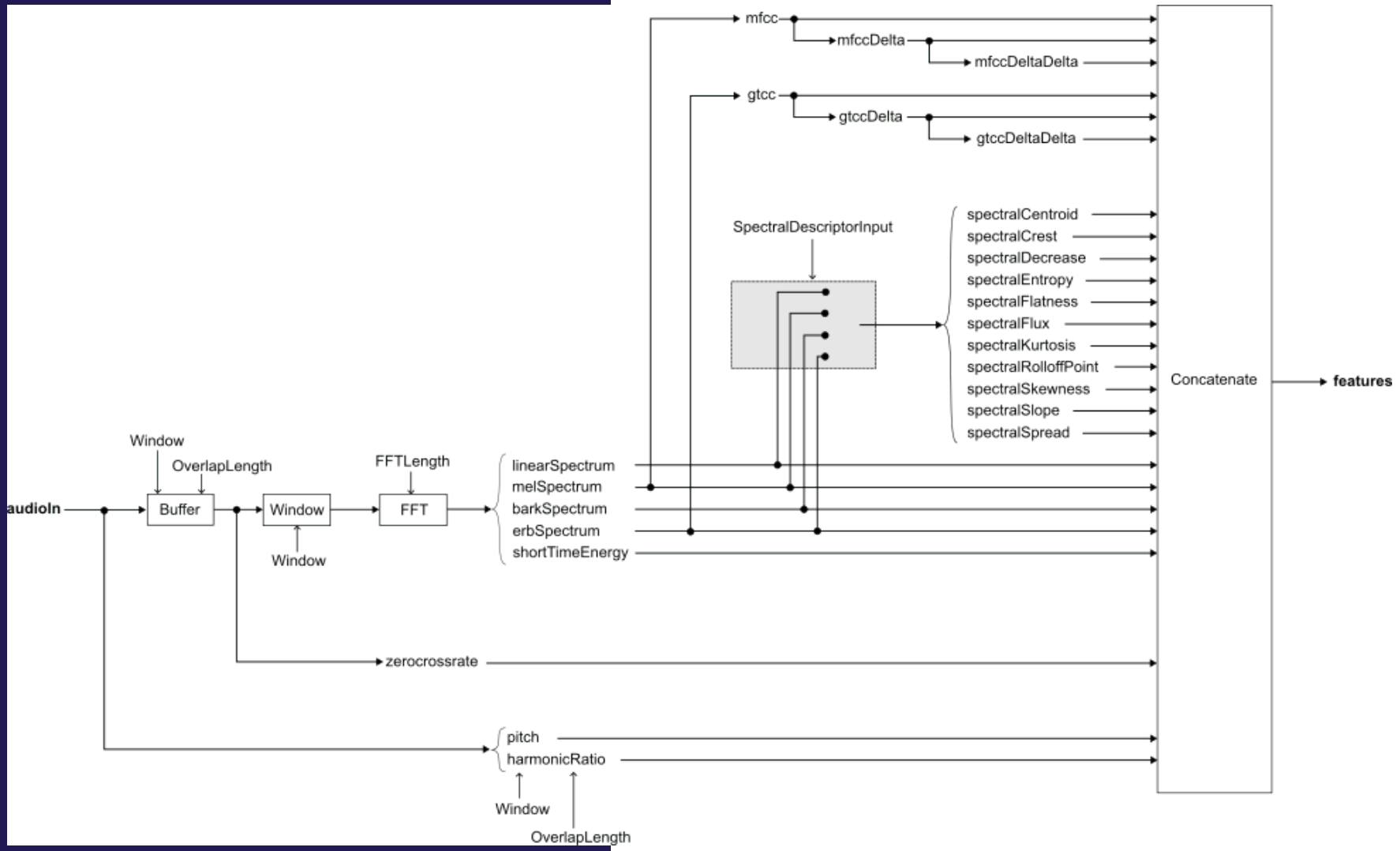
  and/or the livescript tool [https://se.mathworks.com/help/audio/ref/extractaudiofeatures.html]

- **However, not all the features presented can be found in the function – but Matlab code will be provided!**



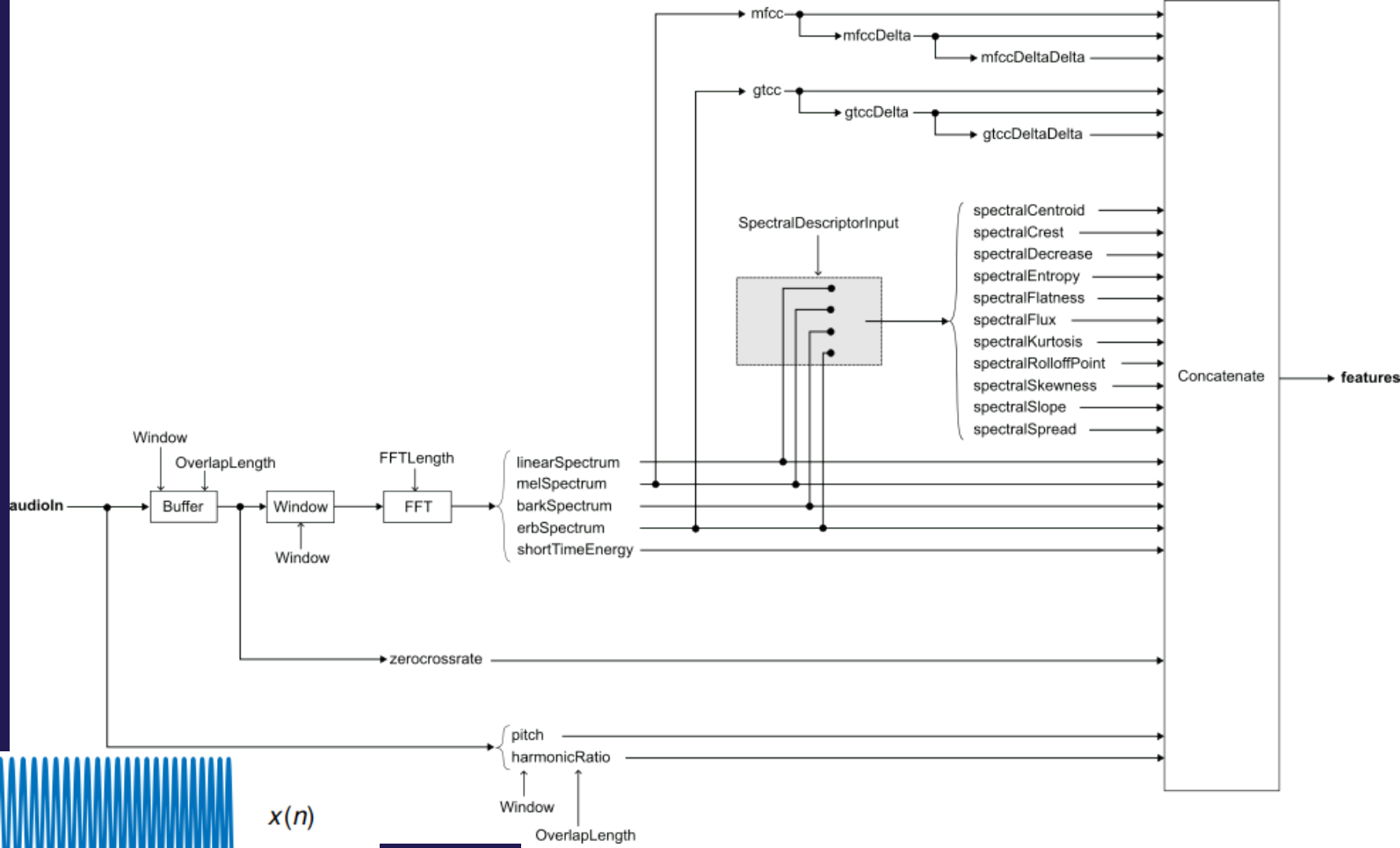
Categories of features in audioFeatureExtractor

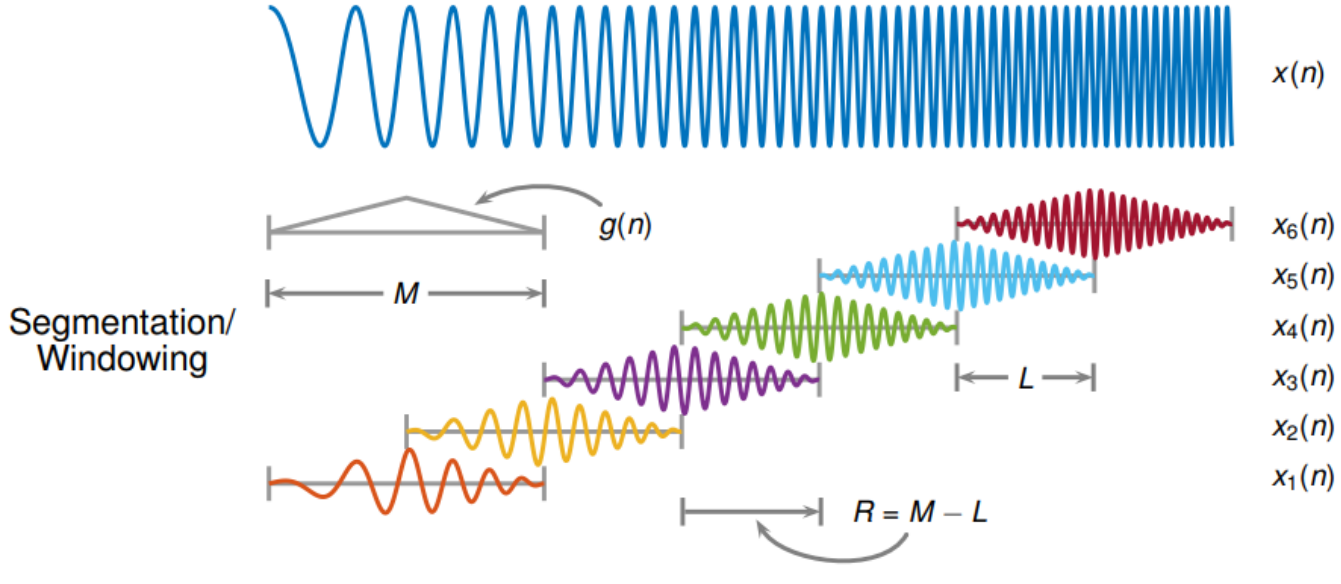
- Time domain
- Frequency domain
- Time-frequency domain
- Cepstral domain



Time window

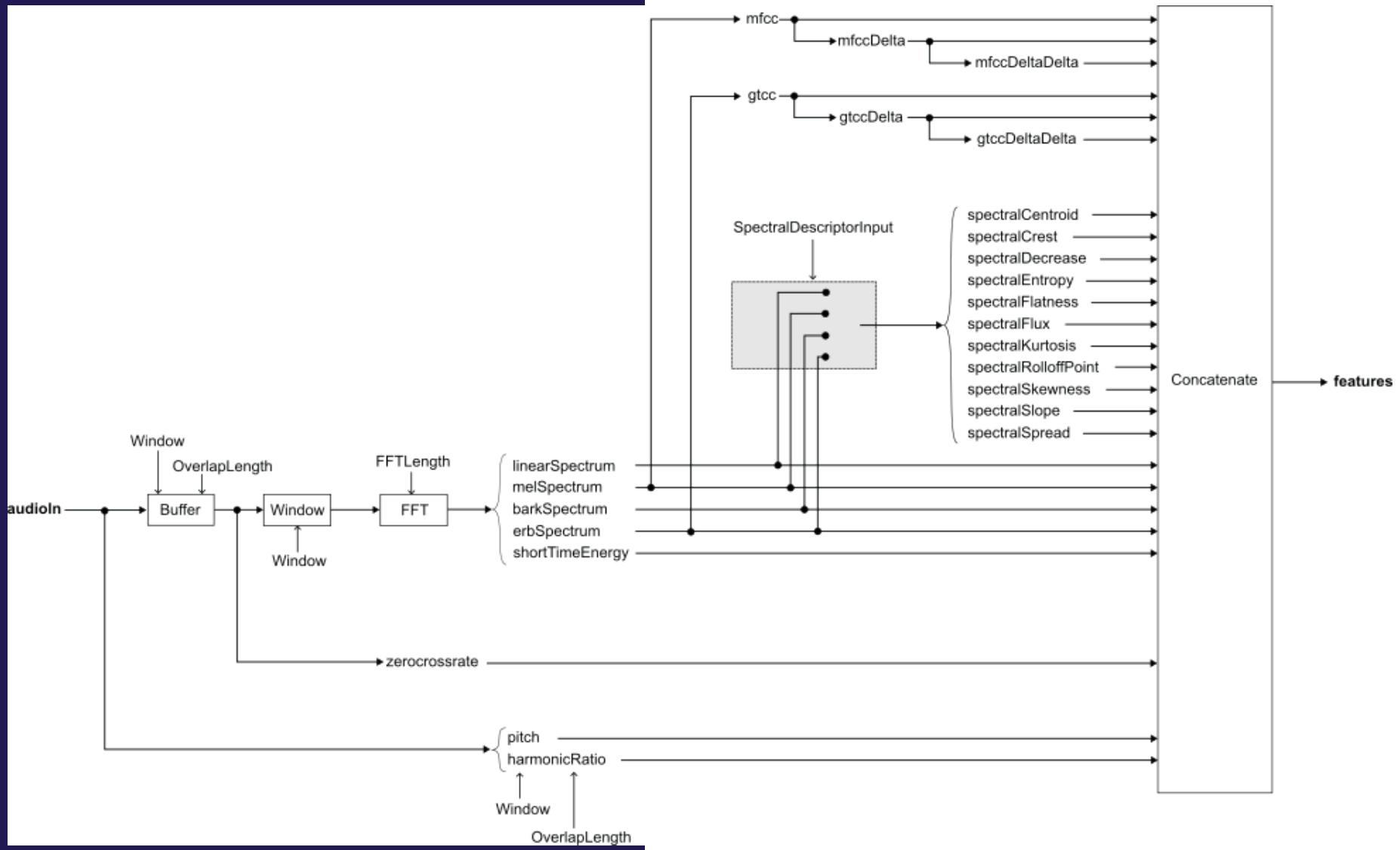
- In order to account for varying sound we use time windows to extract frames
- Features are often calculated for each frame
- Different window types and length can influence the analysis (remember topic 1)





