MUSIC INFORMATION RETRIEVAL

AND THE PRINCIPLES OF AUDIO PROCESSING AND ANALYSIS

Alexander Schindler
Scientist
Information Management
Center For Digital Safety & Security

AIT Austrian Institute Of Technology Gmbh
Alexander.schindler@ait.ac.at

WHAT IS MUSIC IR?

What is Music IR?

- Searching for Music
  - Searching for music on the Web
  - Query by Humming
  - Similarity Retrieval
  - Identity detection (fingerprinting)

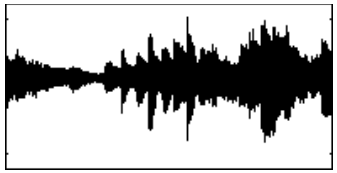
- Extraction of information from music

  → plenty of other tasks!

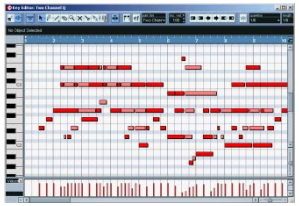
WHAT IS MUSIC?

- **Music**

  Audio: wav, au, mp3, ...

  Symbolic: MIDI, mod, ...

  Scores: Scan, MusicXML

- **Text**
  - Song lyrics
  - Artist Biographies
  - Websites: Fanpages, Blogs, Album Reviews, Genre descriptions

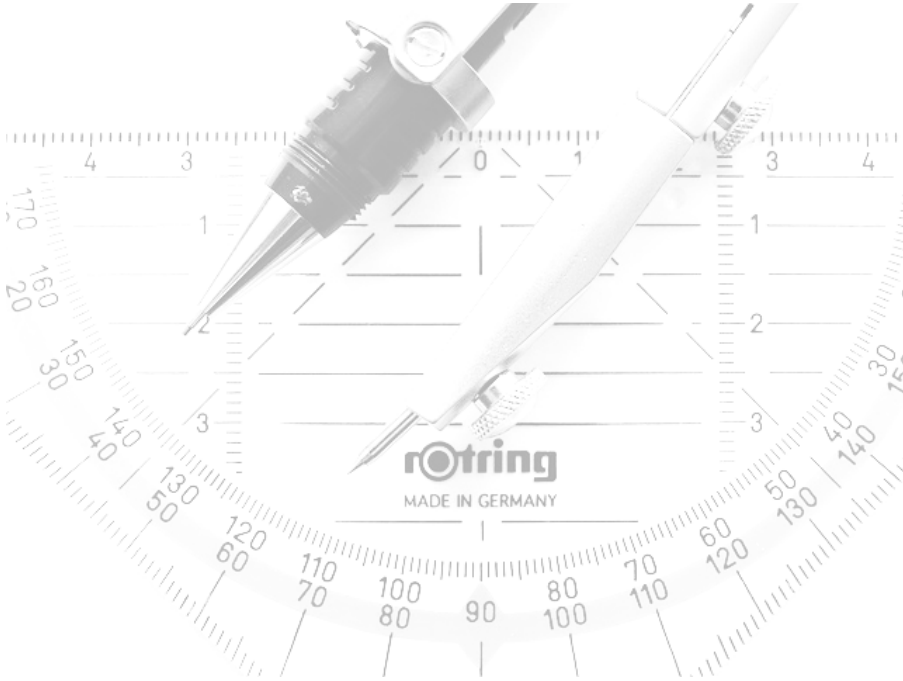
- **Community data**
  - Market basket
  - Tags
  - Social Networks
    - Spotify
    - Last.fm

- **Video/Images**
  - Album covers
  - Music videos

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

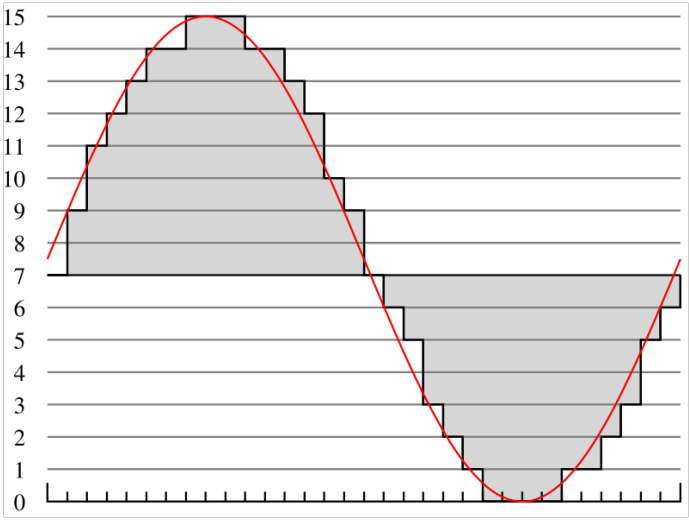
FEATURE EXTRACTION FROM MUSIC

Extracting Music Information



TOO MUCH AUDIO DATA

- Digital Audio
  - Sampling Rate: 44,100 Hz
  - 16-bit resolution for each channel
    - 2 channels for stereo
  - 88,200 Integers per second



EXCERCISE: FIND DOCUMENTS CONTAINING THE WORD „MUSIC“

Document 1:

*“Most of these issues stem from the commercial interest in **music** by record labels, and therefore imposed rigid copyright issues, that prevent researchers from sharing their **music** collections with others. Subsequently, only a limited number of data sets has risen to a pseudo benchmark level, i.e. where most of the researchers in the field have access to the same collection.“*

Document 2:

*“The Echonest Analyzer [5] is a **music** audio analysis tool available as a free Web service accessible over the Echonest API and as a commercially distributed standalone command line tool. The Analyzer implements an onset detector which is used for segmentation.”*

Document 3:

*“The Million Song Dataset (MSD), a collection of one million **music** pieces, enables a new era of research of **Music** Information Retrieval methods for large-scale applications. It comes as a collection of meta-data such as the song names, artists and albums, together with a set of features extracted with the The Echo Nest services, such as loudness, tempo, and MFCC-like features.”*

EXCERCISE: FIND SONGS WITH STRINGS

Song 1:

83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7, 66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2, 60, 9, 23, 43, 21, 98, 34, 29, 62, 26, 72, 38, 98

Song 2:

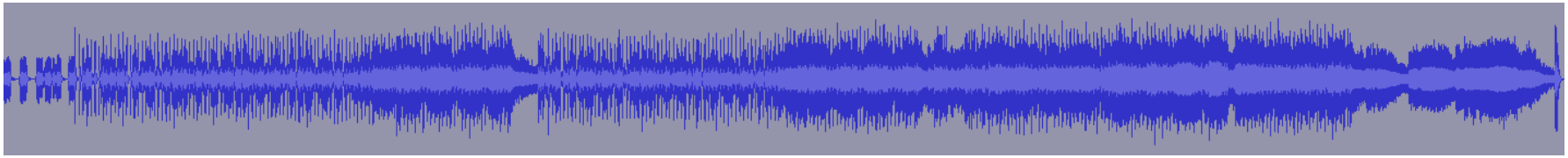
55, 96, 11, 49, 83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7, 66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2, 60, 9, 23, 43, 21, 98, 34, 29, 62, 26, 72, 38, 98, 55, 96, 11, 49, 83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7

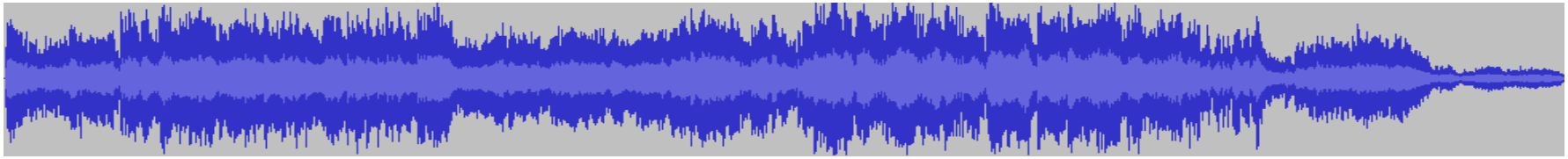
Song 3:

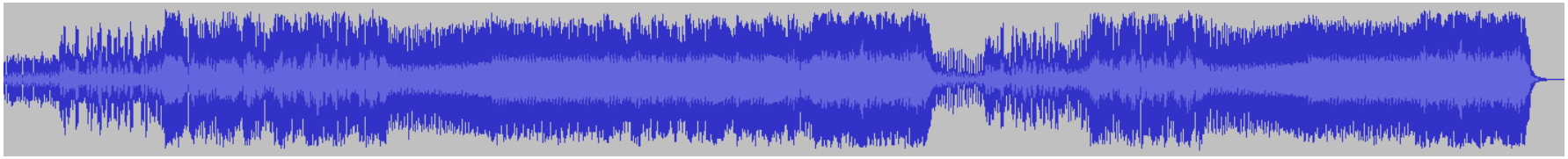
66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2, 60, 9, 23, 43, 21, 98, 34, 29, 62, 26, 72, 38, 98, 55, 96, 11, 49, 83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7, 66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2

AUSTRIAN INSTITUTE OF TECHNOLOGY

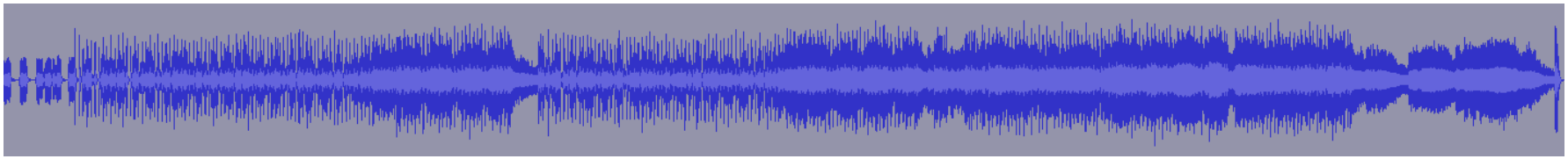
EXCERCISE: SAME GENRE?