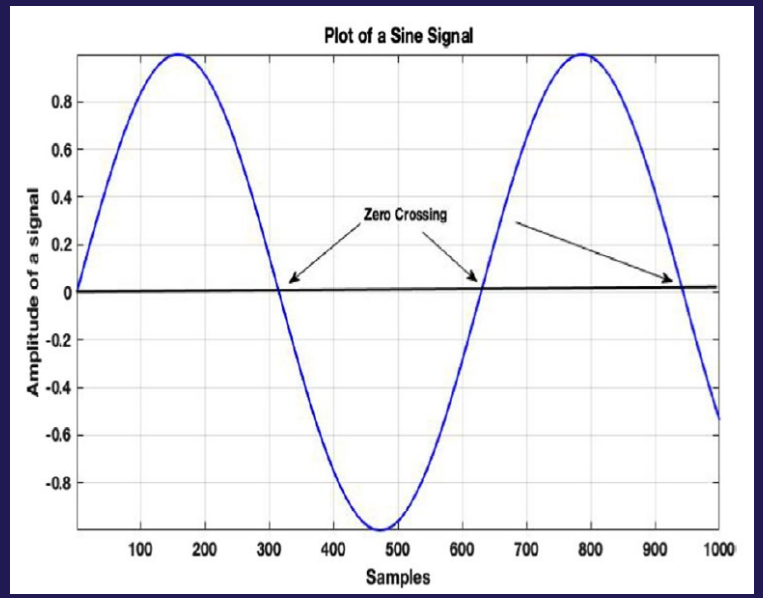
Time domain features

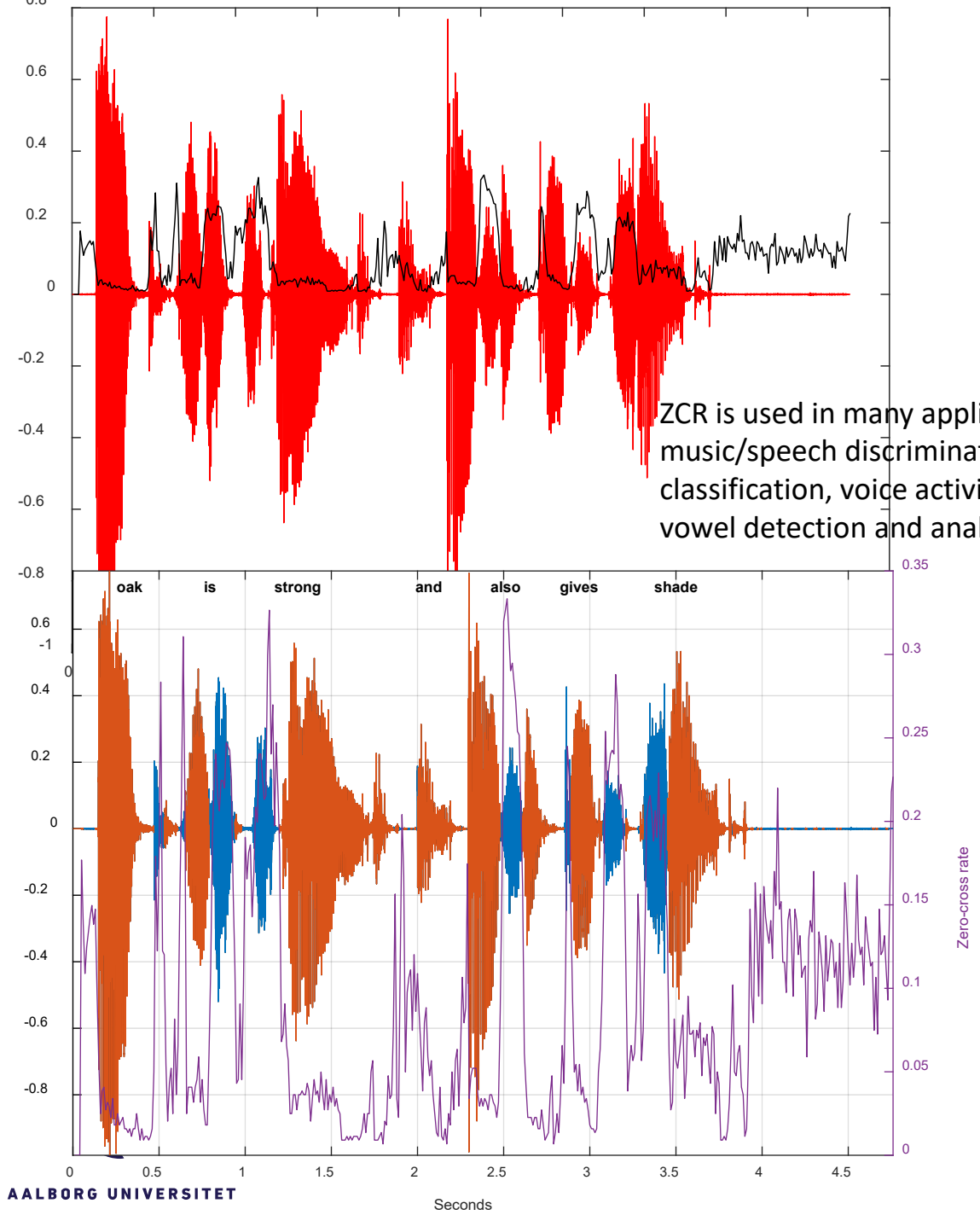
- Zero-crossing rate
- Amplitude/Energy, RMS
- Loudness
- Short-time energy
- Envelope
- Auto-correlation, Cross-correlation
- AR models and LPC



Zero-crossing rate

- How many times the signal cross zero during frame or per second (divided by frame length)
- Example to separate voiced (vowels → low zero crossing rate) and unvoiced (consonants → high zero-crossing rate) from each other
- Can also be used to estimate the fundamental frequency.

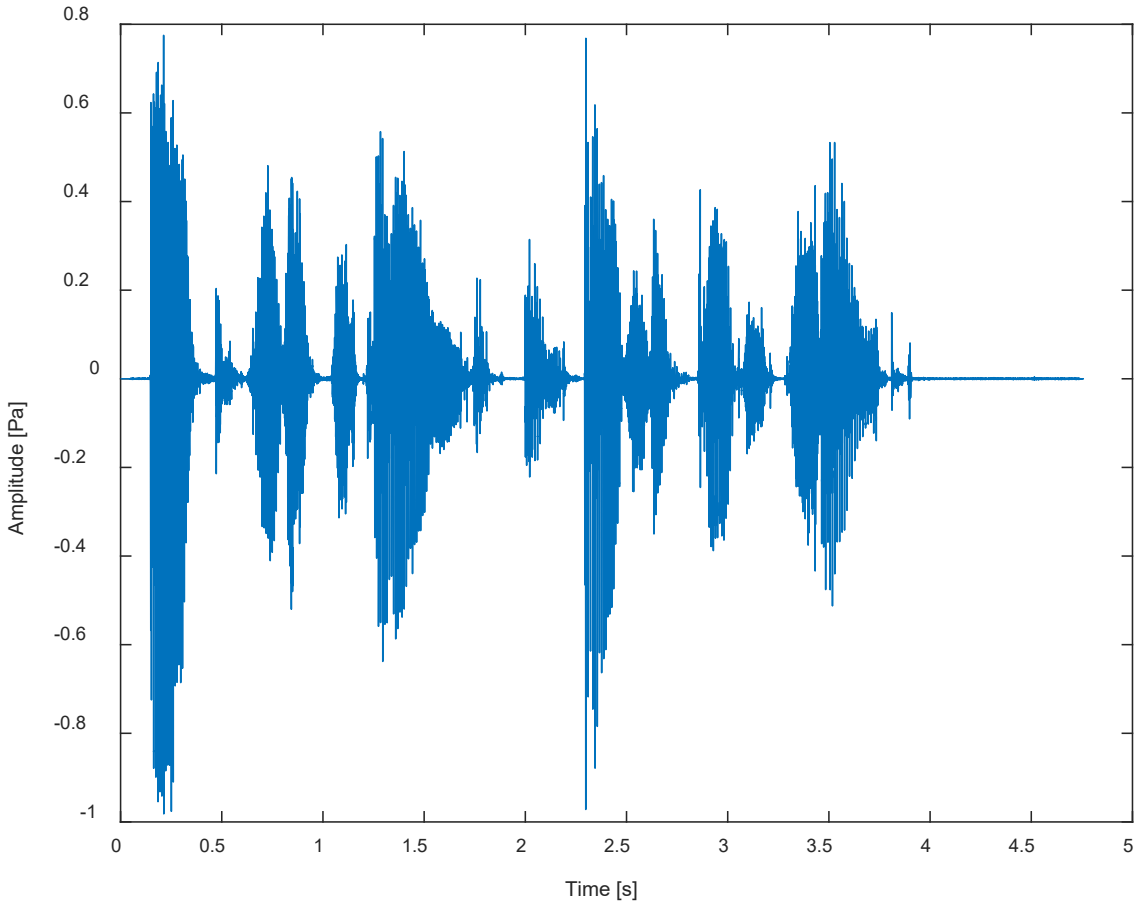




ZCR is used in many applications including music/speech discrimination, music genre classification, voice activity detection and vowel detection and analysis

Amplitude, energy, sound pressure level

- Amplitude in pascal or digital levels
- Root mean square (RMS):
  use rms() in Matlab
- Sound pressure level SPL:
  20*log10(p/20e-6) dB, where p is the signal RMS in pressure



RMS = 0.0939 Pascal

SPL = 73.4 dB

Assuming digital values correspond to Pascal

AALBORG UNIVERSITET

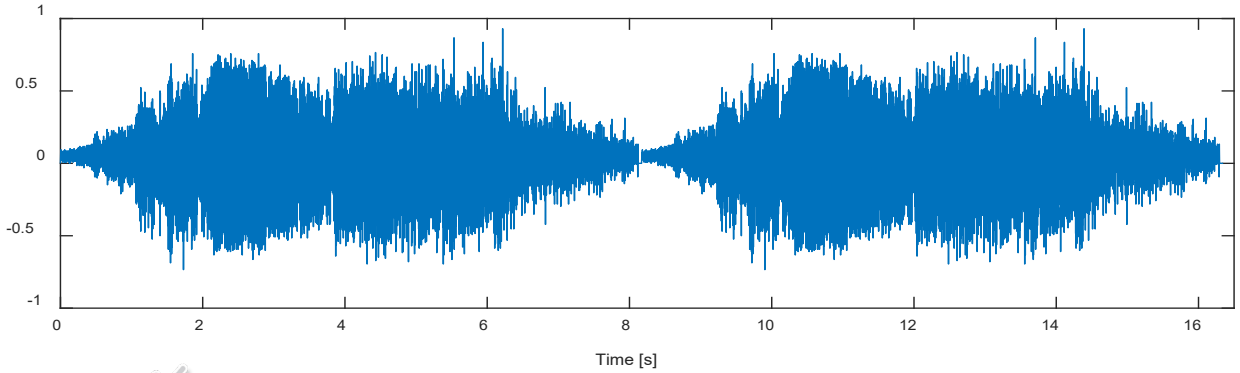
Loudness

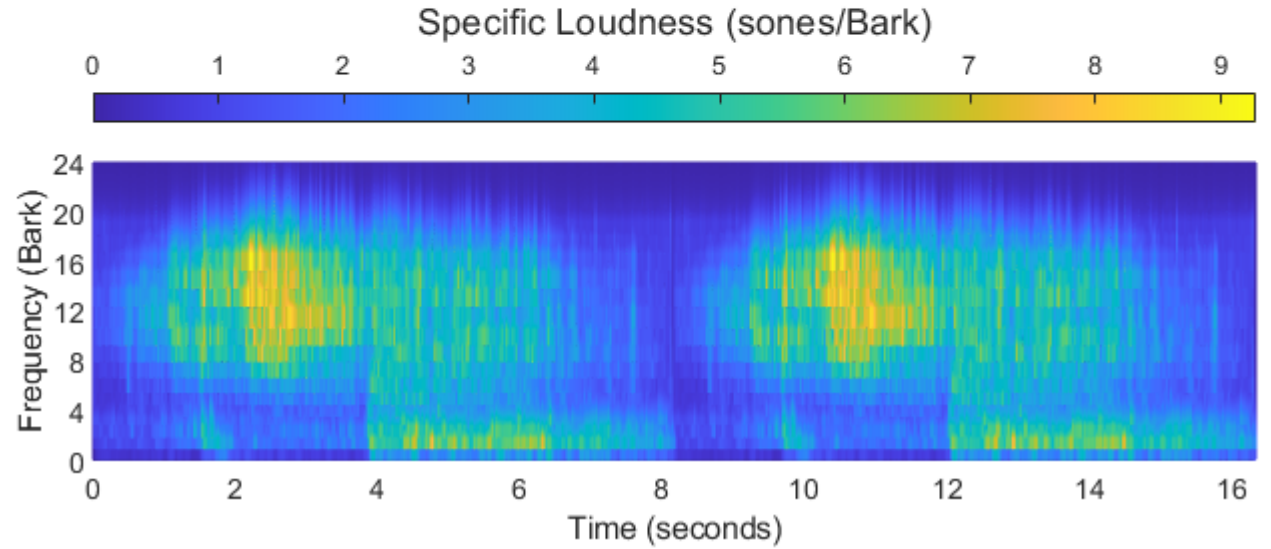
- Loudness describes our perception and can also be estimated closer to human hearing – example in Matlab:

```
[audioIn,fs] = audioread('JetAirplane-16-
11p025-mono-16secs.wav');
t=1/fs:1/fs:length(audioIn)/fs;
plot(t,audioIn)
xlabel('Time [s]')

figure
acousticLoudness(audioIn,fs,'SoundField','dif
fuse','TimeVarying',true)
```

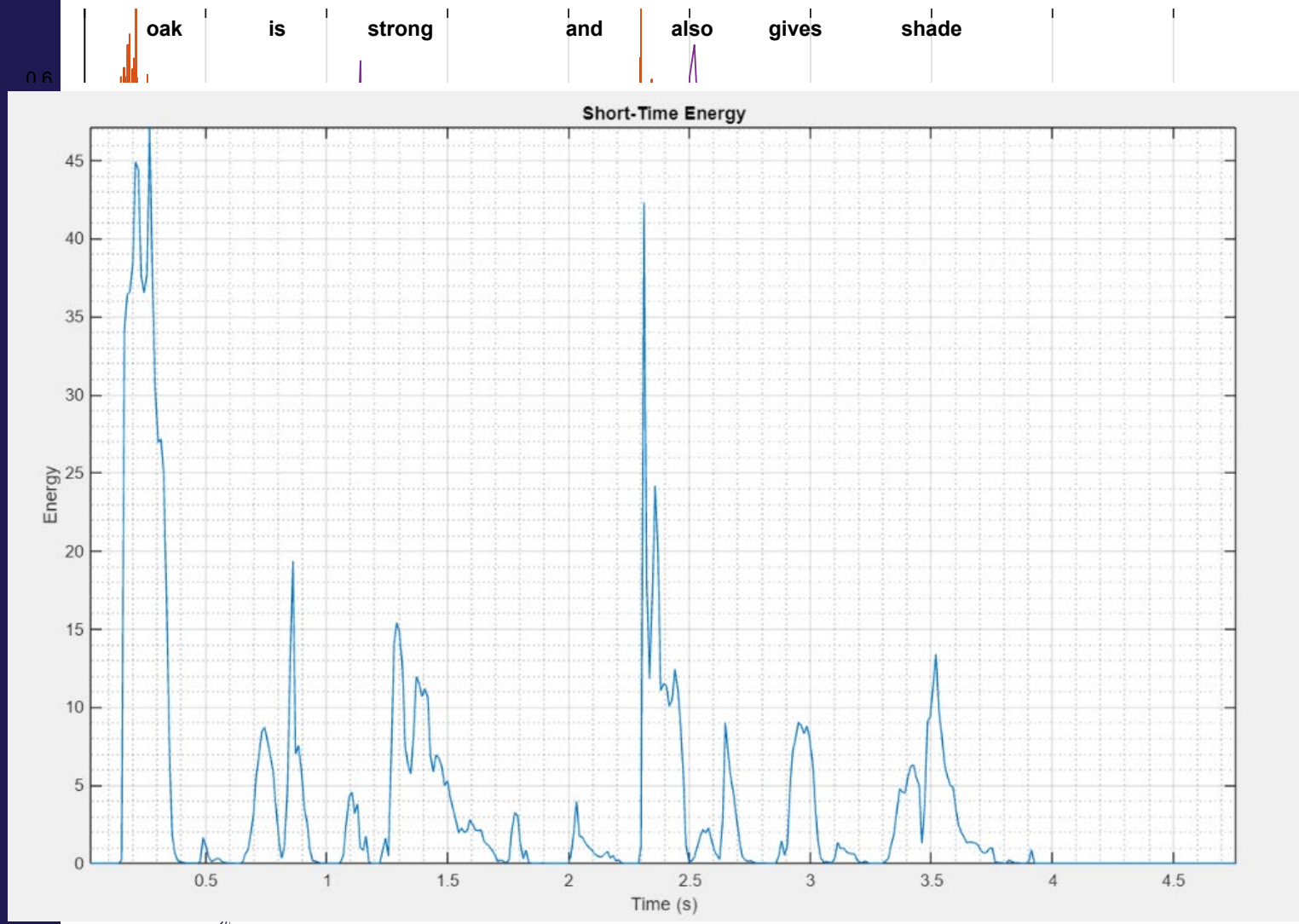
- It is used in speech/music discrimination, speech segmentation and acoustic scene classification.





Short-time energy

- Energy of signal in window:
  sTE = sum(xbw.^2,1)
  , where xbw is the frame (windowed signal)
- Example: The STE is low for unvoiced segments and high for voiced segments.
- Other examples of use are:
  music onset detection, environmental sound detection, vowel detection and analysis and audio based surveillance systems.

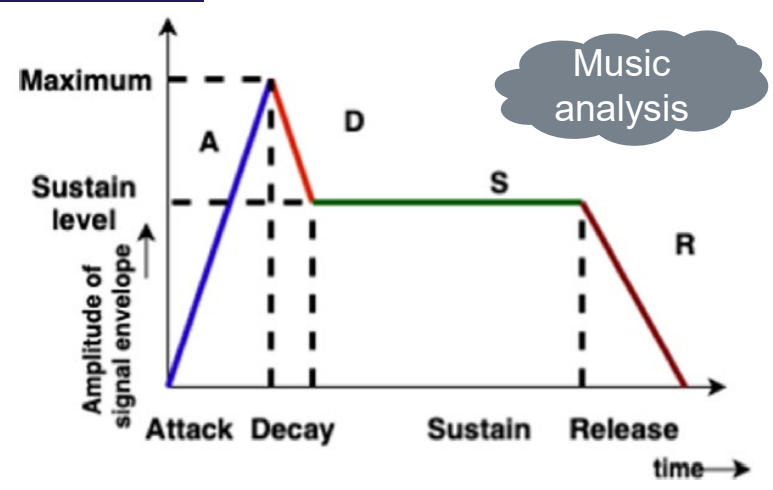


Envelope

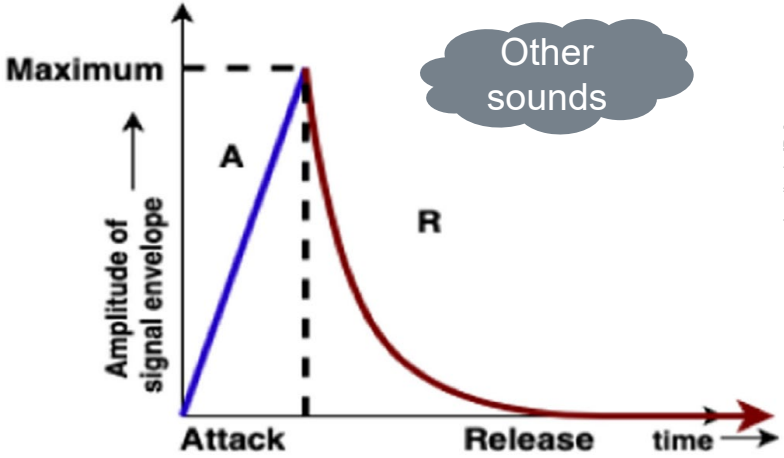
- Shape of amplitude:
  use envelope() Matlab function
- How does amplitude change with time

Algorithm 2: ADSR envelope detection

1. **Result**: ADSR envelope
2. Input: Musical Key number, duration of key pressing
3. $freq = 440 * 2((keynum - 49)/12)$ ▷calculate frequency for given key
4. $t = 0 : 1/sampling\ frequency : duration$
5. $tone = sin(2 * pi * freq * t)$ ▷generate sinusoidal output tone
6. $A = linspace(0, 1, 0, 1 * (length(tone))$ ▷rise 10 percent of signal
7. $D = linspace(1, 0.8, 0.15 * (length(tone))$ ▷drop of 15 percent of signal
8. $S = linspace(0.8, 0.8, 0.6 * (length(tone))$ ▷delay of 60 percent of signal
9. $R = linspace(0.8, 0, 0.15 * (length(tone))$ ▷drop of 15 percent of signal
10. $ADSR = [ADSR]$
11. Multiply ADSR with tone and plot on MATLAB

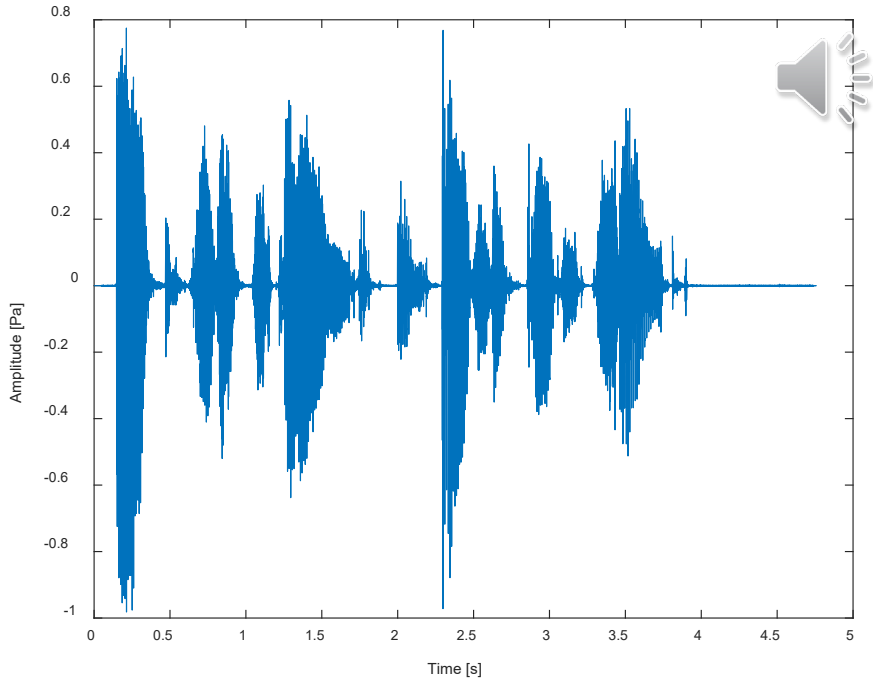


(a) ADSR envelope

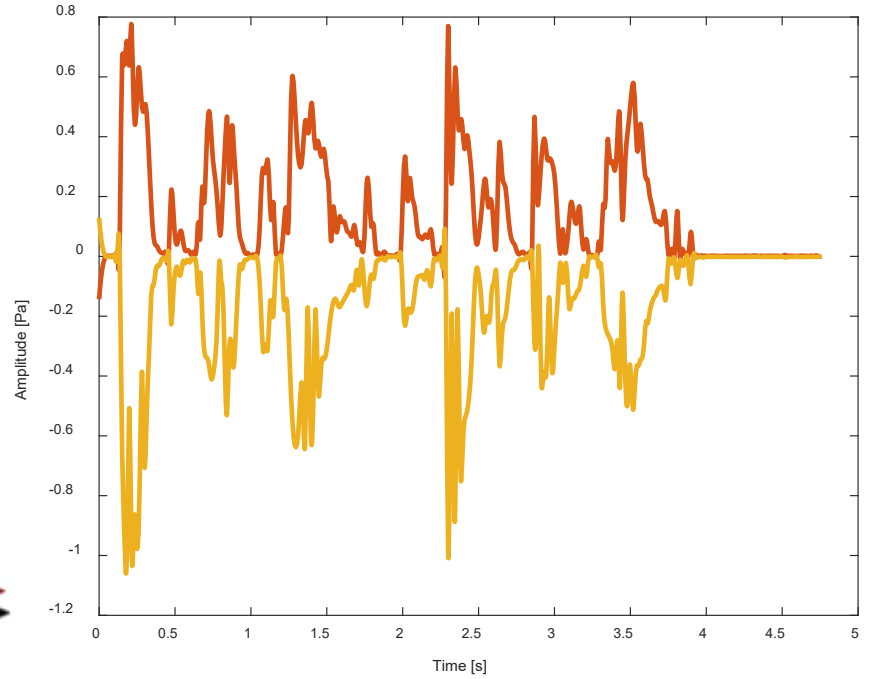


(b) AR envelope

Music analysis

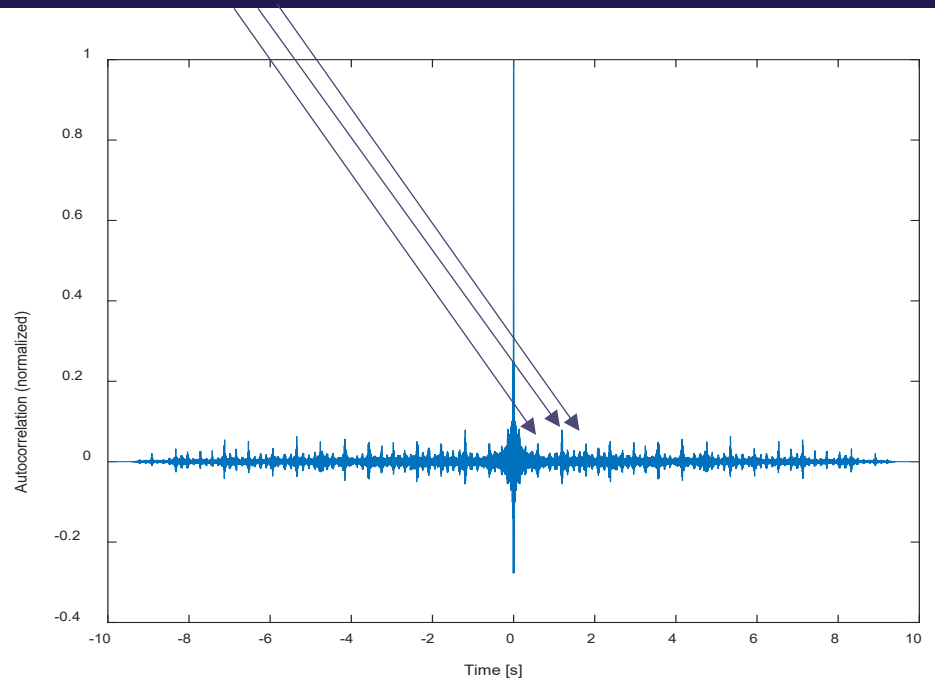


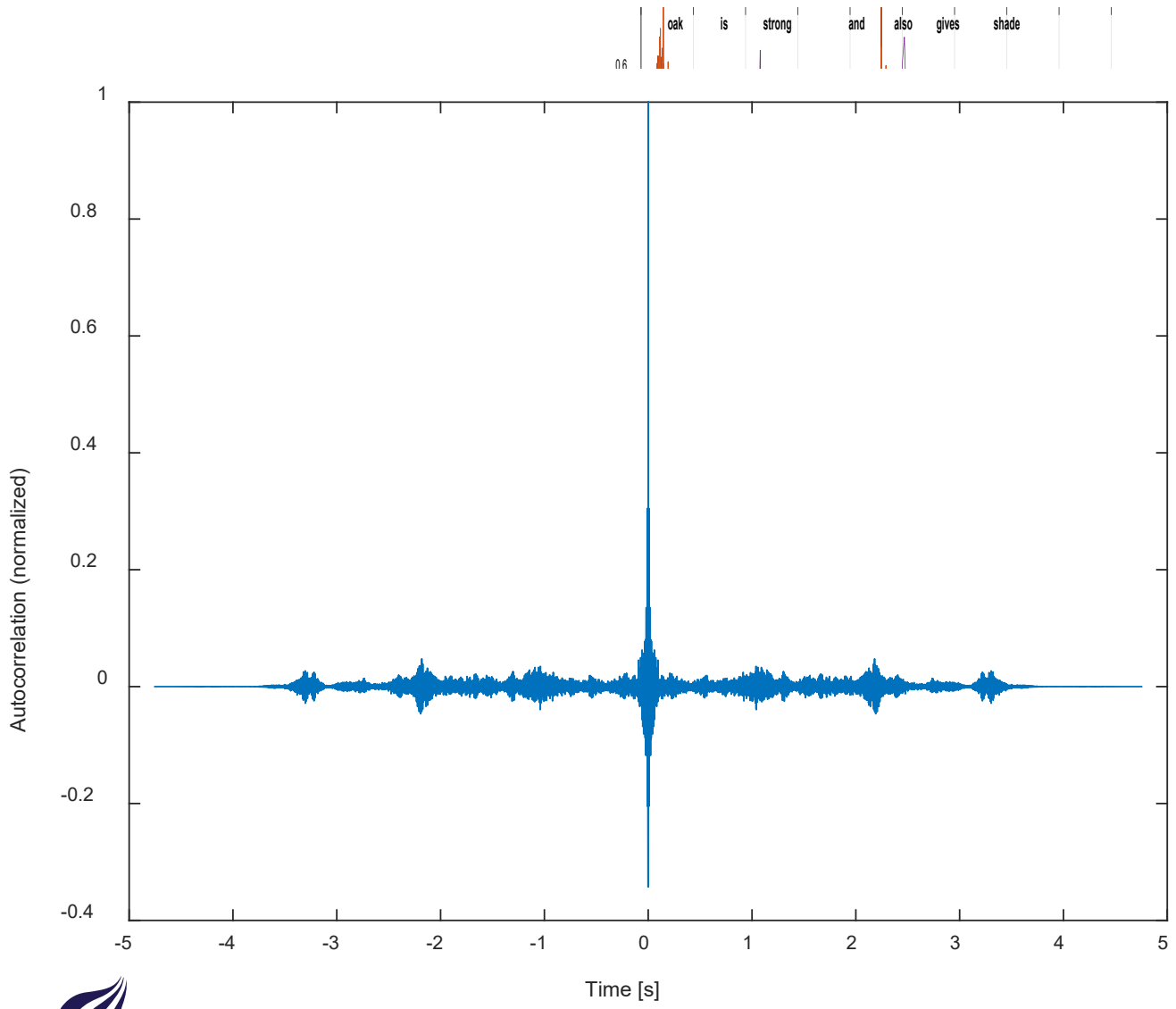
Other sounds



Autocorrelation

- How similar is the signal at different *lags* (delayed versions of itself)?
- Use xcorr() Matlab function:
  [c,lags] = xcorr(x);
- Detect echoes or repetitions in signal
- Find beats in music signals