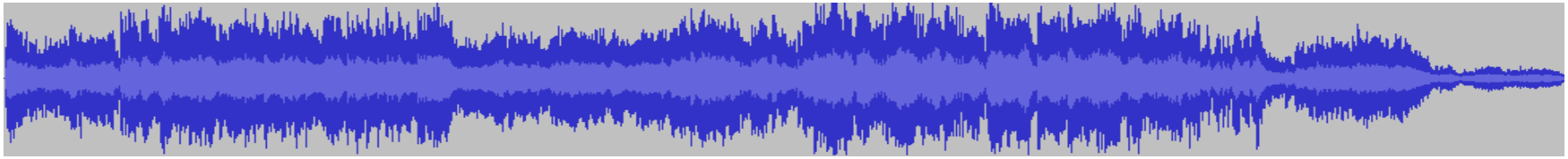




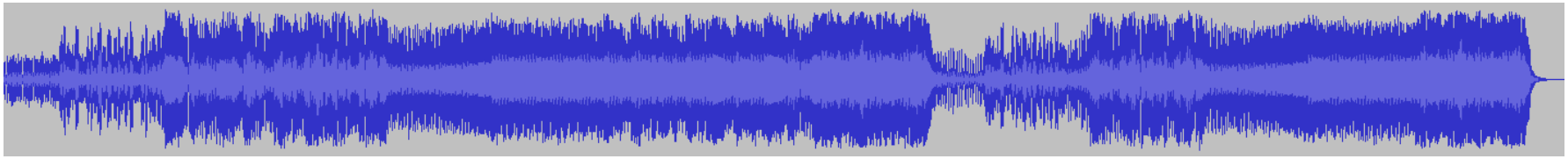
EXCERCISE: IDENTIFY SONGS



AC-DC – Highway to Hell



John Williams – Star Wars Main Theme

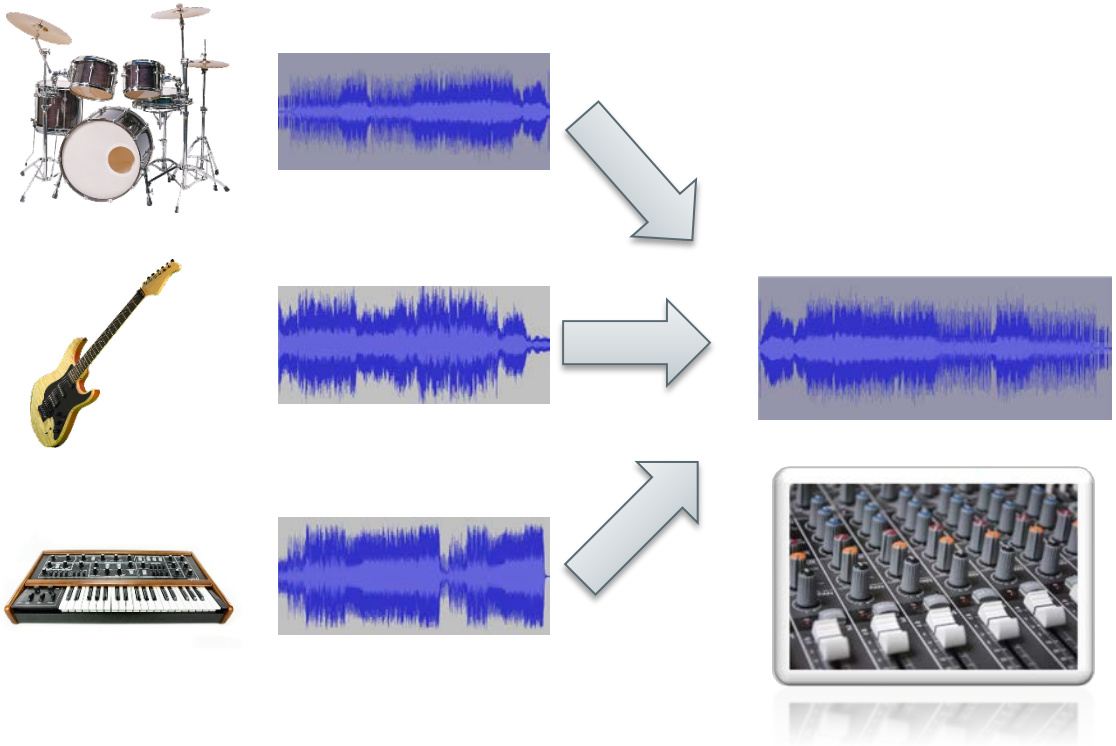


Rihanna feat. Calvin Harris – We Found Love

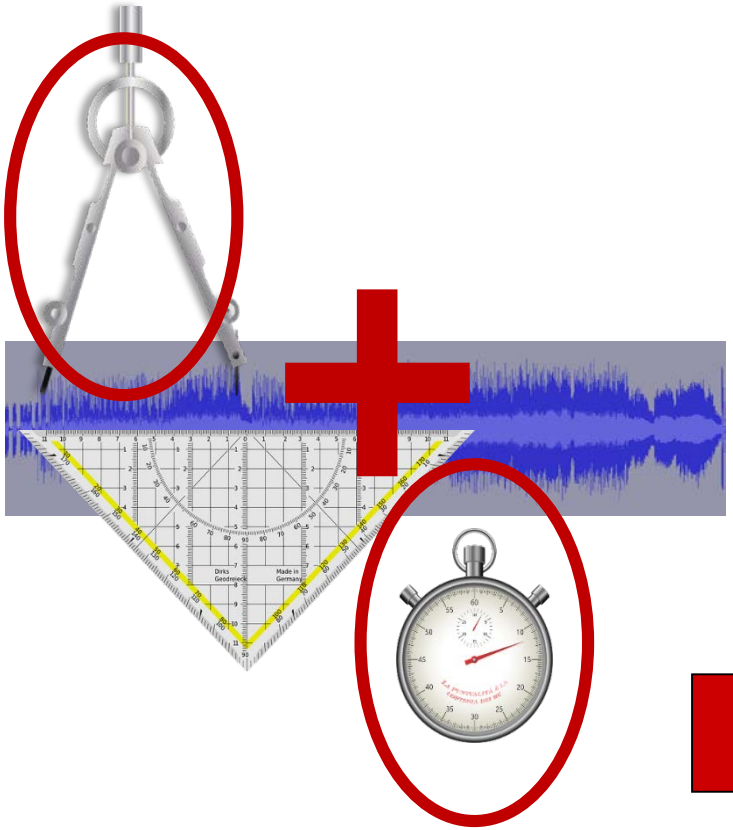
AUDIO FEATURE EXTRACTION

- Reduce audio data by extracting information about:
  - Pitch
  - Timbre
  - Rhythm
  - etc.
- → extract „audio descriptors“

PROBLEM: SOURCE SEPARATION

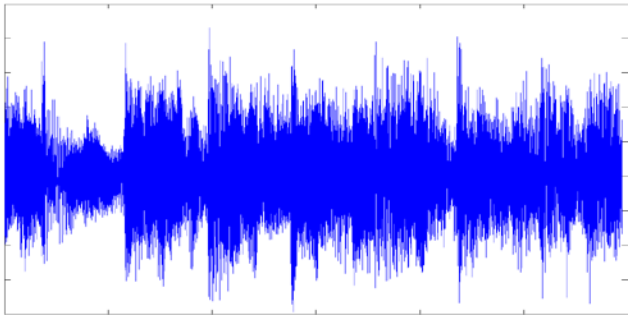


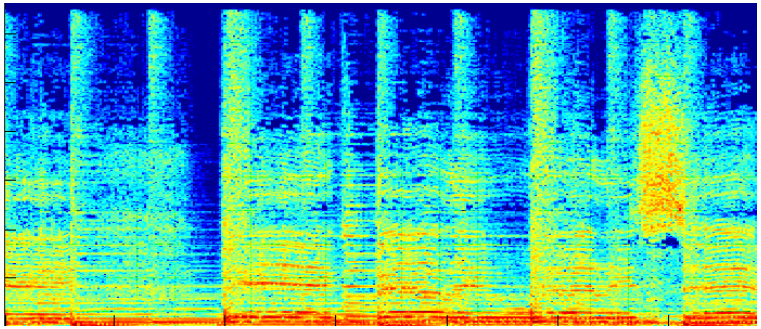
AUDIO FEATURE EXTRACTION



Timbre
Melody
Instrumentation
Harmony
Structure
Complexity
Rhythm
Tempo
Sociocultural Background
Lyrics
Mood

SIGNAL PROCESSING





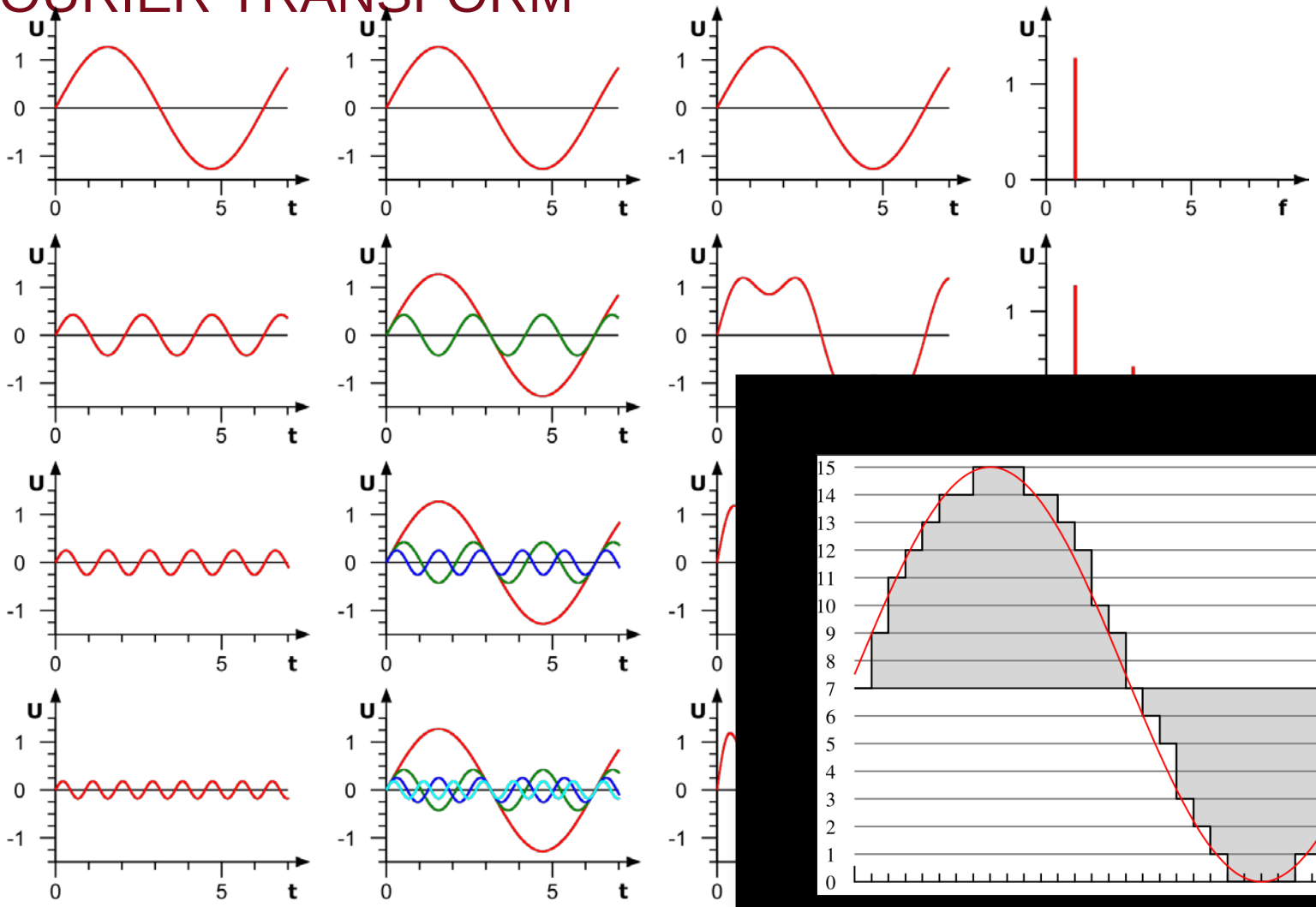
Time Domain
(„Wave Form“)

Frequency Domain
(„Spectrum“)

Time-Frequency Transformation

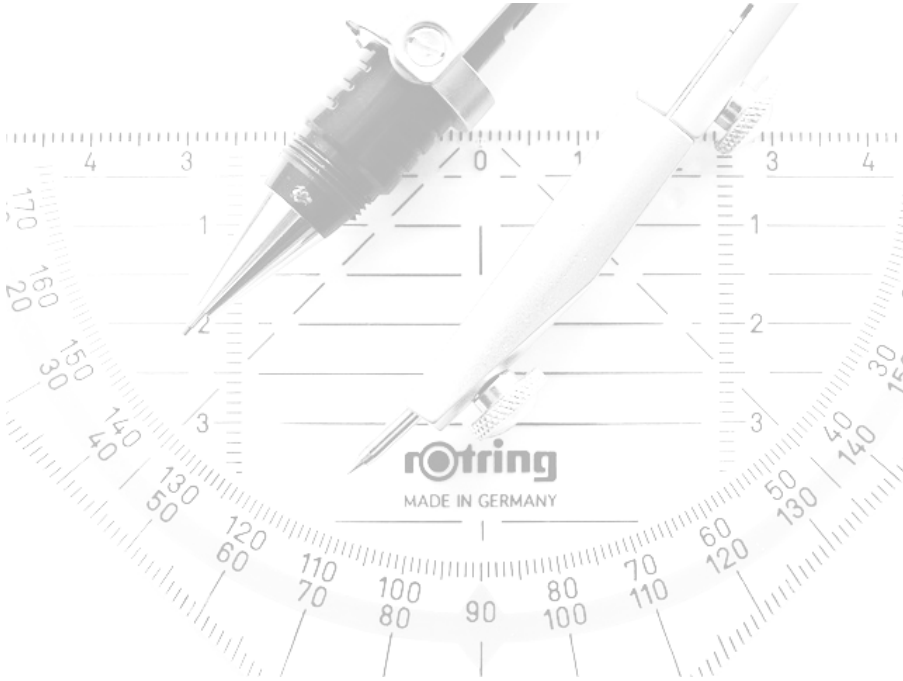
Fourier Transform (FFT)
Discrete Cosine Transform (DCT)
Wavelet Transform