AALBORG UNIVERSITET

BREAK

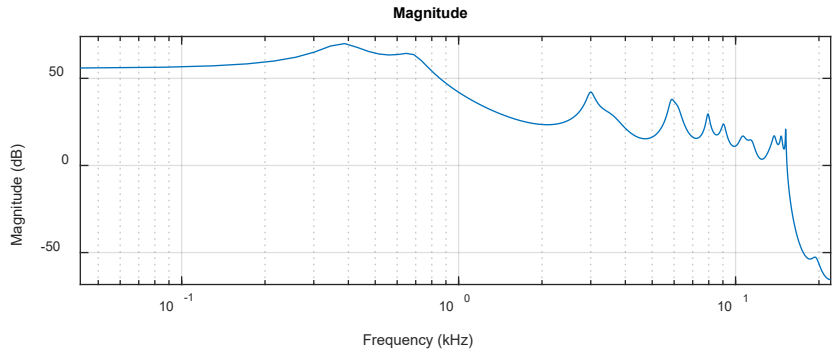
Auto regressive model and LPC

- Auto regressive model can be used to create a linear filter that predicts next values of a signal based on previous signal values
- It creates an all-pole filter that represent the spectral envelope of the signal (often used for speech signals)
- Different AR algorithms exist:
- One that ensures a stable filter is the Burg method:
  ARCoeffs = arburg();
- Another uses the autocorrelation:
  ARCoeffs = LPC();
- LPC coefficients are use for audio segmentation and audio retrieval. It can also be used to “repair” errors or missing segments in an audio stream

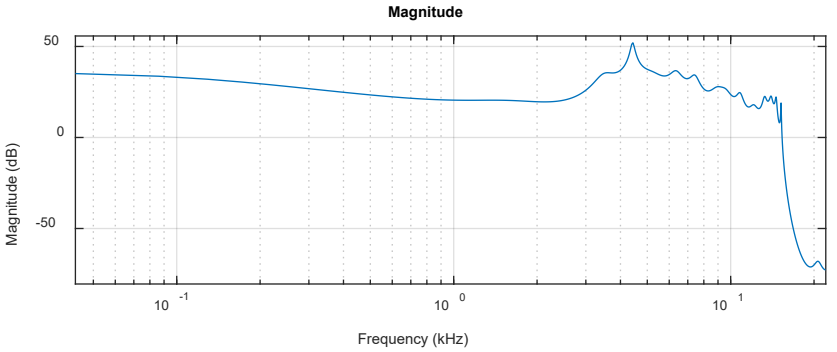
AALBORG UNIVERSITY
DENMARK

AR model order 32 example

“OAK”



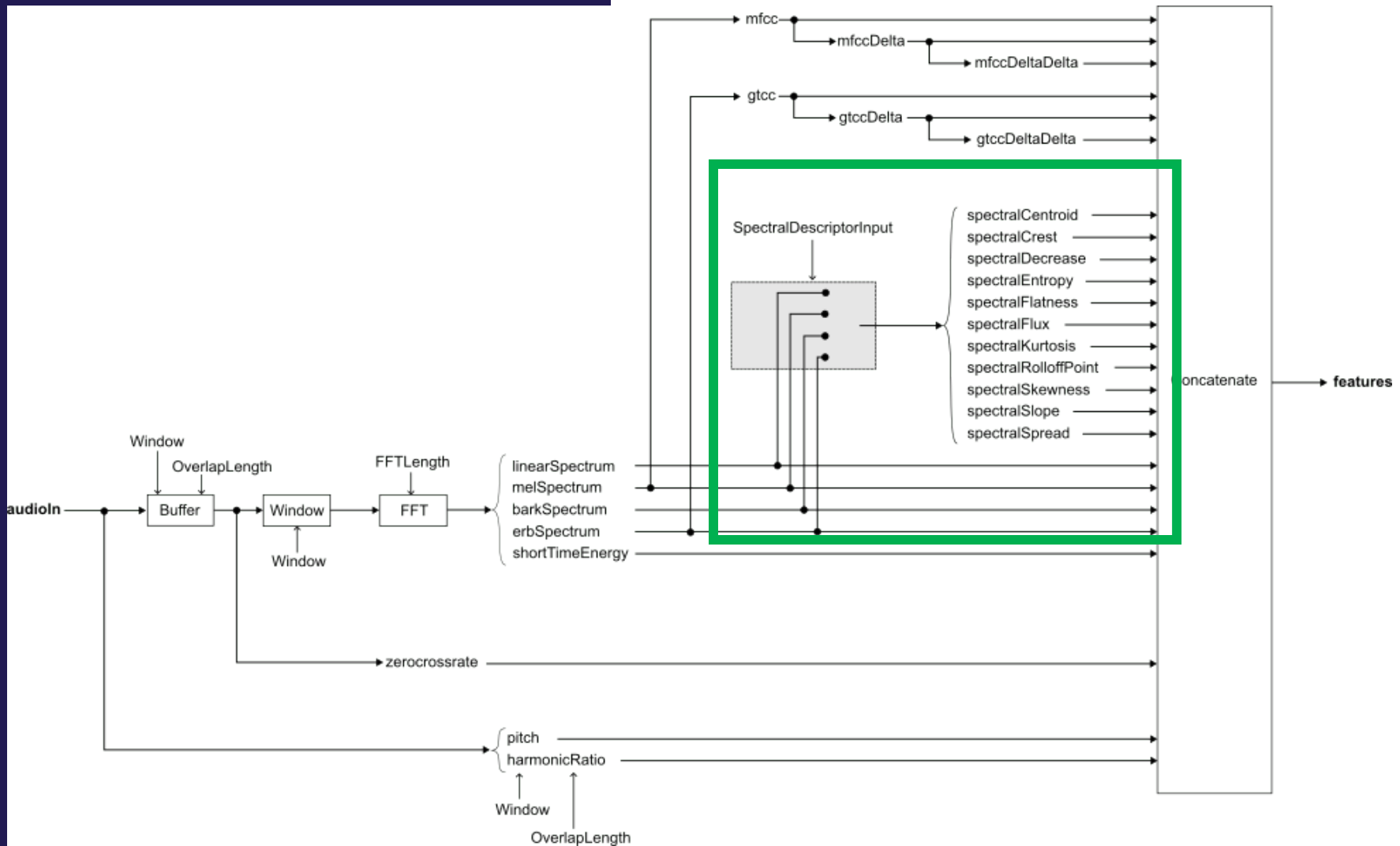
“..ss”



https://se.mathworks.com/help/releases/R2024b/audio/ug/spectral-descriptors.html

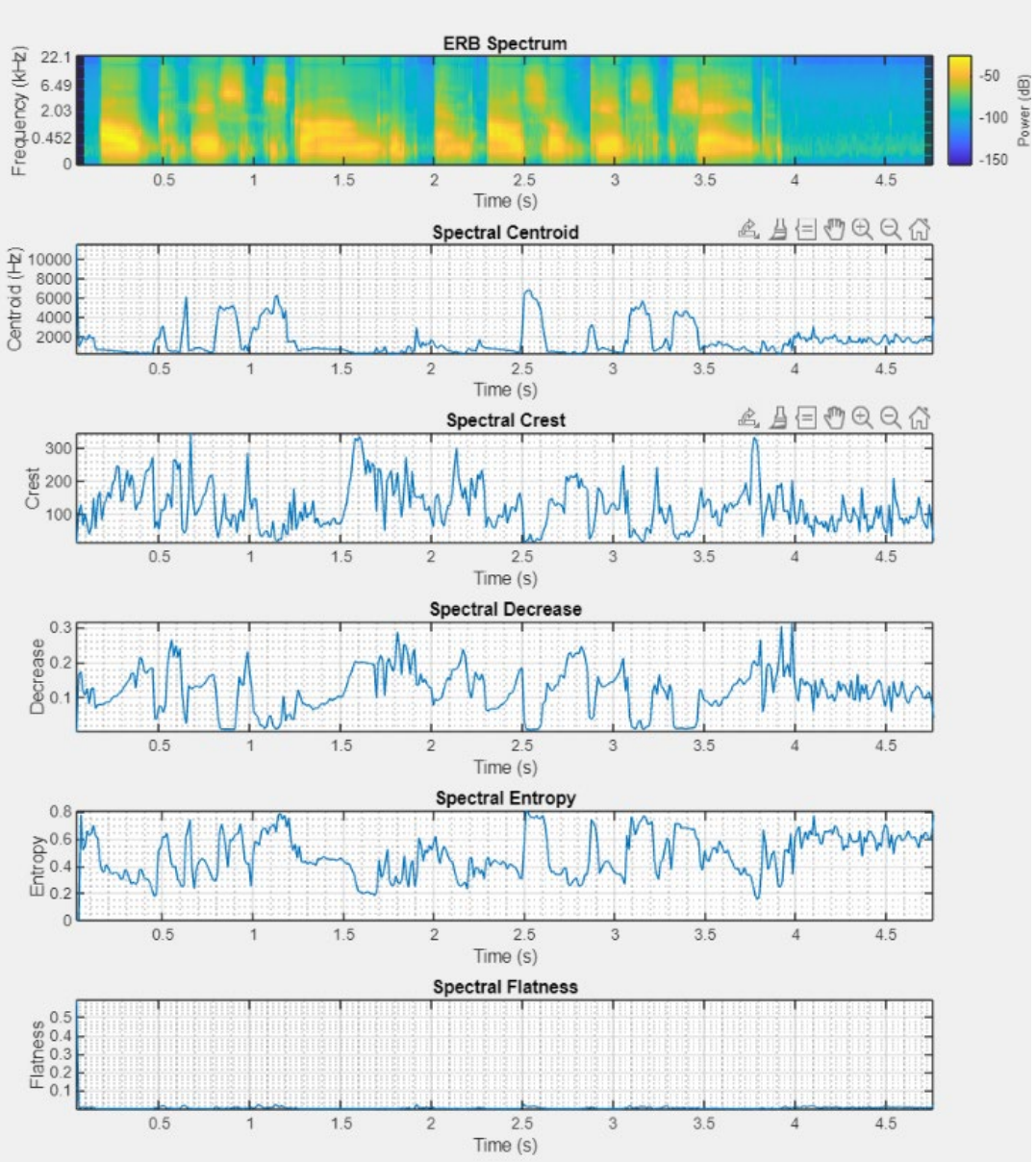
Frequency domain features (based STFT)

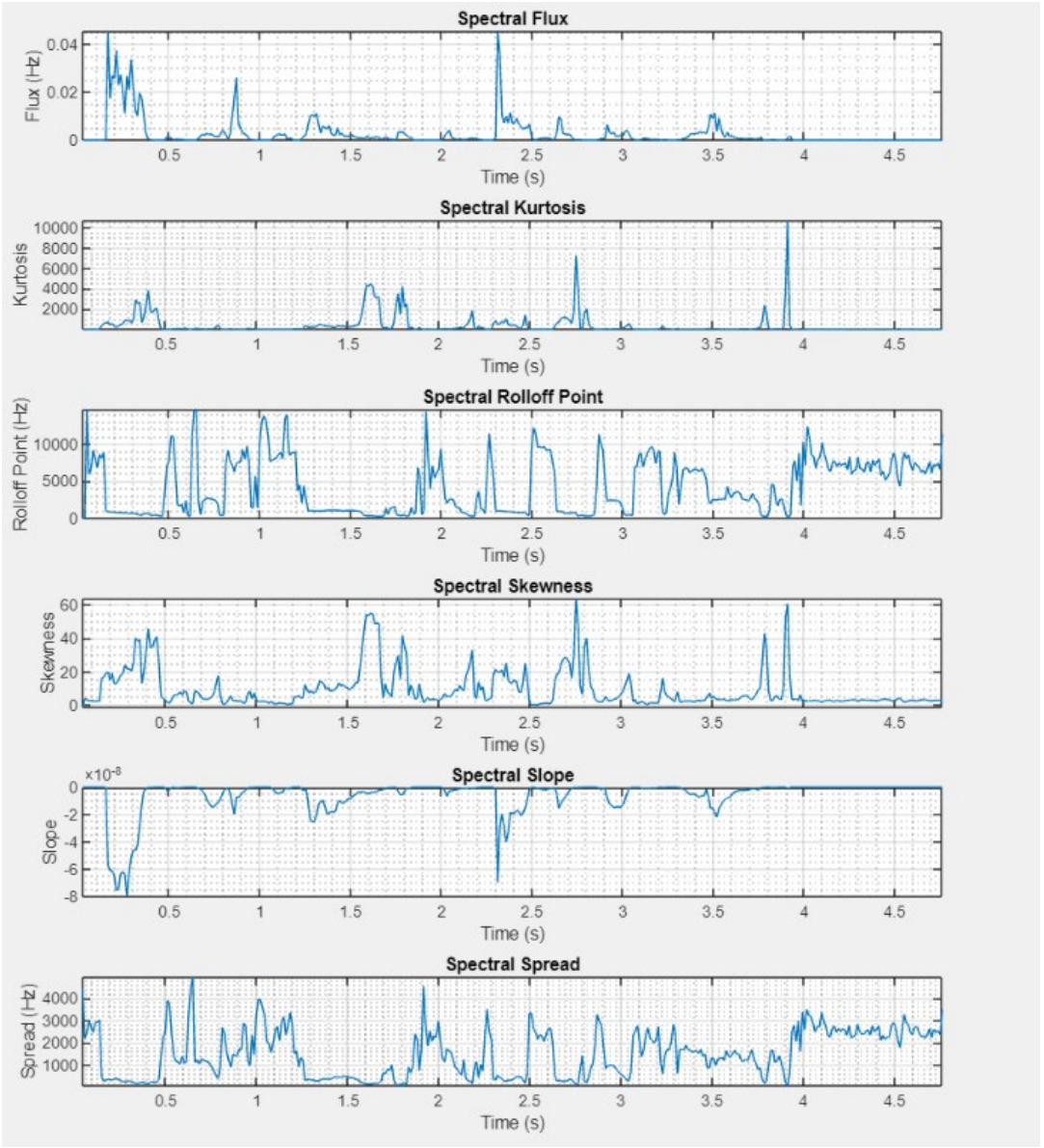
- Spectral centroid
- Spectral crest
- Spectral decrease
- Spectral entropy
- Spectral flatness
- Spectral flux
- Spectral kurtosis
- Spectral roll-off point
- Spectral skewness
- Spectral slope
- Spectral spread



https://se.mathworks.com/help/releases/R2024b/audio/ug/spectral-descriptors.html

Frequency domain features





AALBORG UNIVERSITY
DENMARK

https://se.mathworks.com/help/releases/R2024b/audio/ug/spectral-descriptors.html

Frequency domain features

- Spectral kurtosis: describes the flatness of the spectrum around its mean value. For gaussian distribution the spectral kurtosis has value 0, if kurtosis is less than 0, we observe flat distribution and if spectral kurtosis is greater than 0, we observe sharp peaked spectral. Just like spectral skewness, spectral Kurtosis is also used as a feature in music genre classification and mood classification, fault detection in bearing of electric motors and Parkinson's disease detection from speech.
- Spectral skewness: Spectral skewness measure the symmetry of the spectrum around its arithmetic mean value. This feature would be equal to zero for silent segments and high for voiced parts. Skewness equal to zero describes symmetric distribution, skewness less than zero indicates more energy to the right side of spectral distribution and skewness greater than zero describes more energy components are present on the left side of the spectrum. Used in mood detection and music genre classification, fault detection in motor bearings and Parkinson's disease detection from speech

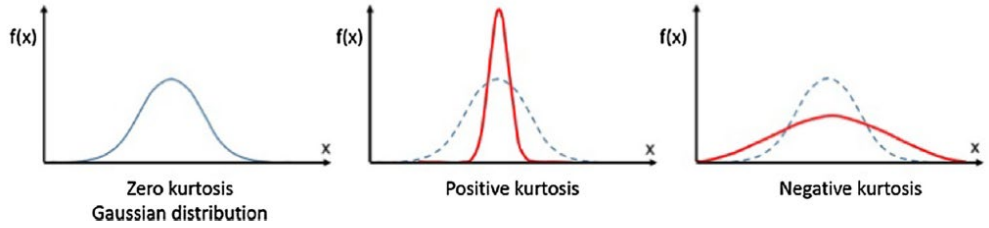


Fig. 16. Kurtosis for different type of spectrum.

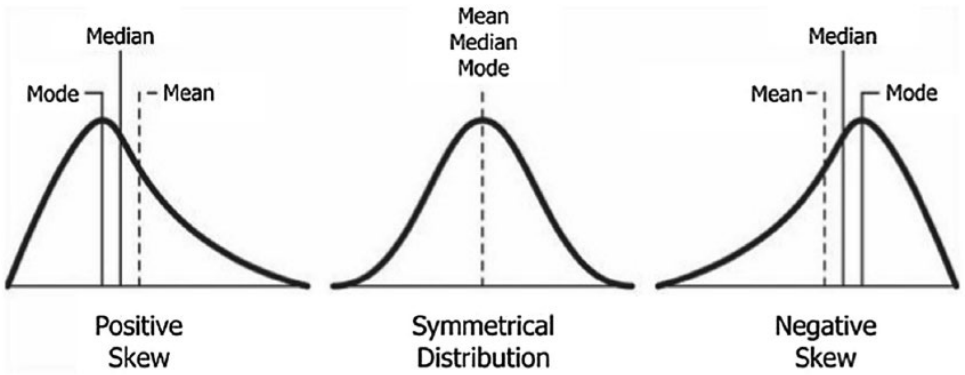


Fig. 15. Skewness for different type of spectrum.

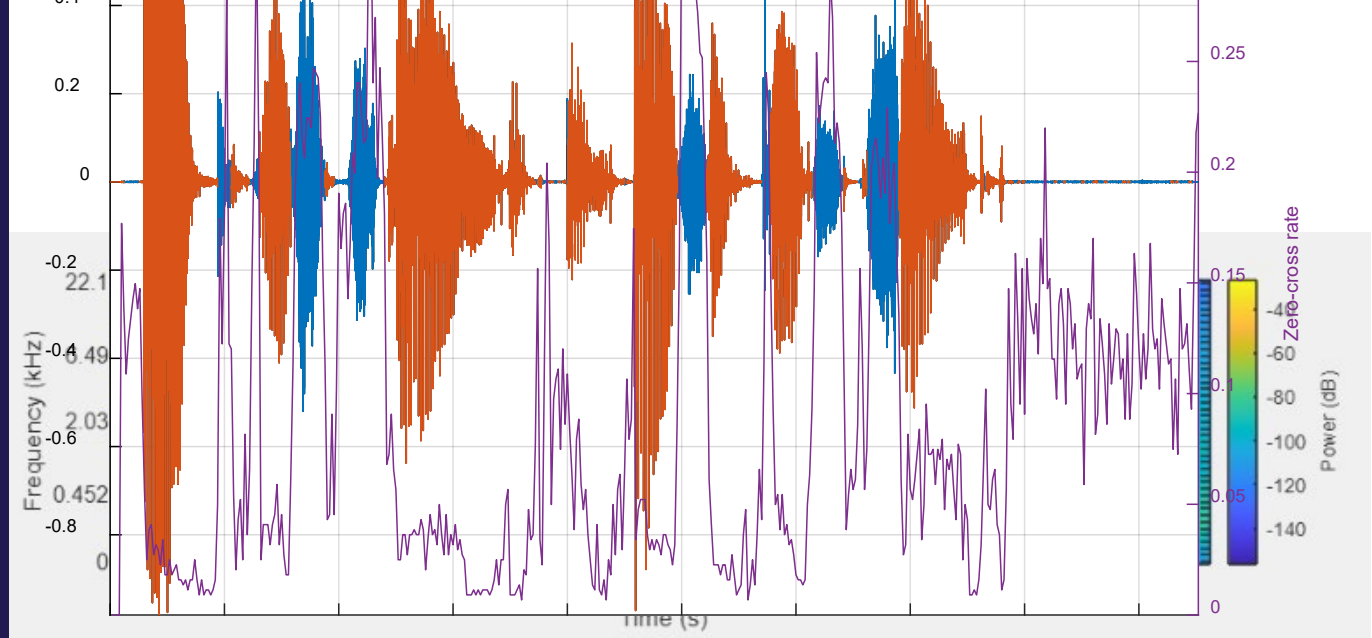
https://se.mathworks.com/help/releases/R2024b/audio/ug/spectral-descriptors.html

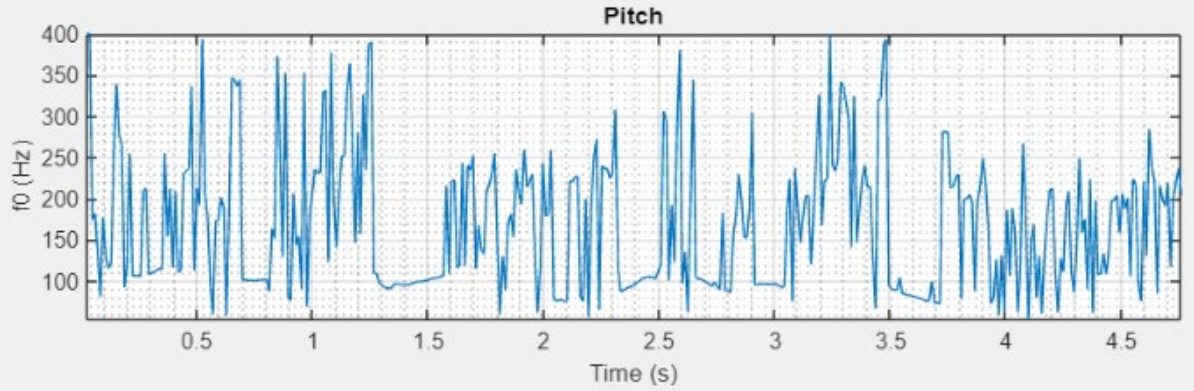
Frequency domain features

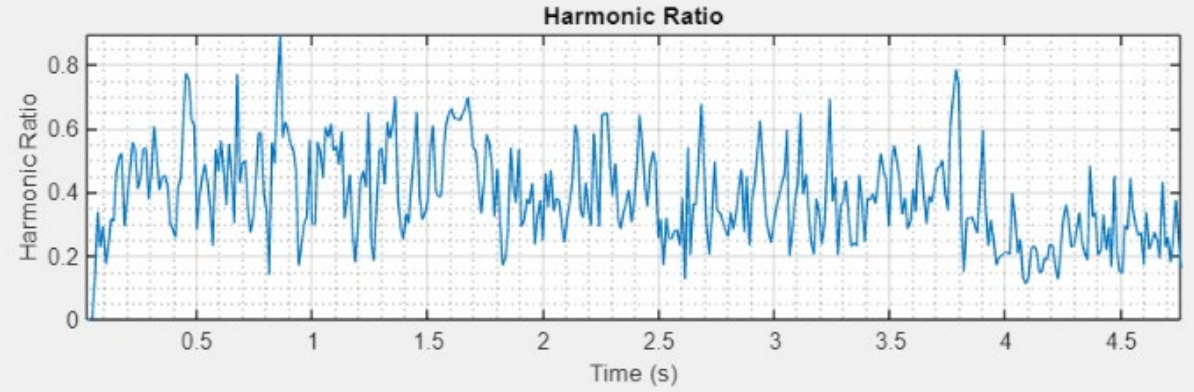
- **Spectral centroid:** indicates where the center of mass of the spectrum is located. It is an excellent measure of brightness of sound signal and used to measure timbre of music, music classification and music mood classification.
- **Spectral spread:** also called Spectral Dispersion is closely related to the bandwidth of the signal. It can be described as average deviation of the rate-map around its centroid. Noise like signals have wide spectral spread than the pure tonal sounds which has small spectral spread. It is generally, wide for music and environmental sounds and comparatively narrow for speech like sounds
- **Spectral flatness:** is the measure of the uniformity in the frequency distribution. It is calculated as the ratio of the geometric mean to the arithmetic mean. It can be used to distinguish between harmonic and noise like sounds. For harmonic sounds the spectral flatness is close to zero and for noise like sounds the value of spectral flatness is close to one. It is employed in music onset detection, music classification and audio fingerprinting.
- **Spectral crest:** In contrast to the spectral flatness, spectral crest factor determines how peaked is the power spectrum of the sound signal. It is also used to distinguish between harmonic/tonal sounds and noise like sounds. It is higher for harmonic/tonal sounds and lower for noise like sounds. This feature is also used for audio fingerprinting and music classification.
- **Spectral entropy:** measures the peakiness of the spectrum. It has been used successfully in voiced/unvoiced decisions for automatic speech recognition. Because entropy is a measure of disorder, regions of voiced speech have lower entropy compared to regions of unvoiced speech. Has also been used to discriminate between speech and music
- **Spectral flux:** is defined as 2-norm of the frame-to-frame spectral amplitude difference vector. It points the sudden changes in the frequency energy distribution of sounds. This feature is used in speech/music discrimination, music genre classification and environmental sound classification.
- **Spectral roll-off point:** is the frequency where 95% of the signal energy is contained below this frequency. Used in speech/music classification, music genre classification, musical instrument classification and audio-based surveillance systems
- **Spectral slope:** is the measure of slope of the amplitude of the signal and it is computed by linear regression. This feature is used in speech analysis and in identifying speaker from a speech signal.
- **Spectral decrease:** explains the average spectral-slope of the rate-map representation, putting a strong emphasis on low frequencies. Is commonly used, along with slope in the analysis of music. In particular, it has been shown to perform well as a feature in instrument recognition.

Pitch & harmonic ratio

- **Pitch: fundamental frequency**
  Used in music onset detection, environmental sound, recognition and audio retrieval.
  Music key detection and genre classification etc.
- **Harmonic ratio:**
  The harmonic ratio indicates the ratio of energy in the harmonic portion of audio to the total energy of the audio
- Used in e.g. in analyzing pathological voices and music-related applications