FOURIER TRANSFORM



AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

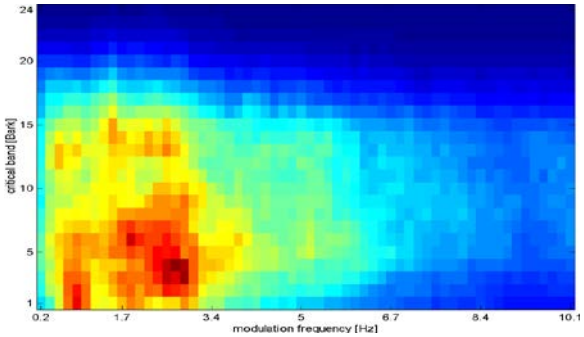
FEATURE EXTRACTION FROM MUSIC



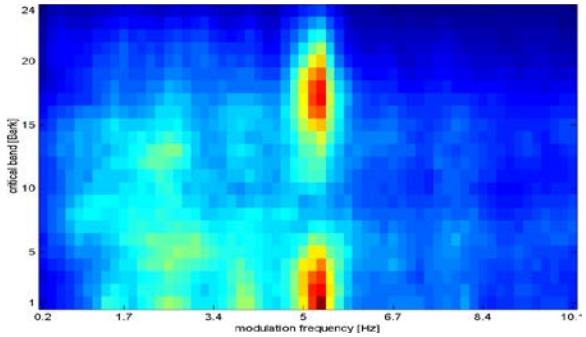
By example…

RHYTHM PATTERN (RP)

- fluctuations on critical frequency bands (a.k.a. Fluctuation Pattern)
- covers rhythm in the broad sense

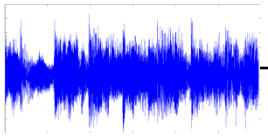


Classical



Rock

RHYTHM PATTERN (RP)



Segmentation (6 s)

Spectrogram (FFT)

Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone

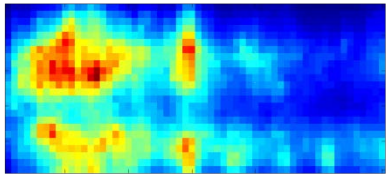
Modulation Amplitudes

Fluctuation Strength

Filtering/Smoothing

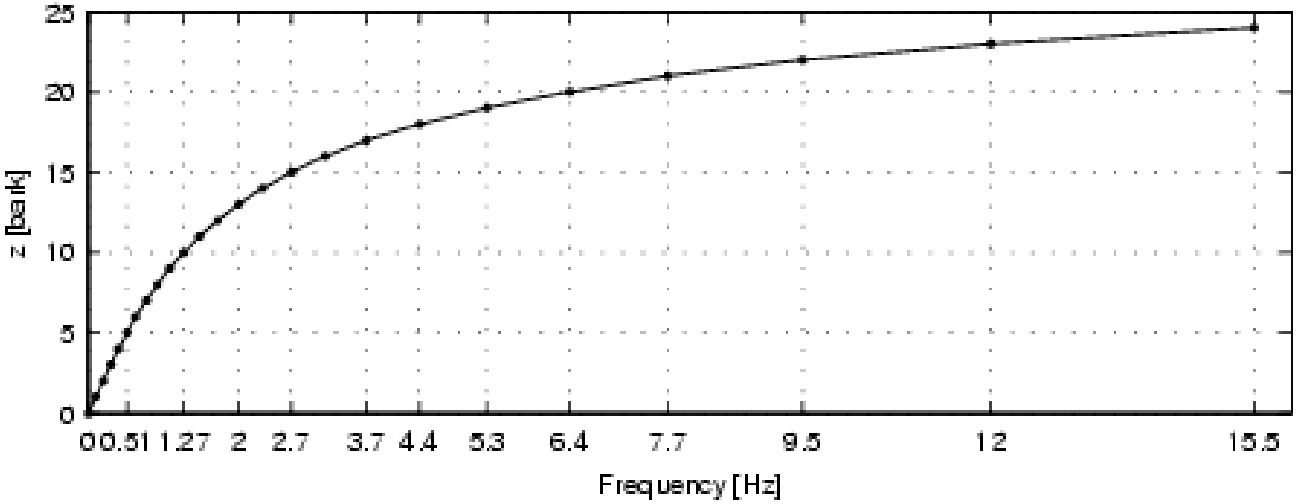
Rhythm Pattern features

(1440 dimensions)



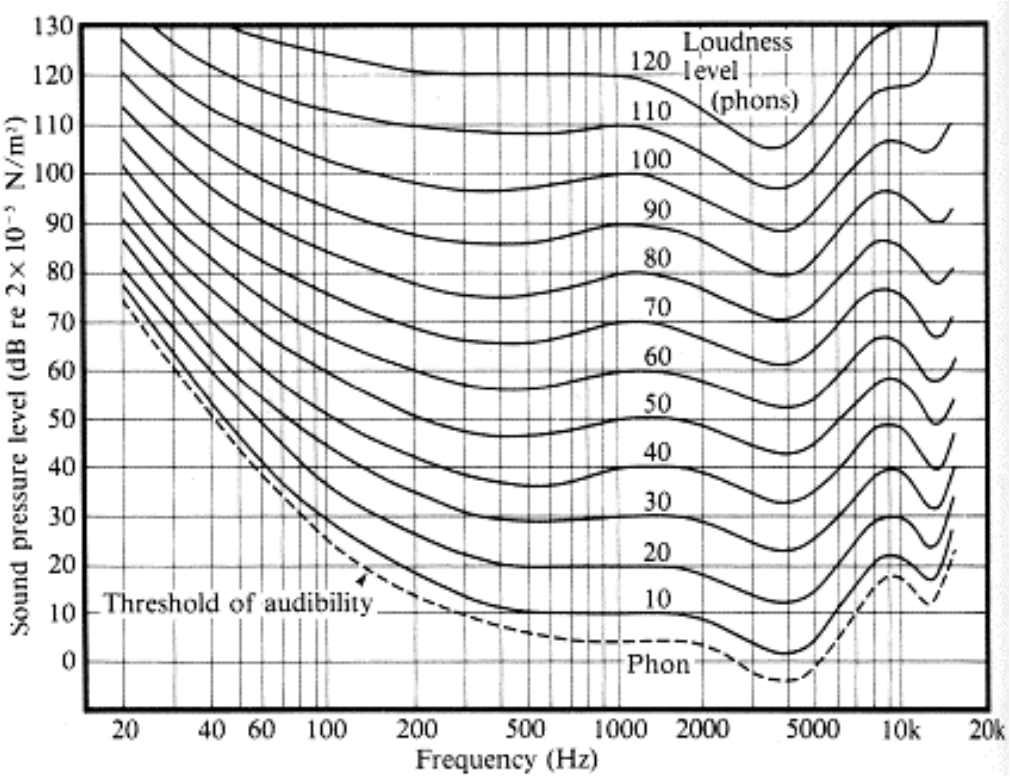
BARK SCALE

- psychoacoustical scale (related to Mel scale)
- 24 „critical bands“ of hearing (non-linear)
- proposed by Eberhard Zwicker in 1961



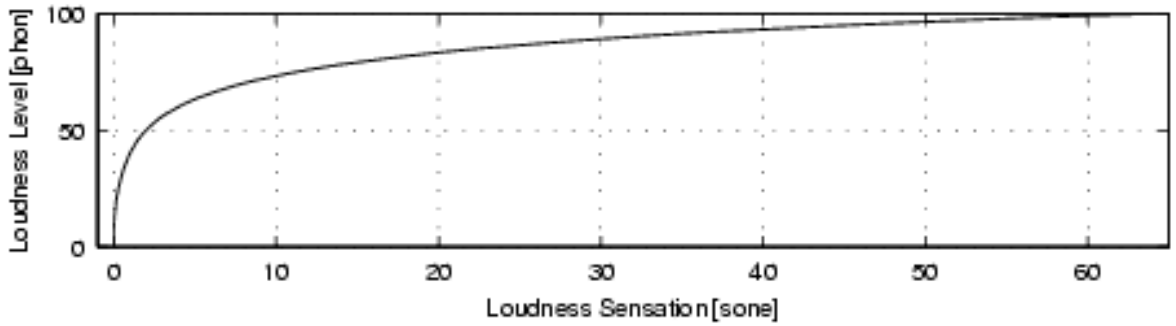
Equal loudness curves (Phon)

- Relationship between sound pressure level in decibel and hearing sensation is not linear
- Perceived loudness depends on frequency of the tone
- equal loudness contours for 3, 20, 40, 60, 80, 100 phon



on-line test: http://www.phys.unsw.edu.au/jw/hearing.html

Sone Transformation



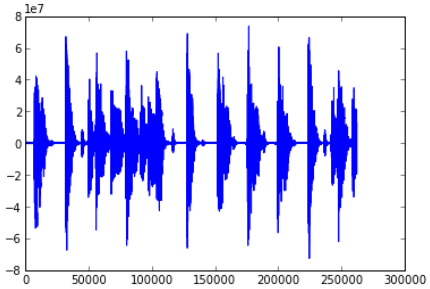
Sone	1	2	4	8	16	32	64
Phon	40	50	60	70	80	90	100

- Perceived loudness measured in Phon does not increase linearly
- Transformation into Sone
- Up to 40 phon slow increase in perceived loudness, then drastic increase
- Higher sensibility for certain loudness differences

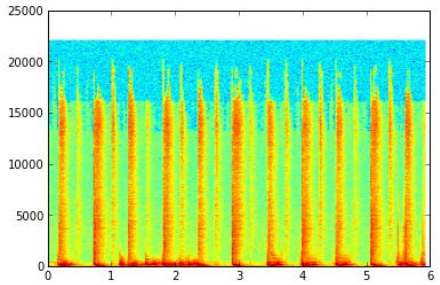
RHYTHM PATTERN (RP): 2 EXAMPLES

Queen – Another One Bites The Dust (first 6 seconds)