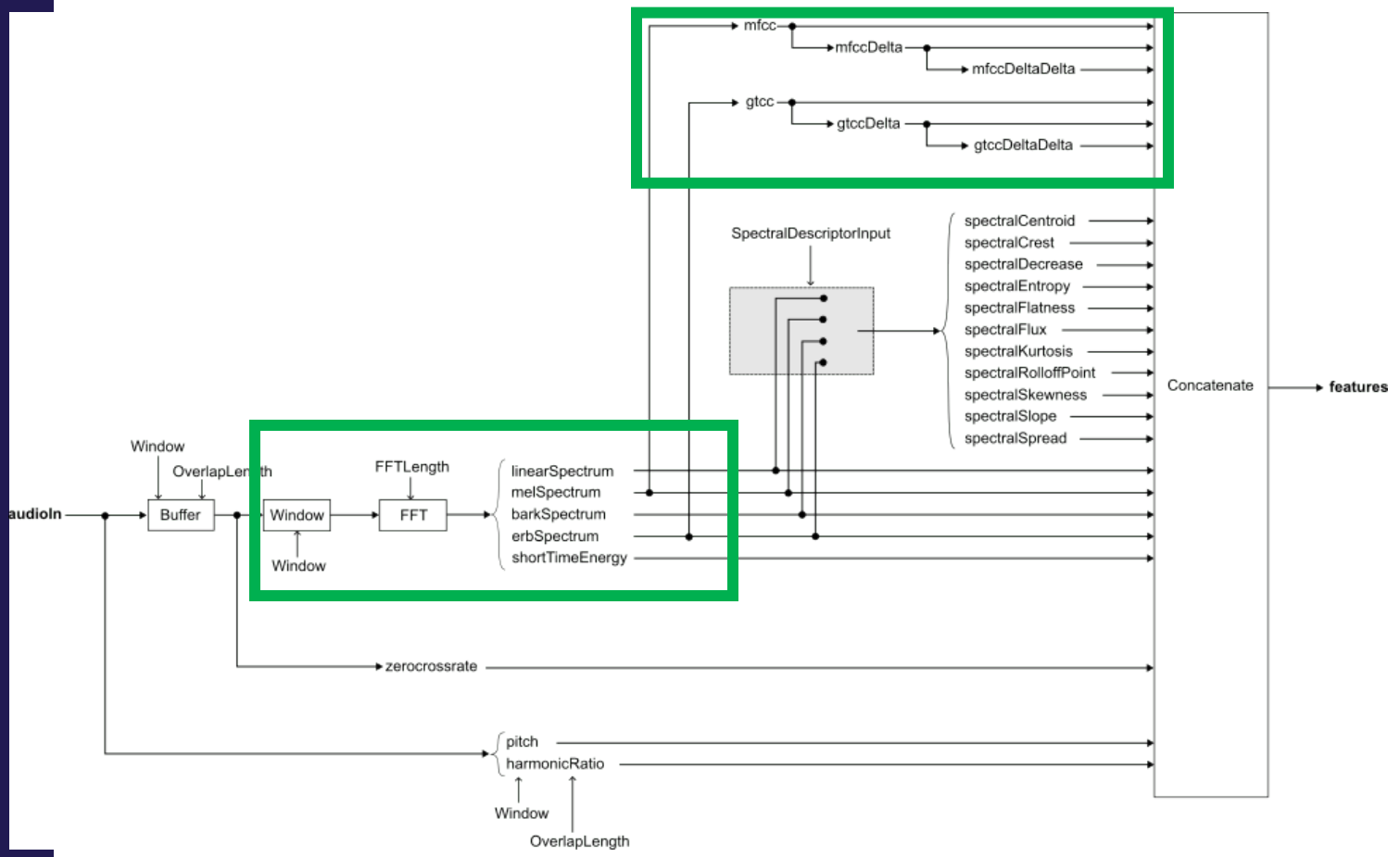






Time-frequency domain features

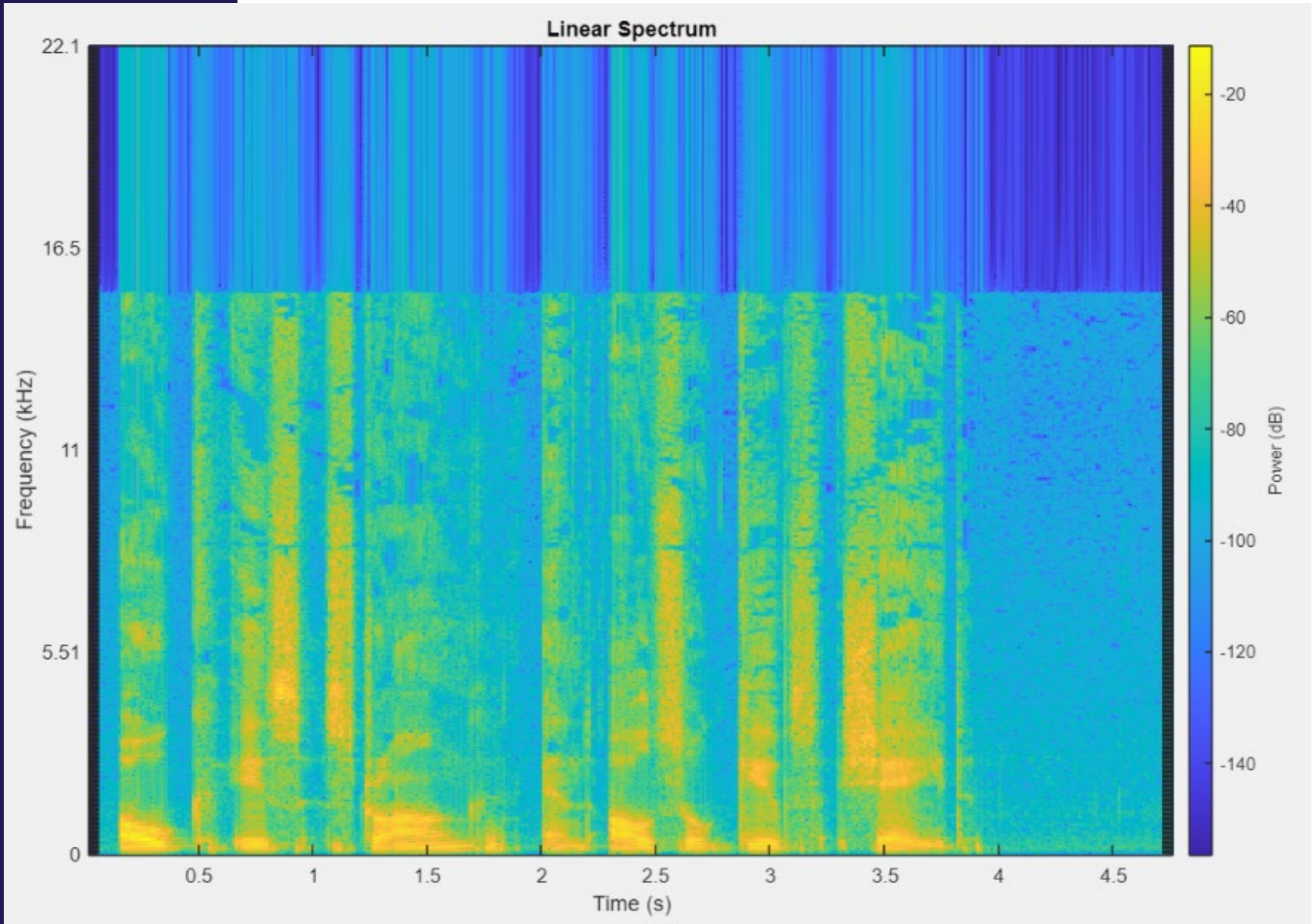
- Spectrogram based
- Cepstral based:
- Mel-frequency cepstral coefficients (MFCCs)
- Gammatone cepstral coefficients (GTCC)



AALBORG UNIVERSITET

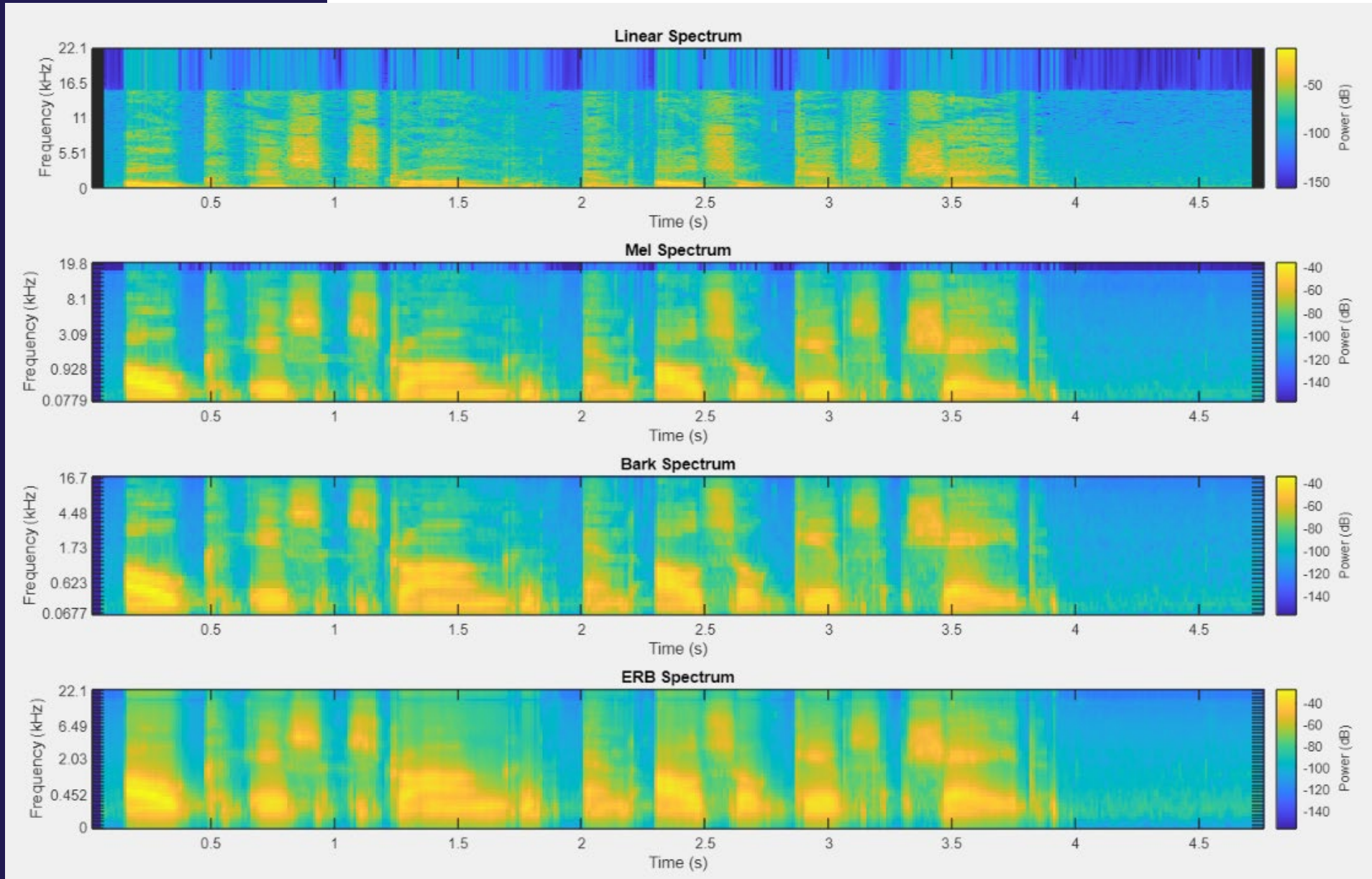
Linear Spectrogram

- Uses Short-time-Fourier transform (STFT)
- Time/frequency resolution depend on window size!



Spectrograms

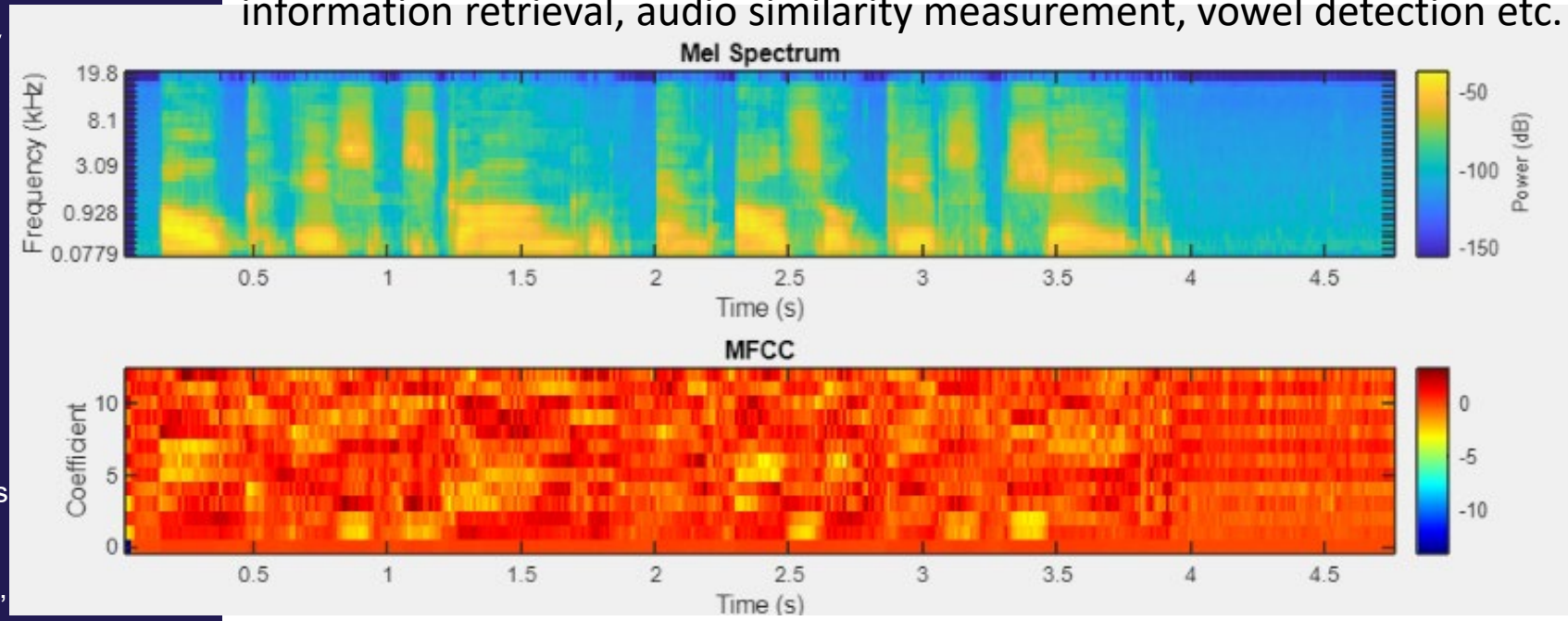
- Linear spectrogram: linear frequency spacing, (relates poorly to human hearing)
- Mel spectrogram: equal distance of pitches in human hearing
- Bark spectrogram: bark scale (Zwicker critical band frequency spacing)
- ERB spectrogram: equivalent rectangular bandwidth frequency spacing → better than Bark according to our current knowledge of the human hearing.

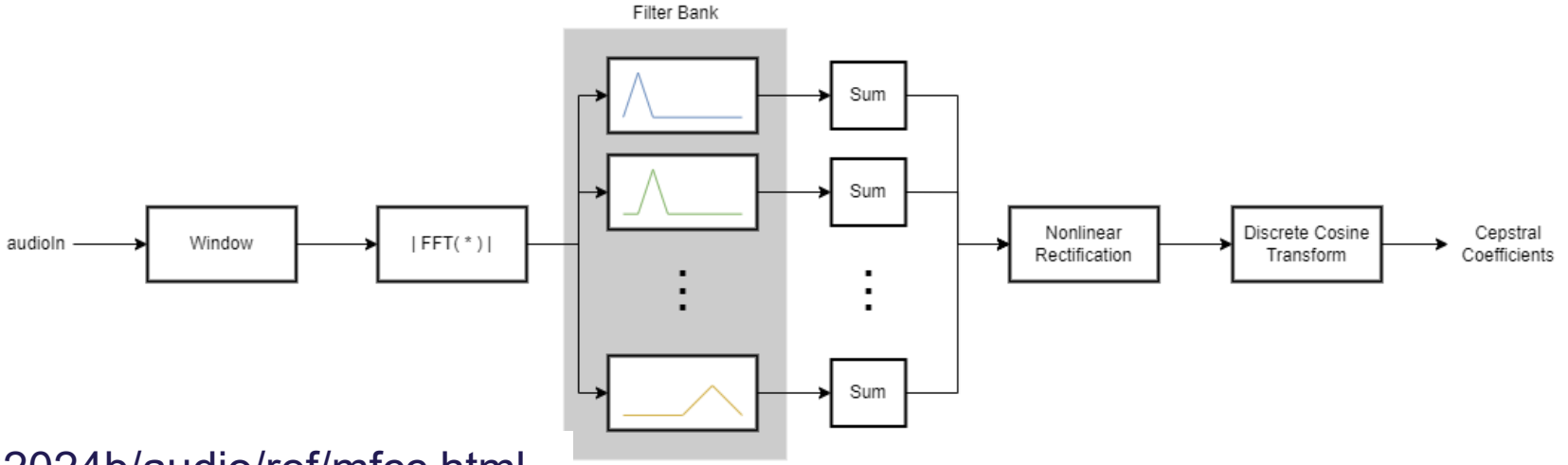


Mel-frequency cepstral coefficients

- Mel-frequency cepstrum coefficients are popular features extracted from speech signals for use in e.g. recognition tasks.
- In the source-filter model of speech, cepstral coefficients are understood to represent the filter (vocal tract).
- The vocal tract frequency response is relatively smooth, whereas the source of voiced speech can be modeled as an impulse train. As a result, the vocal tract can be estimated by the spectral envelope of a speech segment.
- The motivating idea of mel-frequency cepstral coefficients is to compress information about the vocal tract (smoothed spectrum) into a small number of coefficients based on an understanding of our hearing.
- Although there is no hard standard for calculating the coefficients, the basic steps are outlined by the diagram

MFCCs has been widely used in speech recognition, speech enhancement, speaker recognition, music genre classification, music information retrieval, audio similarity measurement, vowel detection etc.