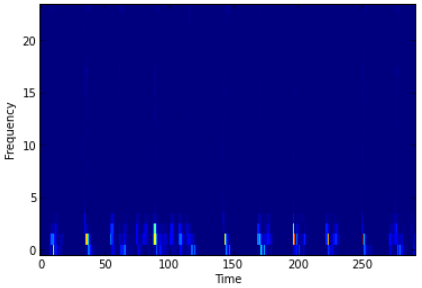
PCM Audio Signal



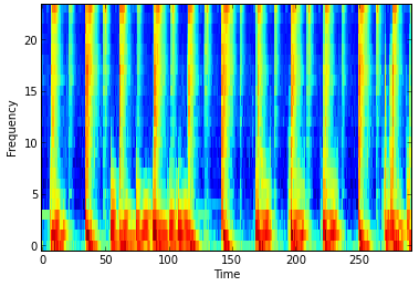
Power Spectrum



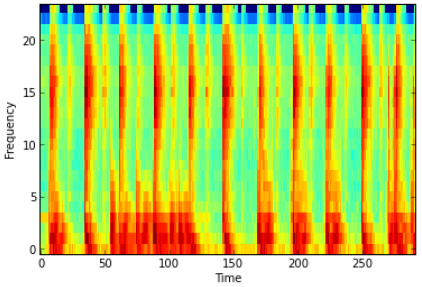
Bark Scale



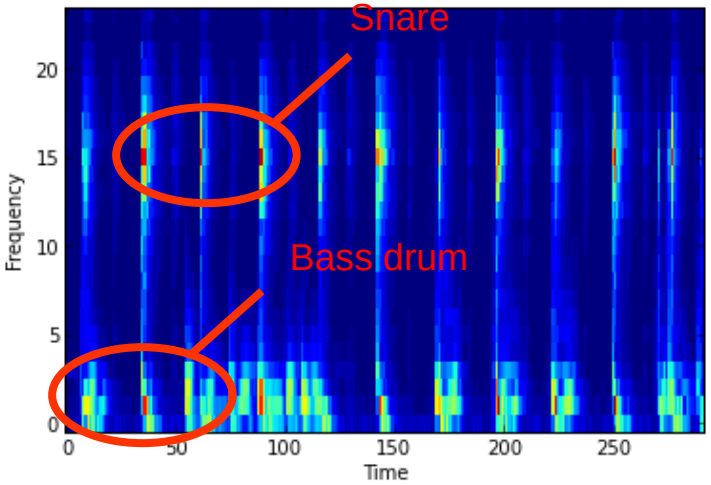
Decibel



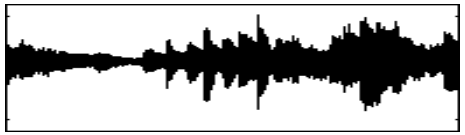
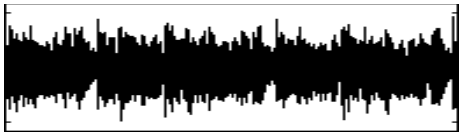
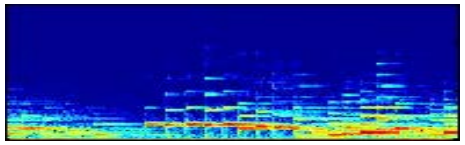
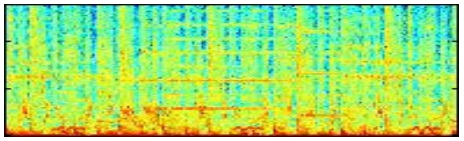
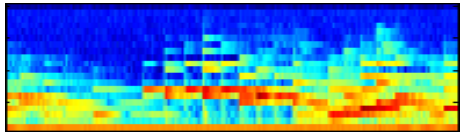
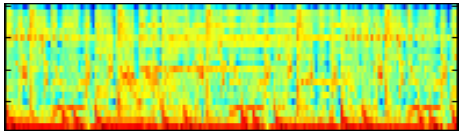
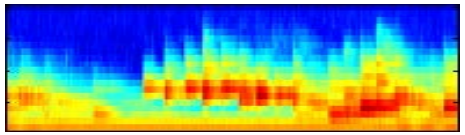
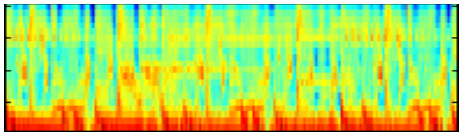
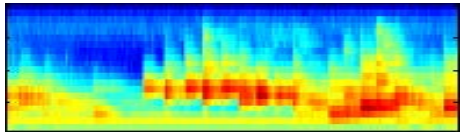
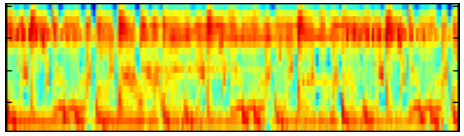
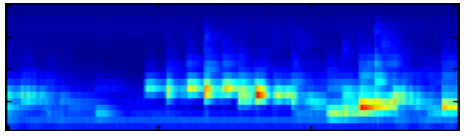
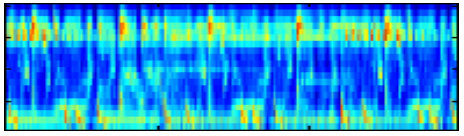
Phon



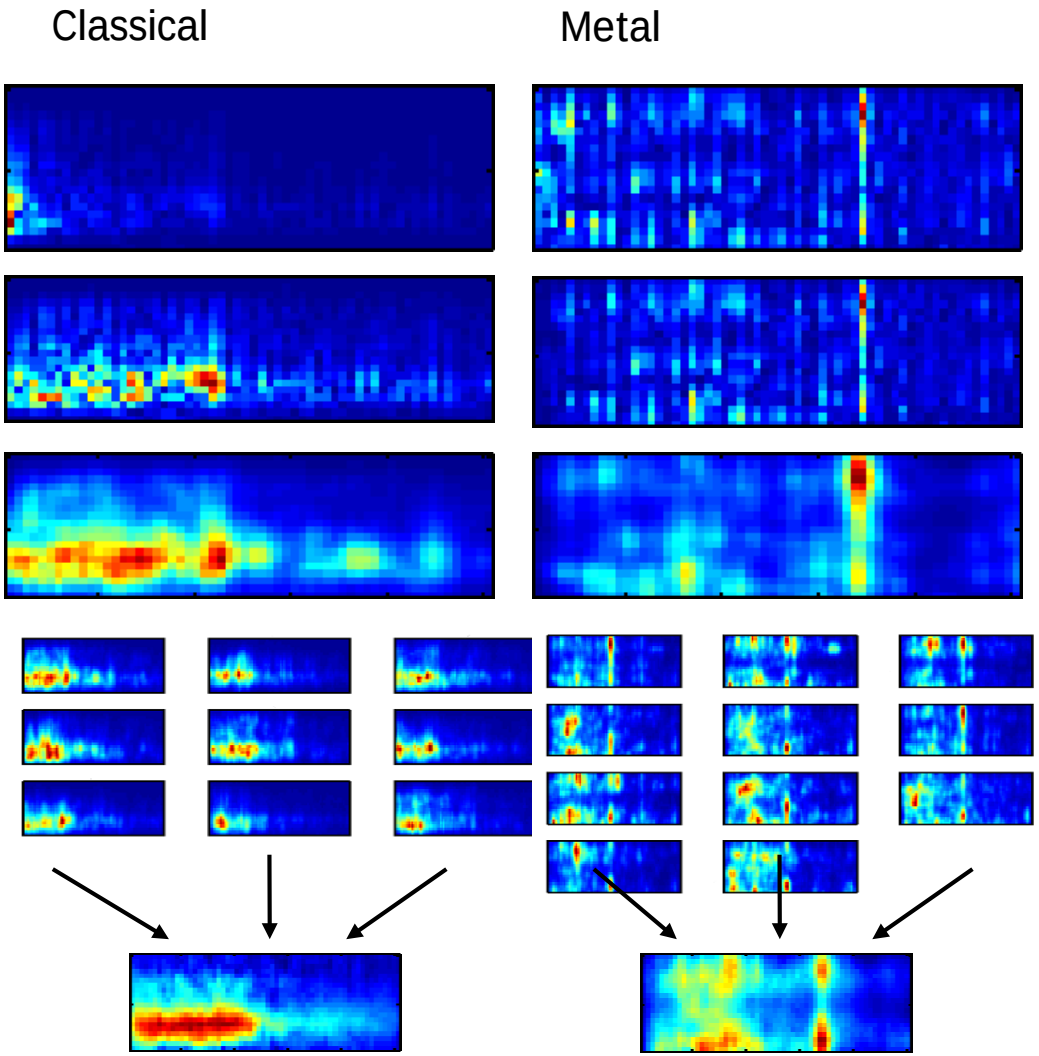
Sone



RHYTHM PATTERN (RP): 2 EXAMPLES

Classical	Metal	
		**PCM Audio Signal**
		Power Spectrum
		Frequency Bands
		Masking Effects
		Phon
		Sone

RHYTHM PATTERN (RP): 2 EXAMPLES



modulation amplitude spectrum (“cepstrum”)

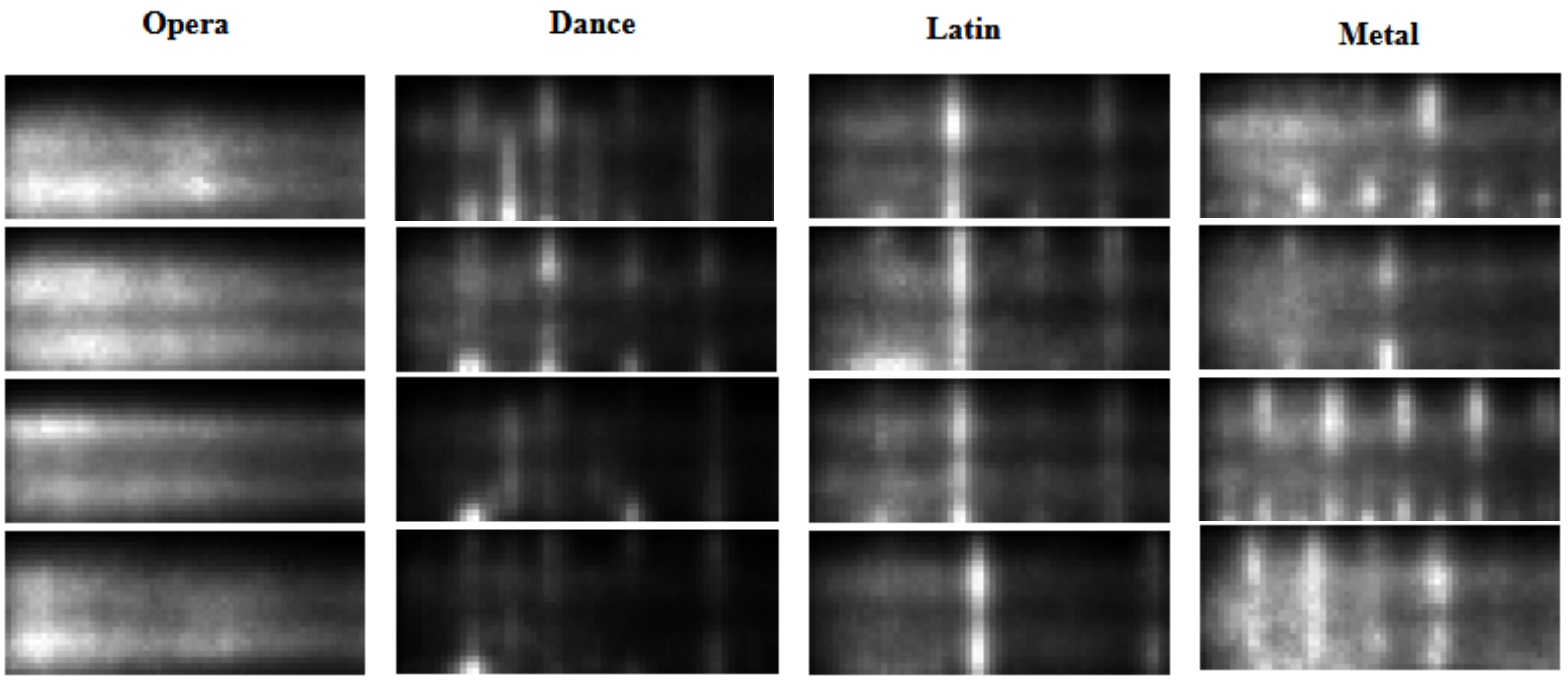
Fluctuation Strength

Filter (Gradient, Gauss)

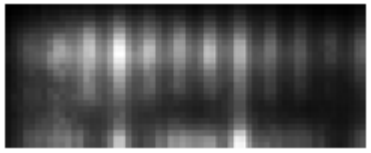
Median

24*60=
1.440-dim feature vec.

RP PER GENRE

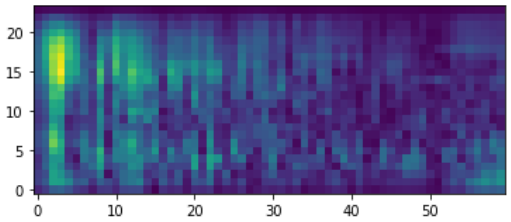


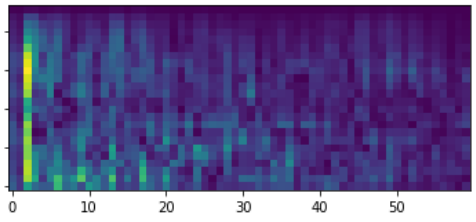
Modulated Synthesizer



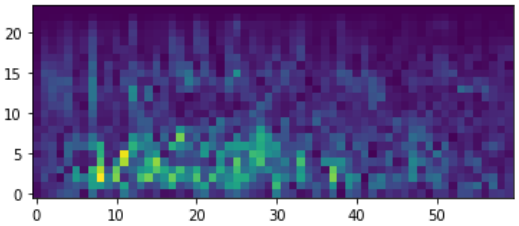
SOUND EVENTS

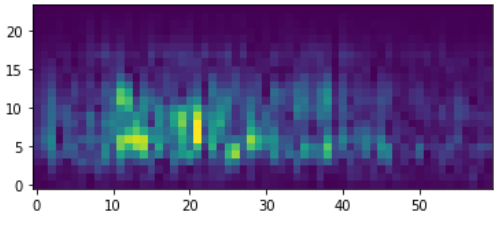
- Gunshots



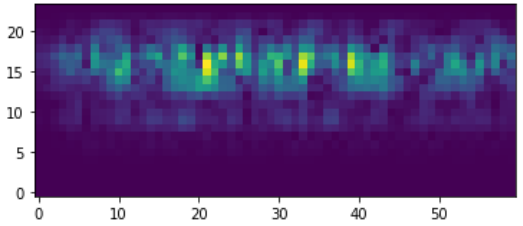


- Emergency Vehicles

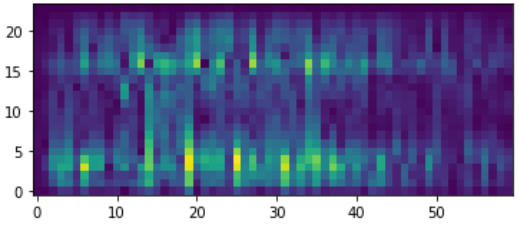




▸ Fire alarm



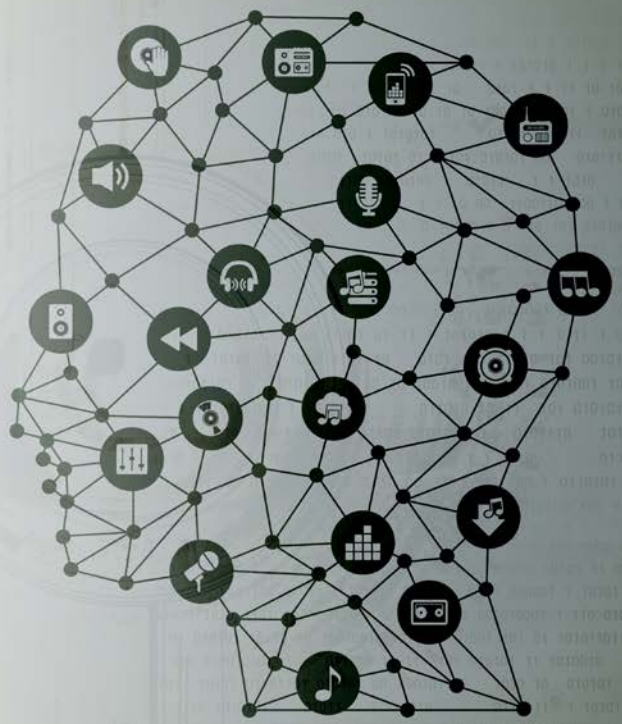
▸ Interview



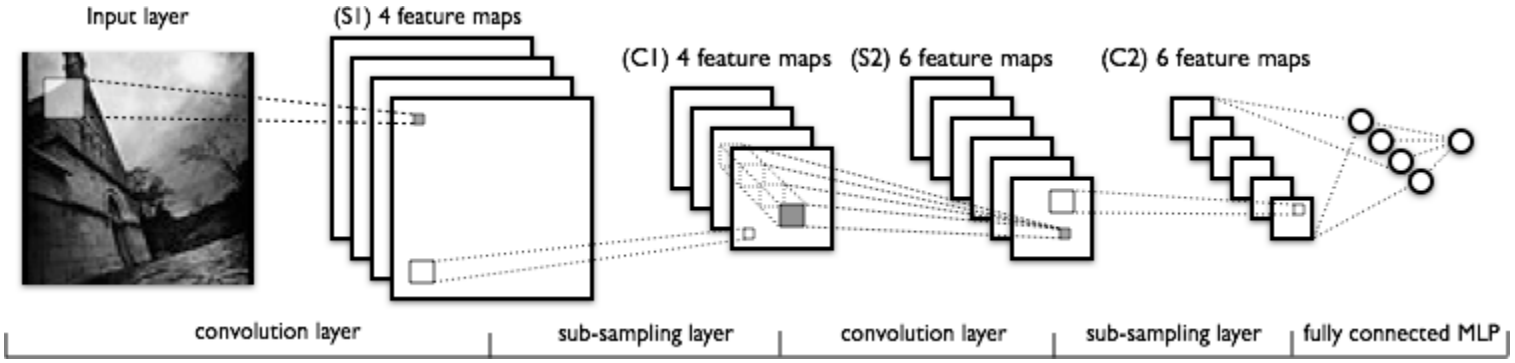
AUSTRIAN INSTITUTE OF TECHNOLOGY

DEEP LEARNING

for
Music Information Retrieval



CONVOLUTIONAL NEURAL NETWORK (CNN)



http://deeplearning.net/tutorial/lenet.html

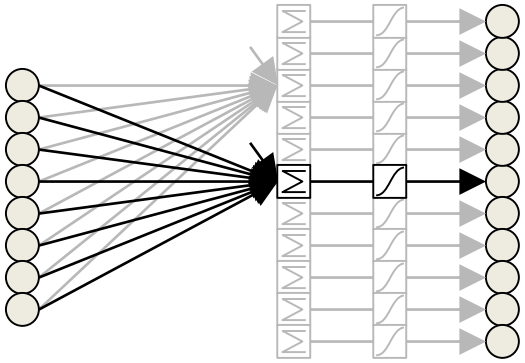
Combines three types of layers:

- **Convolutional layer:** performs 2D convolution of 2D input with multiple learned 2D kernels
- **Subsampling layer:** replaces 2D patches by their maximum (“max-pooling”) or average
- **Fully-connected layer:** computes weighted sums of its input with multiple sets of learned coefficients

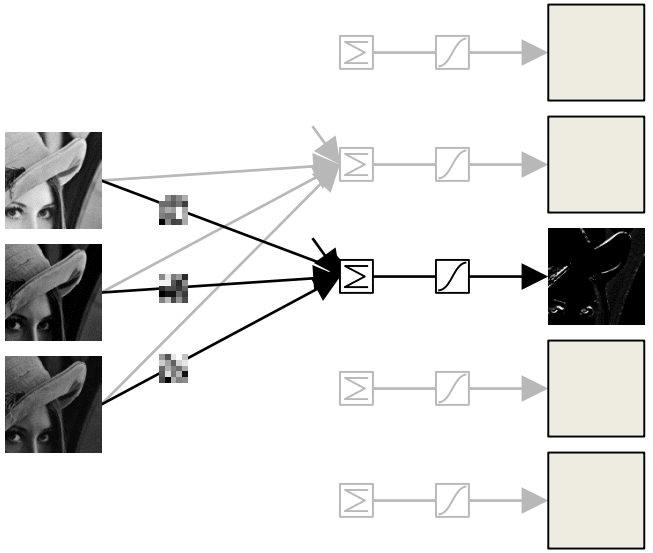
Applies a nonlinear function after each linear operation (without, a deep network would be linear despite its depth).

Full vs. Convolutional Layer / Network

- Fully-connected layer:
- Each **input** is a **scalar** value, each **weight** is a **scalar** value, each output is the sum of inputs **multiplied** by weights.

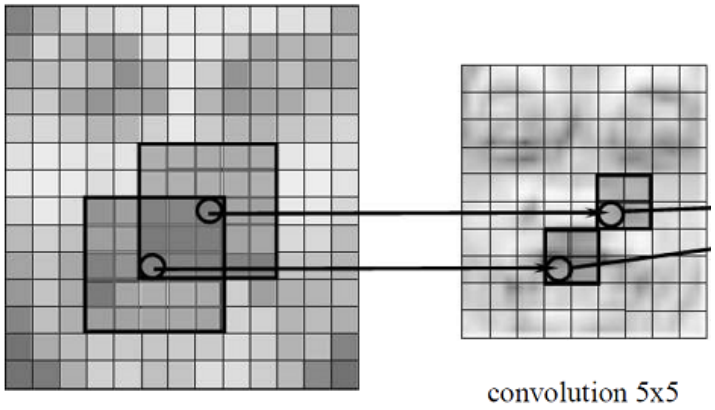


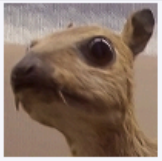
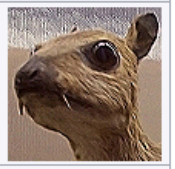
- Convolutional layer:
- Each **input** is a **tensor** (e.g., 2D), each **weight** is a **tensor**, each output is the sum of inputs **convolved** by weights.