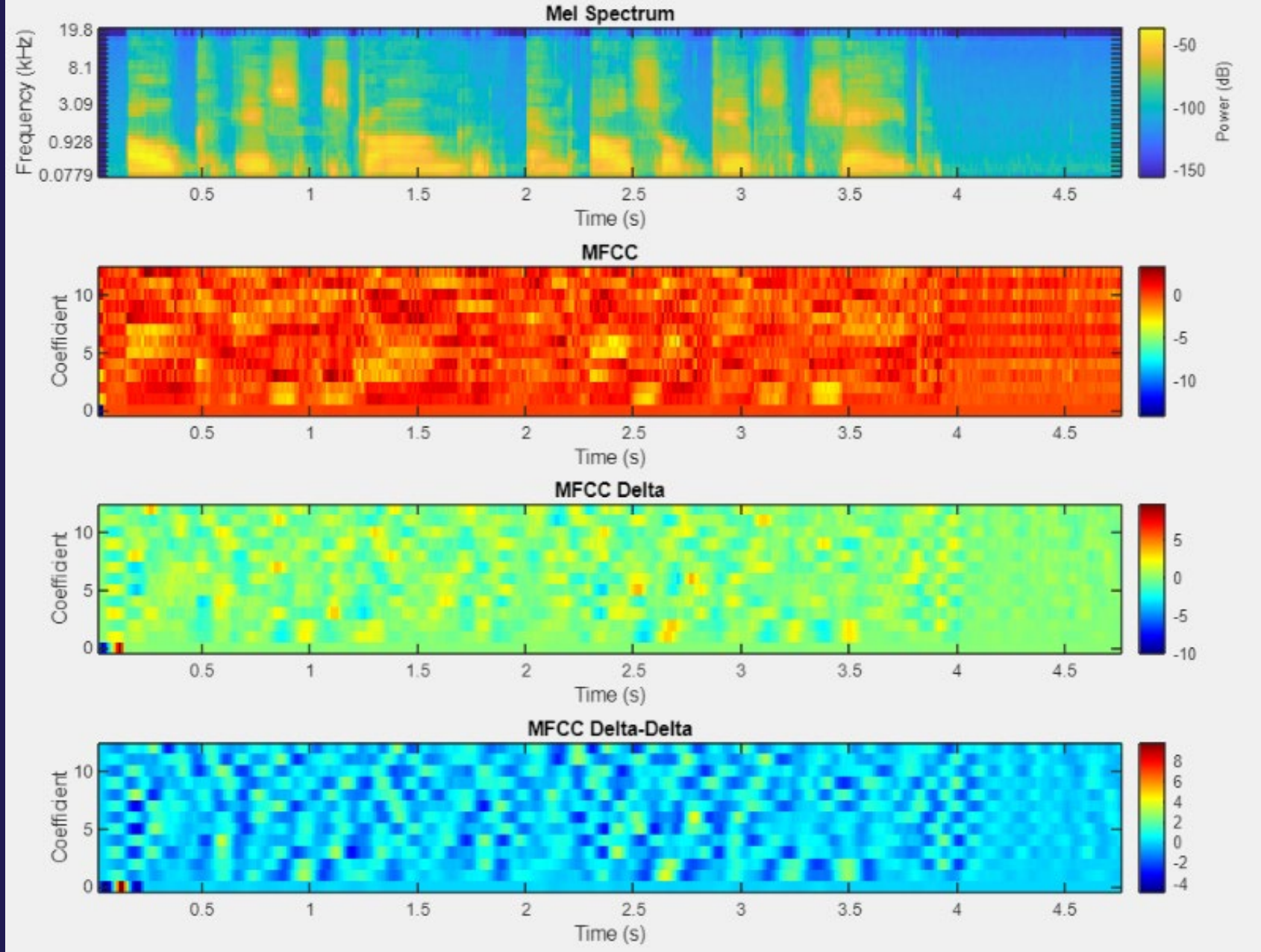




https://se.mathworks.com/help/releases/R2024b/audio/ref/mfcc.html

Mel-frequency cepstral coefficients

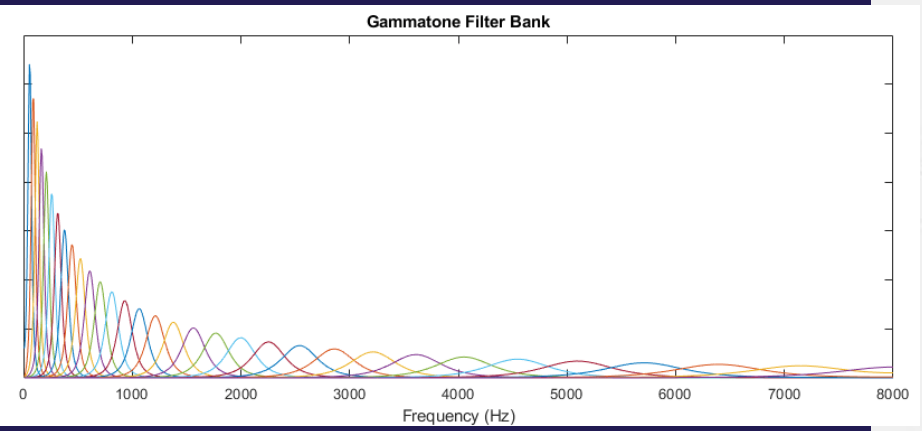
MFCCs has been widely used in speech recognition, speech enhancement, speaker recognition, music genre classification, music information retrieval, audio similarity measurement, vowel detection etc.



- MFCCDelta: Change in coefficients from one frame to the next
- MFCCDeltaDelta : Change in MFCCDelta from one frame to the next

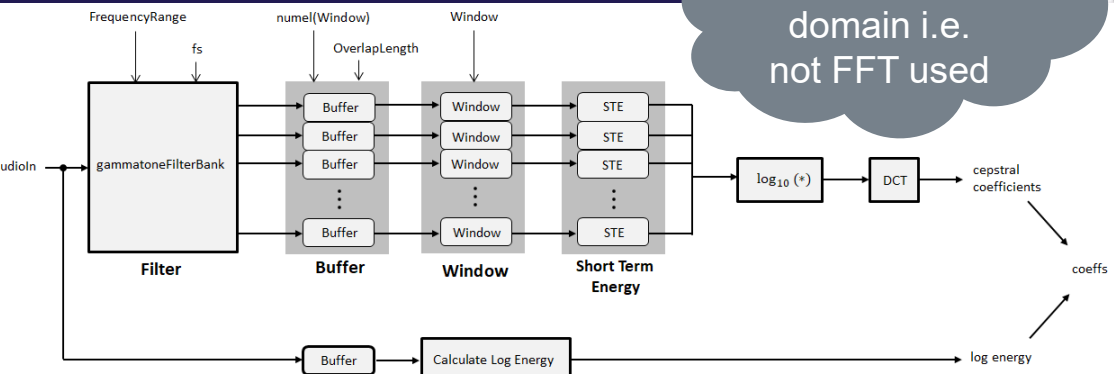
Gammatone - cepstral coefficients

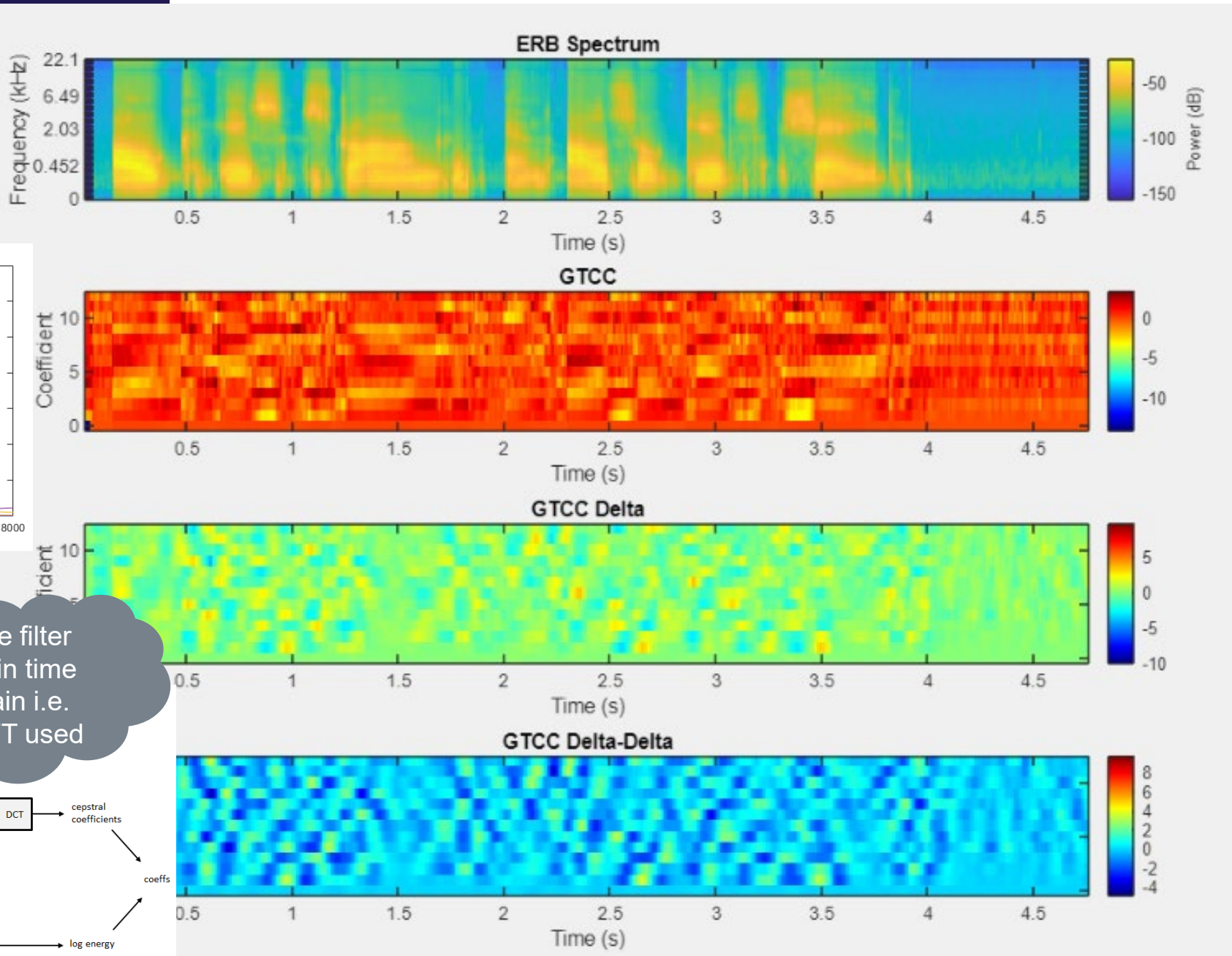
GTCCs similar to MFCCs



- GTCCDelta: Change in coefficients from one frame to the next

Notice filter bank in time domain i.e. not FFT used





Deep features example

- Deep learning is a powerful technique to extract high level features from low level information. The features extracted from the hidden layers of various deep learning models is known as deep features.
- The deep features could be extracted from any deep leaning model like convolutional neural networks (CNNs), deep neural networks (DNNs), recurrent neural networks (RNNs), Deep stacked auto-encoder (SAE), unidirectional long short term memory network (LSTM), bi-directional long short term memory (BLSTM) and other similar models.
- What the feature represent depend on the neural network, and where the feature is taken from.