Motivation for Convolutions

- Apply local filter kernels
- These kernels are the neurons that are learned

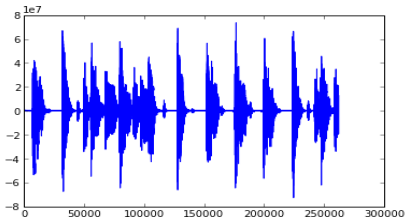


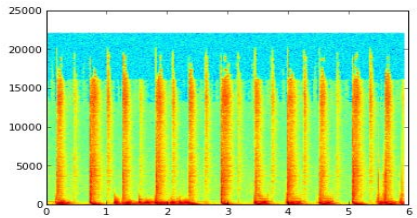
Operation	Kernel	Image result
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	

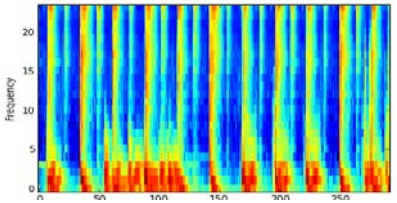
Images: http://sanghyukchun.github.io/75/
https://en.wikipedia.org/wiki/Kernel_(image_processing)

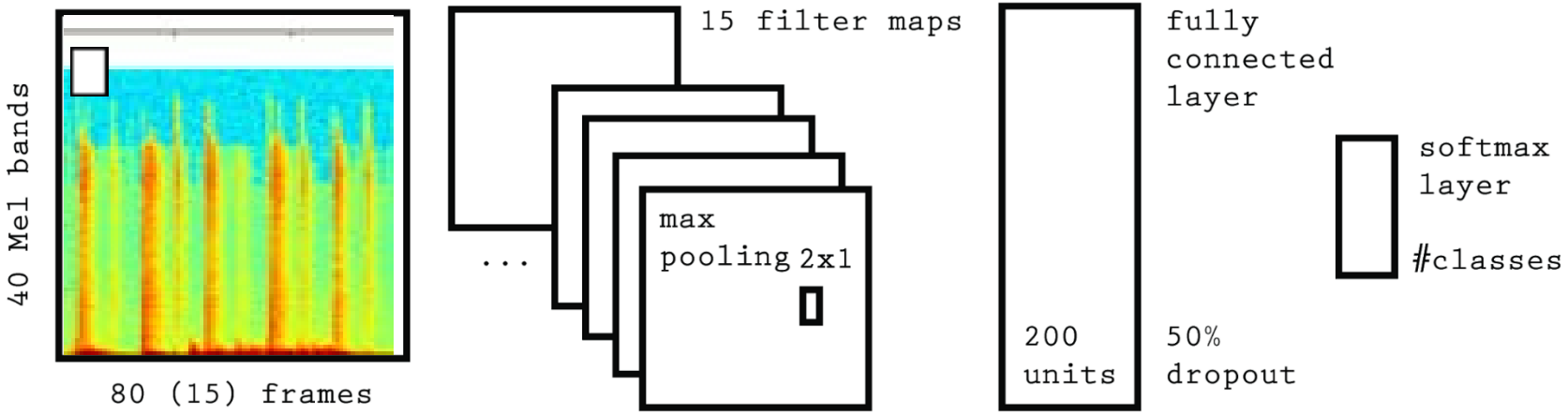
DEEP LEARNING FOR MUSIC IR

Pre-Processing: Waveform → Spectrogram → 40 Mel bands → Log scale







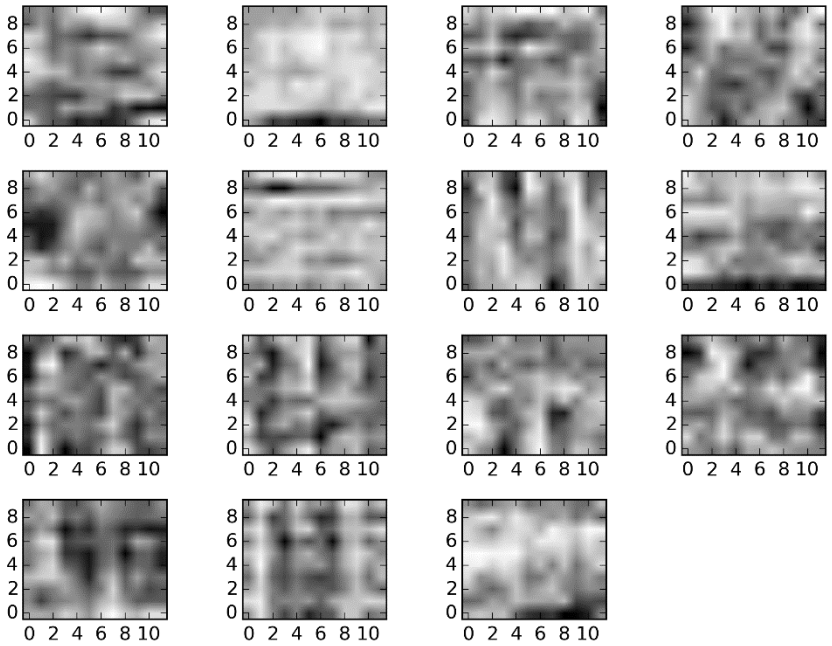


Winning algorithm MIREX 2015 music/speech classification task (99.73%) by Thomas Lidy

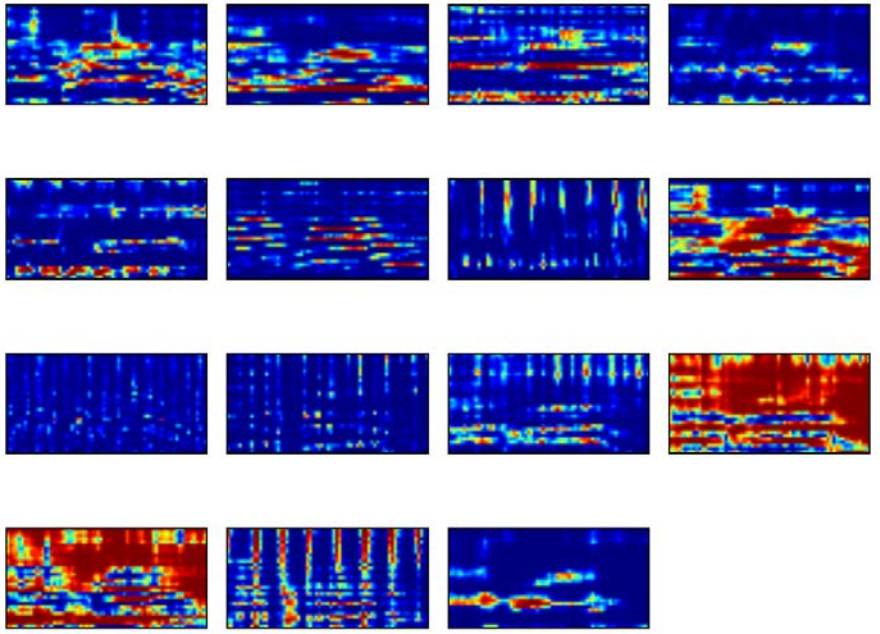
VISUALIZING CNN FILTERS LEARNED FOR MUSIC/SPEECH CLASSIFICATION

AUSTRIAN INSTITUTE OF TECHNOLOGY

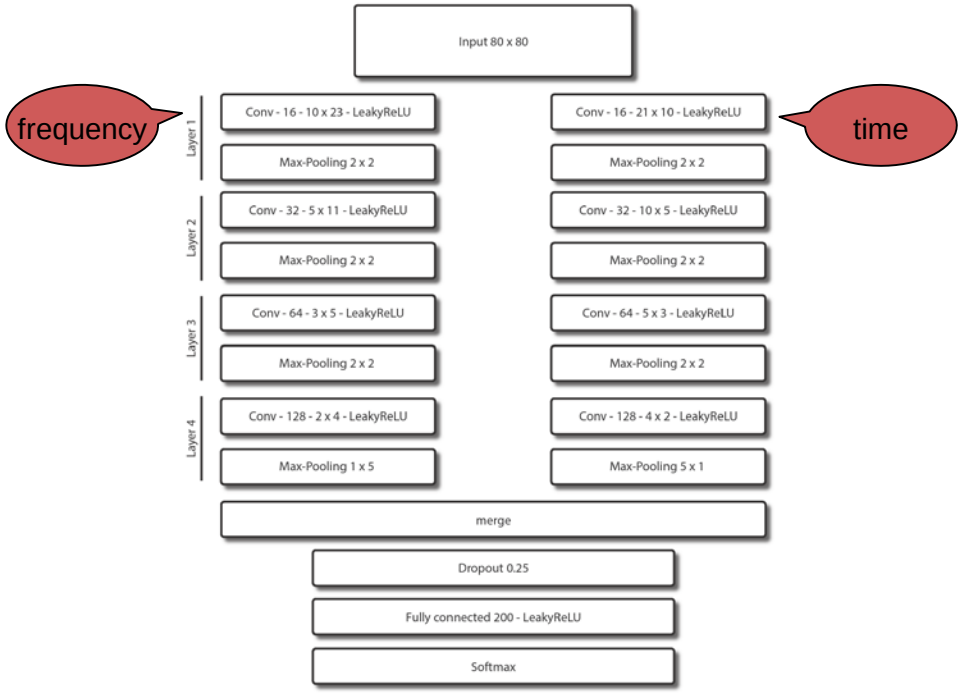
Learned Filter Weights



Convolved Spectrograms



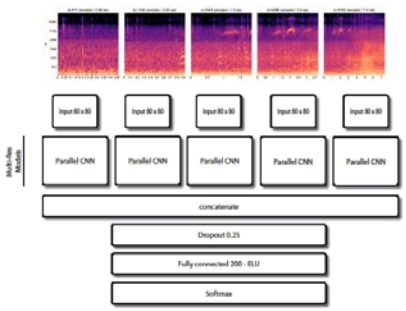
DEEP VS. SHALLOW

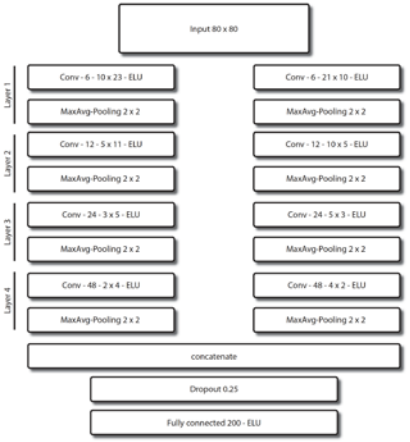


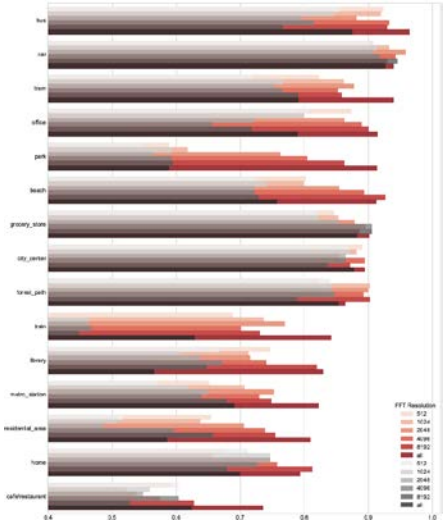
	100 epochs		200 epochs	
	Shallow	**Deep**	**Shallow**	**Deep**
GTZAN	78.1	**78.6**	**80.8**	80.6
ISMIRgenre	**85.5**	84.1	84.9	**85.1**
Latin	92.4	**94.4**	93.5	**95.1**
MSD	63.9	**67.2**	/	/

MULTI-RESOLUTION CONVOLUTIONAL NEURAL NETWORKS

- Network Architecture







fft win size	instance raw	grouped raw	instance augmented	grouped augmented
512	64.14 (2.84)	70.32 (2.96)	69.06 (4.33)	76.63 (4.44)
1024	66.32 (2.58)	71.27 (3.06)	71.70 (5.46)	77.06 (5.46)
2048	66.83 (1.52)	70.23 (1.99)	76.24 (2.53)	80.46 (3.30)
4096	69.50 (2.83)	71.92 (3.23)	79.20 (3.03)	81.66 (3.29)
8192	69.66 (2.58)	71.47 (2.95)	82.26 (2.40)	83.73 (2.63)
grouped single		73.12		83.19
multi-res	72.23 (4.15)	74.30 (4.81)	85.22 (2.11)	87.29 (2.02)
multi-res do	69.39 (2.77)	72.05 (3.26)	82.51 (2.37)	86.04 (3.03)

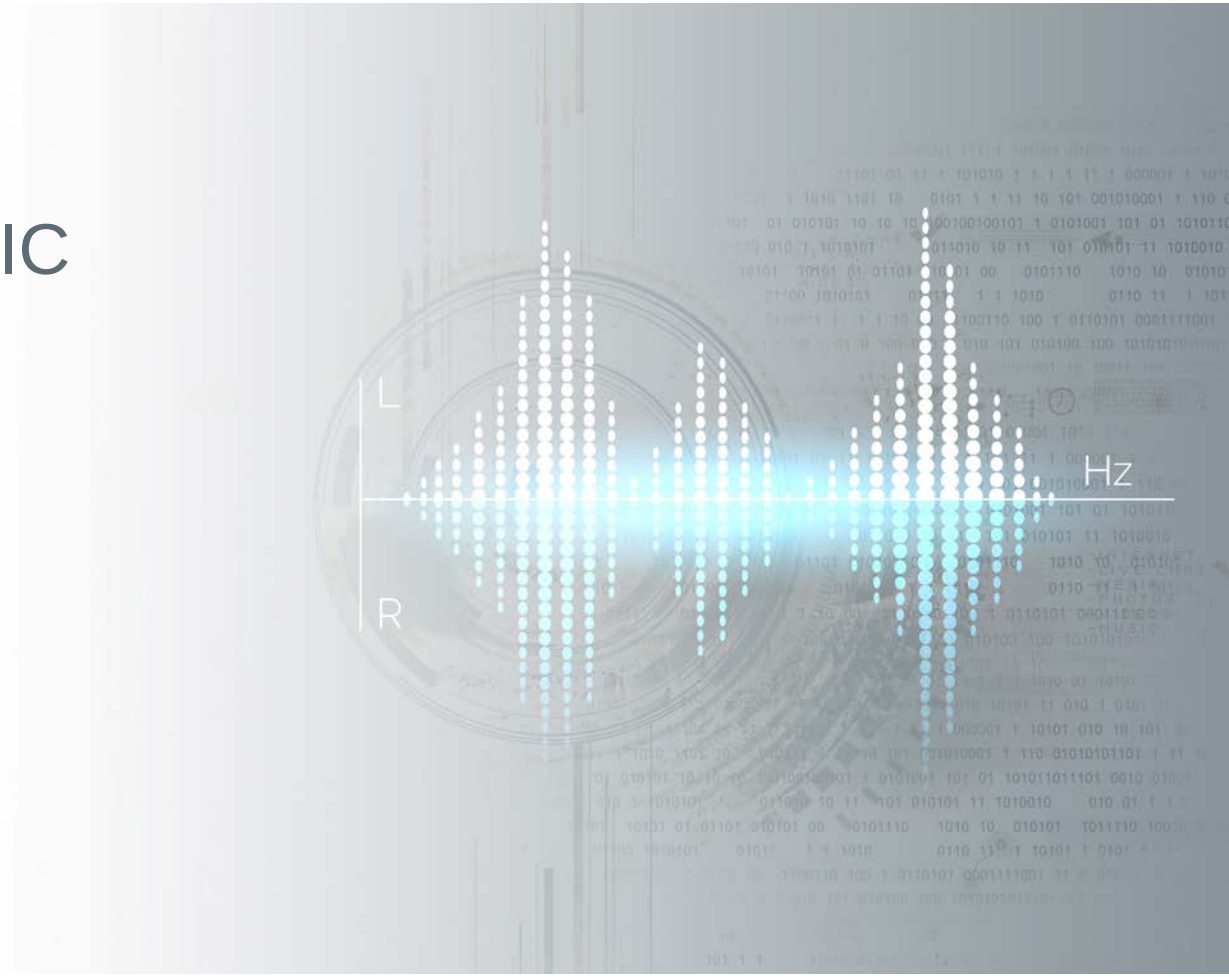
SUCCESS

- DCASE 2016
  - Task winner: Domestich Audio Tagging
- MIREX 2016
  - Classical Composer Identification
  - Latin Genre Classification
  - Music Mood Classification
  - KPOP Genre (Annotated by Korean Annotators) Classification
  - KPOP Genre (Annotated by American Annotators) Classification
  - KPOP Mood (Annotated by Korean Annotators) Classification
  - KPOP Mood (Annotated by American Annotators) Classification

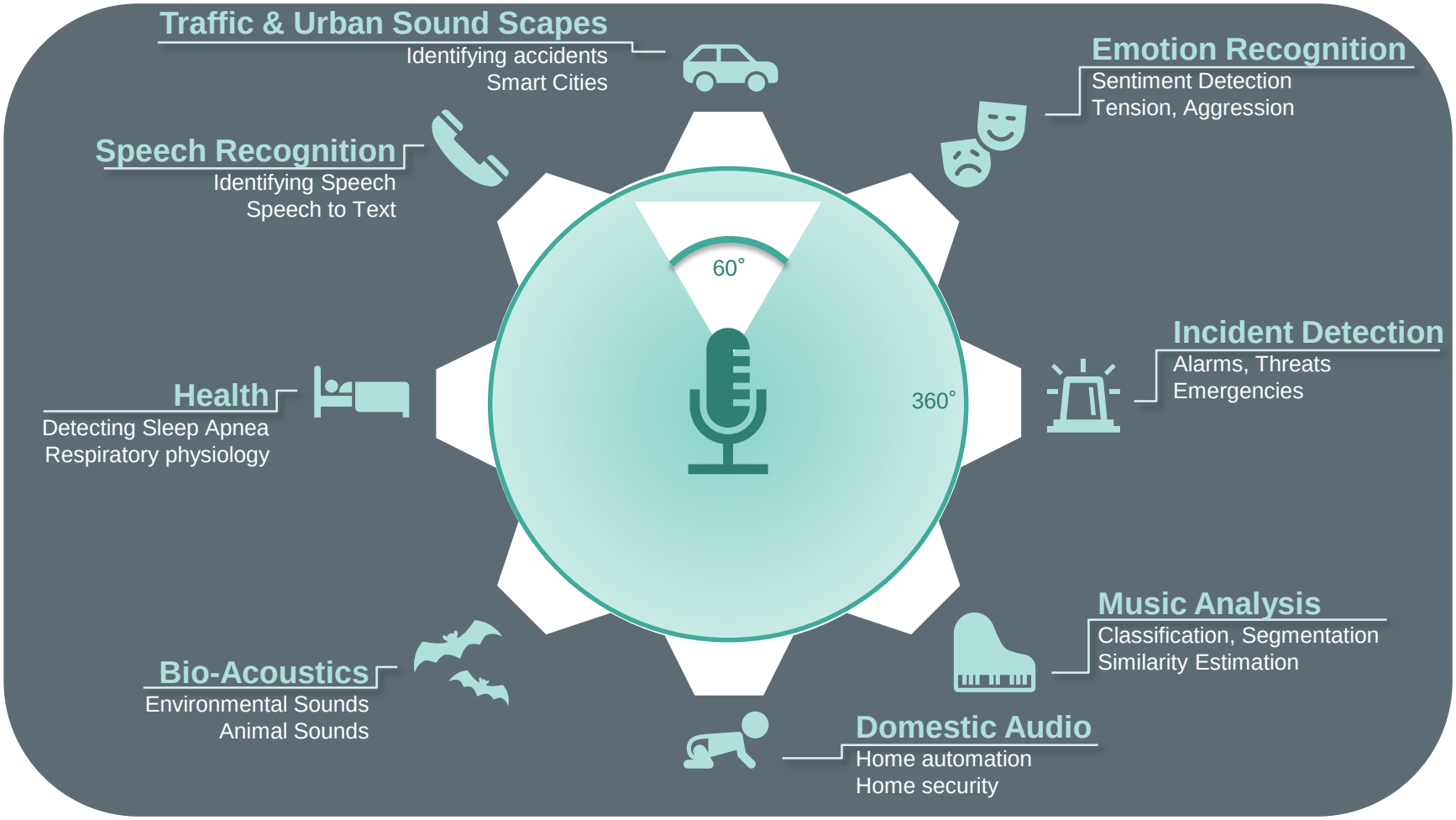
AUSTRIAN INSTITUTE OF TECHNOLOGY

AUDIO AND MUSIC RESEARCH

Tasks and Domains



AUDIO ANALYSIS



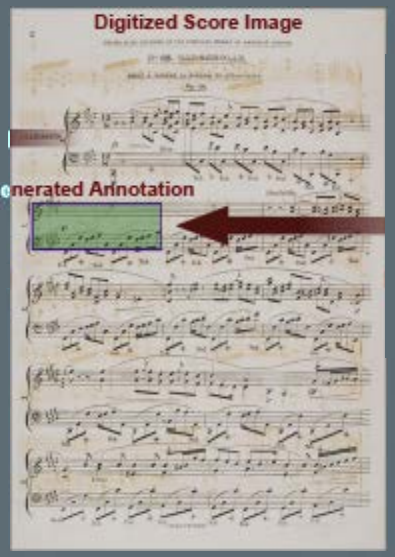
MUSIC IR – TASKS

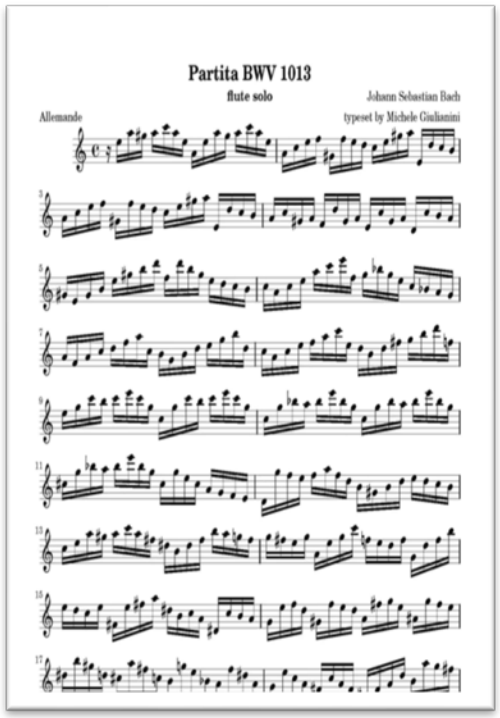
- **Genre classification**
- **Mood classification**
- **Music Recommendation**
- **Artist identification**
- Artist similarity
- Cover song detection
- **Rhythm and beat detection**
- **Score following**
- **Chord detection**
- **Organization of music**
- **Audio Fingerprinting**
- **Audio segmentation**
- Instrument detection
- Automatic source separation
- **Onset detection**
- **Optical music recognition**
- Melody transcription …..

SCORE FOLLOWING

Audio to Score Alignment

Europeana-Sounds Project
Scalable Computing





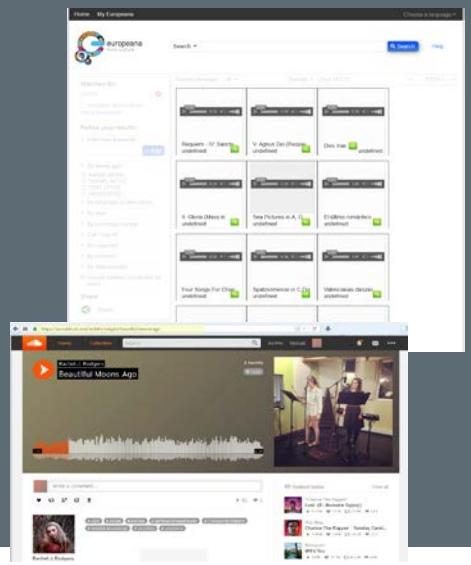
AUDIO SIMILARITY ESTIMATION

Finding Related Content

Artificial Intelligence
Multimedia-Analysis

- **Music Similarity**
  - Complex Concept
  - Example: Europeana-Sounds
    - Cultural Simlarity (Countries, Languages, Religion)
    - Historical Similarity (Epochs)
    - Sound Quality (Wax tapes, Shellacs, Digital born)

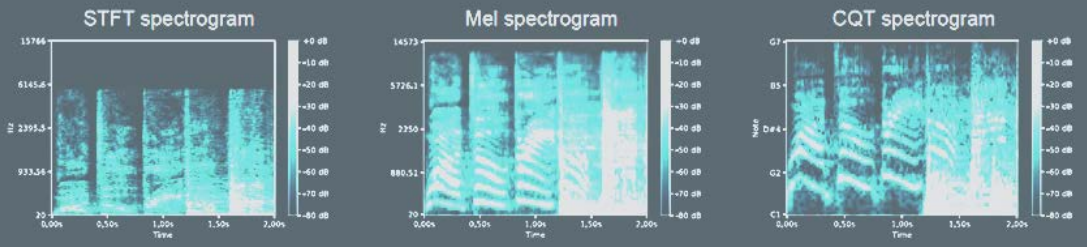
- **Sound Scene Simlarity**
  - Example: Forensic Video-Investigation
    - Finding similar video-sequences based on audio signature
    - Immananet localization / finding videos recorded nearby