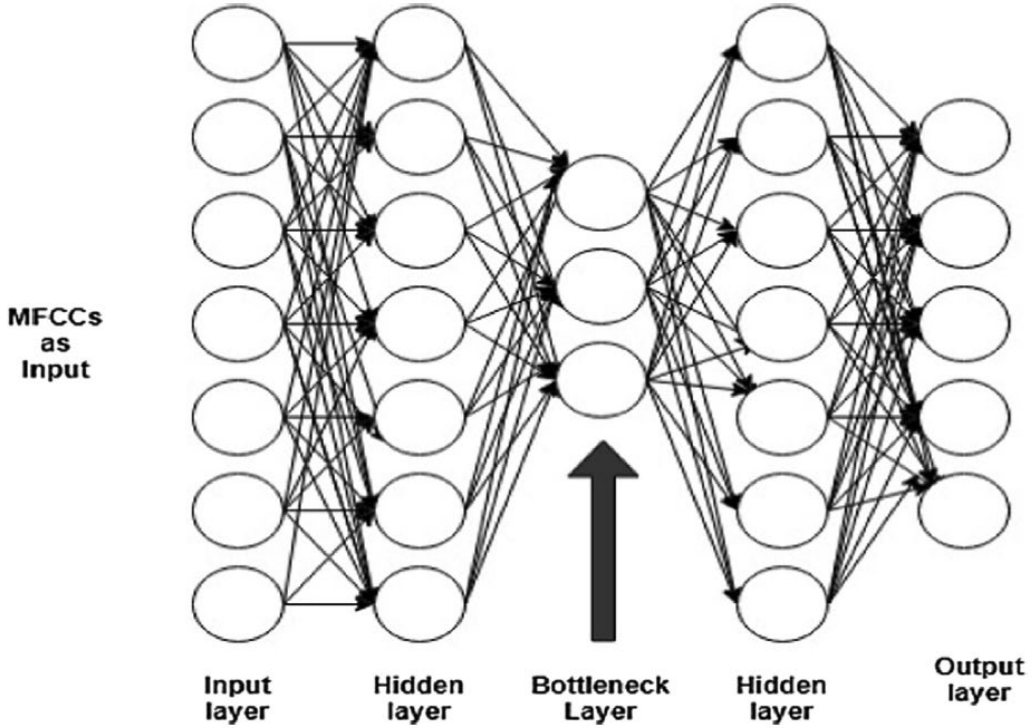
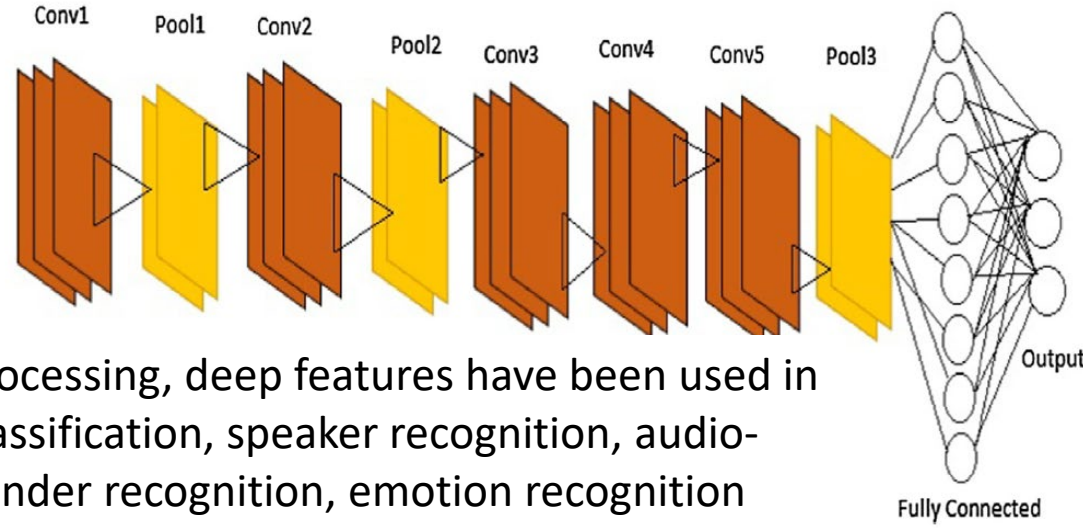


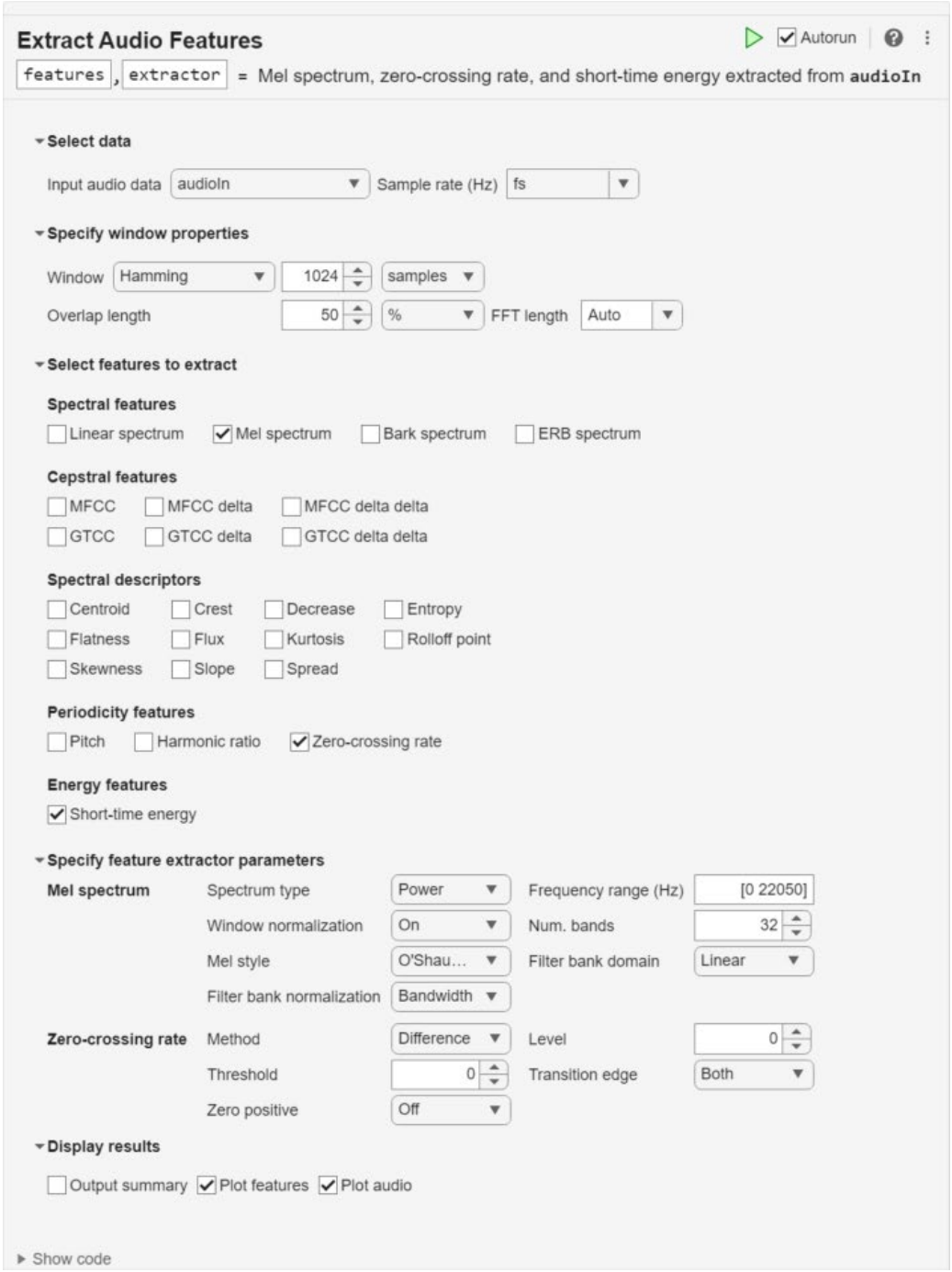
Fig. 18. Deep feature from bottleneck layer.



In audio signal processing, deep features have been used in acoustic scene classification, speaker recognition, audio-video analysis, gender recognition, emotion recognition and spoofing detection.

Extract audio features in Matlab

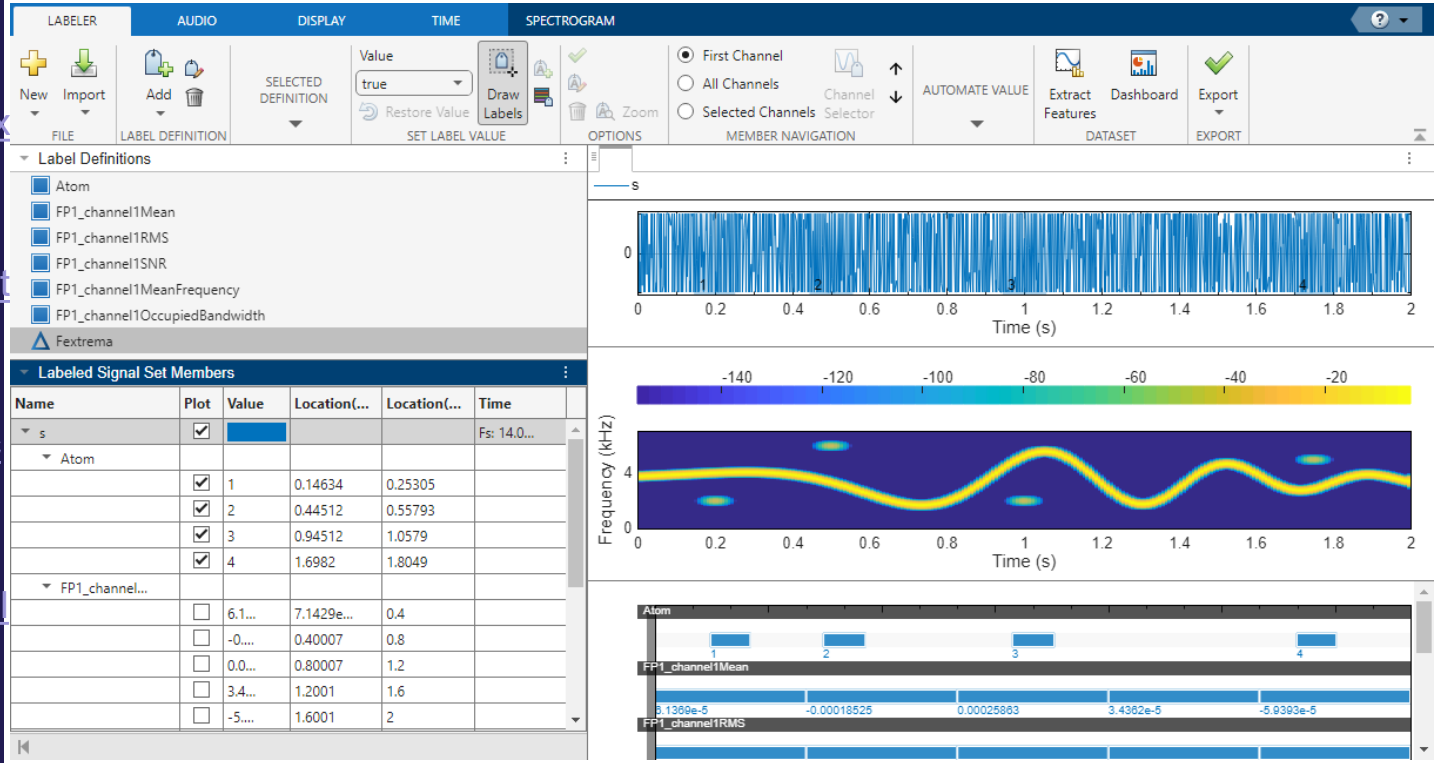
- Usin the Matlab function audioFeatureExtractor **and/or the livescript tool** https://se.mathworks.com/help/audio/ref/extractaudiofeatures.html



AALBORG UNIVERSITET

Other useful functions/apps for audio and AI

- signalTimeFeatureExtractor
  https://se.mathworks.com/help/signal/ref/signaltimefeatureextractor.html
- signalFrequencyFeatureExtractor
  https://se.mathworks.com/help/signal/ref/signalfrequencyfeatureextractor.html
- signalTimeFrequencyFeatureExtractor
  https://se.mathworks.com/help/signal/ref/signaltimefrequencyfeatureextractor.html
- audioDatastore
  https://se.mathworks.com/help/audio/ref/audiodatastore.html
- signalLabeler
  https://se.mathworks.com/help/signal/ref/signallabeler-app.html
- audioDataAugmenter
  https://se.mathworks.com/help/audio/ref/audiodataaugmenter.html



AALBORG UNIVERSITET

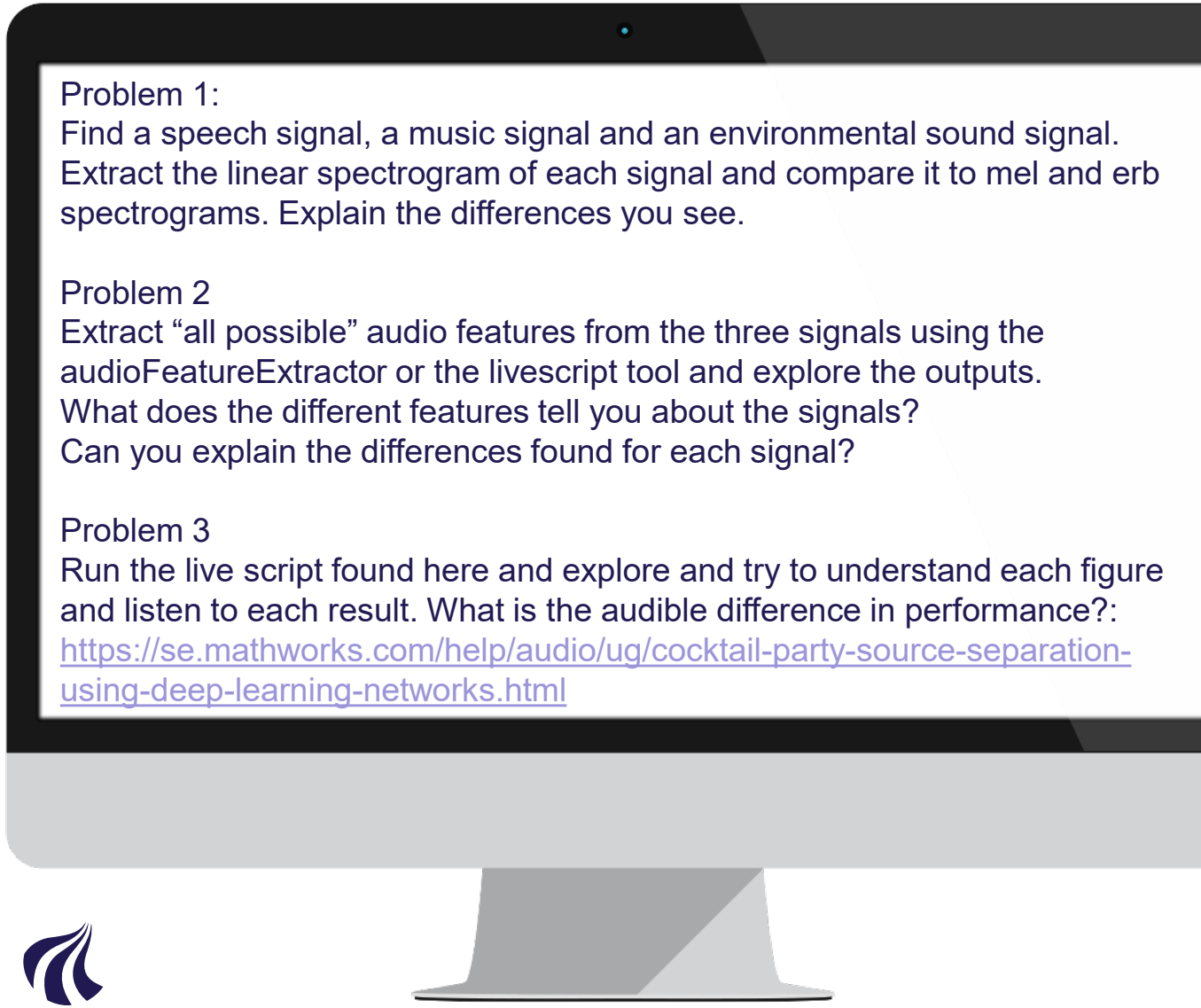
Group rooms?

#A3

101, 102, 103 104 & 106 ?

Extract audio features

- Use the Matlab function
  audioFeatureExtractor
  and/or the livescript tool
  https://se.mathworks.com/help/audio/ref/extractaudiofeatures.html



THE END