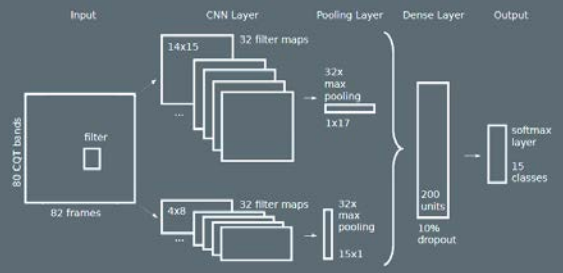
SCENE ANALYSIS

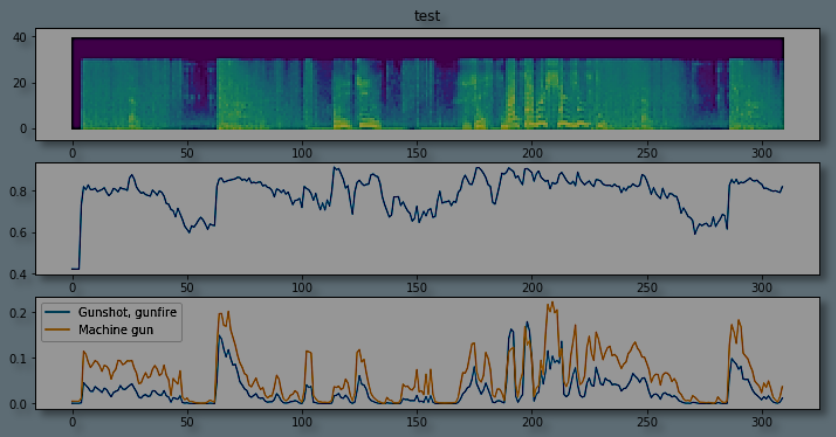
Audio-Visual Scene Understanding

Artificial Intelligence
Multi-media Analysis

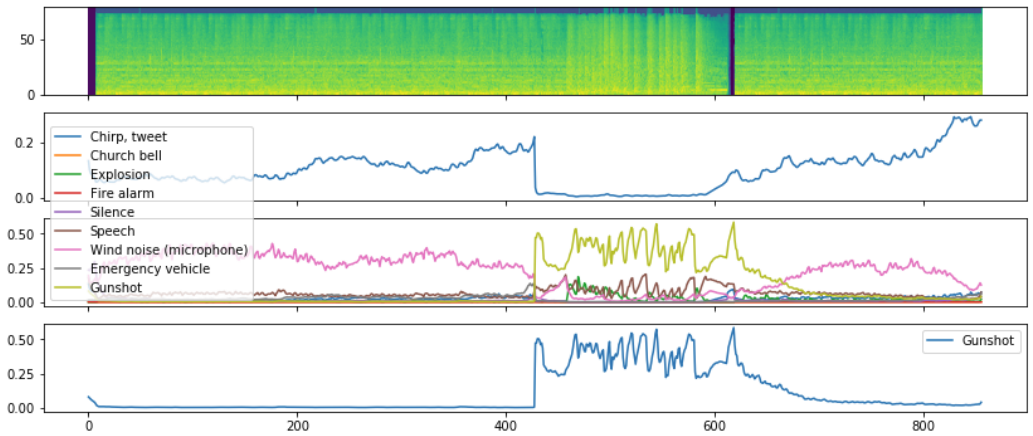


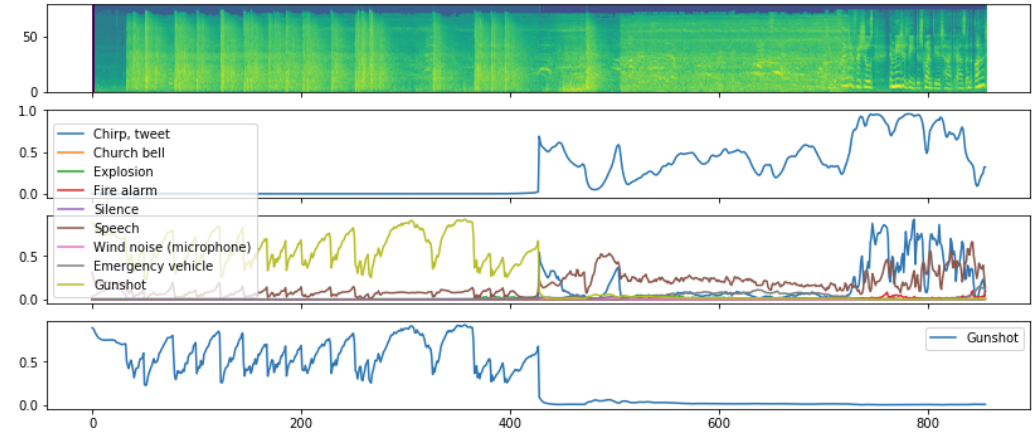
- **Acoustic Scene Classification**
  - Identify different acoustic scenes (Bus, Train, Urban Park, Bar)
  - Identify different activites (talking, walking, reading, children playing)
- **Audio Event Detection**
  - Detect acoustic events in audio-stream (onsets/length)
  - Identify detected Events (Gunshots, explosions, baby cry)
- **Audio-Visual Scene Understanding**
  - Combining acoustic with visual information
  - Improved interpretability of current scene
  - Multi-task learning





RESULTS: AUDIO EVENT DETEKTION





BIRD SONG IDENTIFICATION

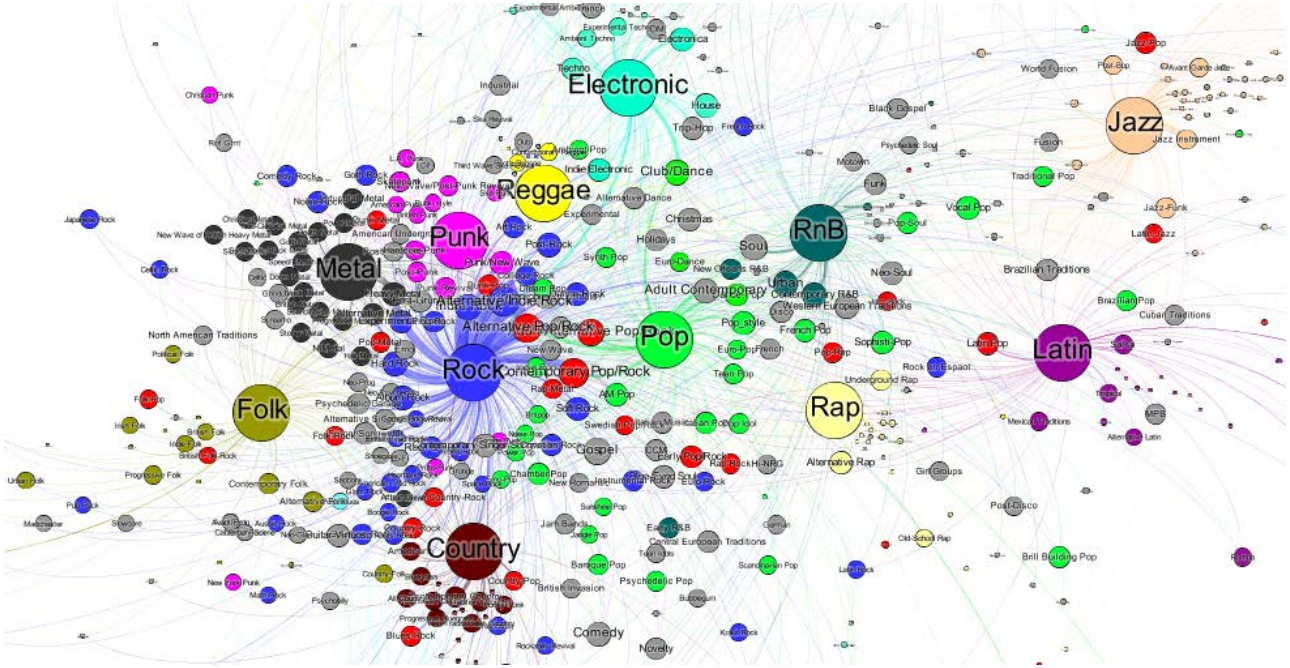
Bio-Acoustics

Artificial Intelligence
Ecology, Environmental Protection

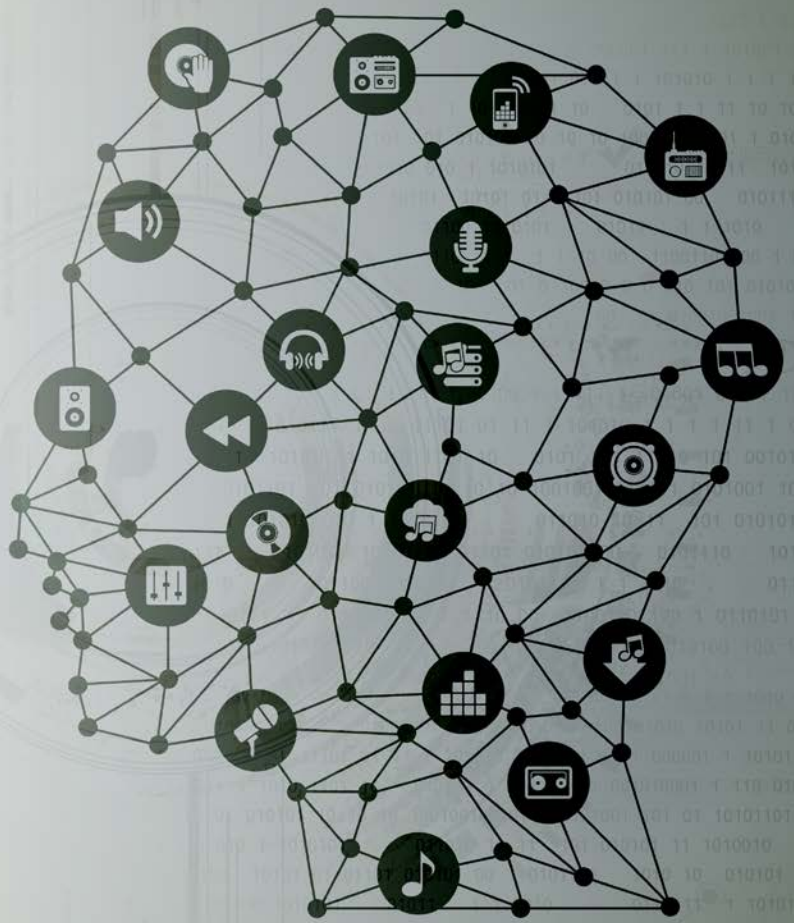
- **Large Scale Classification Problem**
  - Identify 1500 different bird species
  - Large variations in audio quantity and quality per species
- **Multi-Modal Neural Network Approaches**
  - Combining Audio-Information with Geo-Information
  - Normalizing temporal information by region (e.g. Dusk/Dawn)
- **Sophisticated Data-Augmentation**
  - Time-stretching (faster/slower)
  - Pitch-shifting (higher/lower)
  - Mixing recordings from same species
  - Mixing surrounding recordings (1° E/W/N/S Neighborhood)
  - Random cuts (cut-shuffle-merge)
  - Volume-shifting
  - Noise Overlays

CURRENT RESEARCH

- Large scale evaluation
  - Correlation between Genres/Styles/Moods
  - Network Analysis
  - Train Neural Networks more efficiently



MULTI-MODAL-ANALYSIS



MULTI-MEDIA UNDERSTANDING

Audio-Visual Video Analysis and Classification

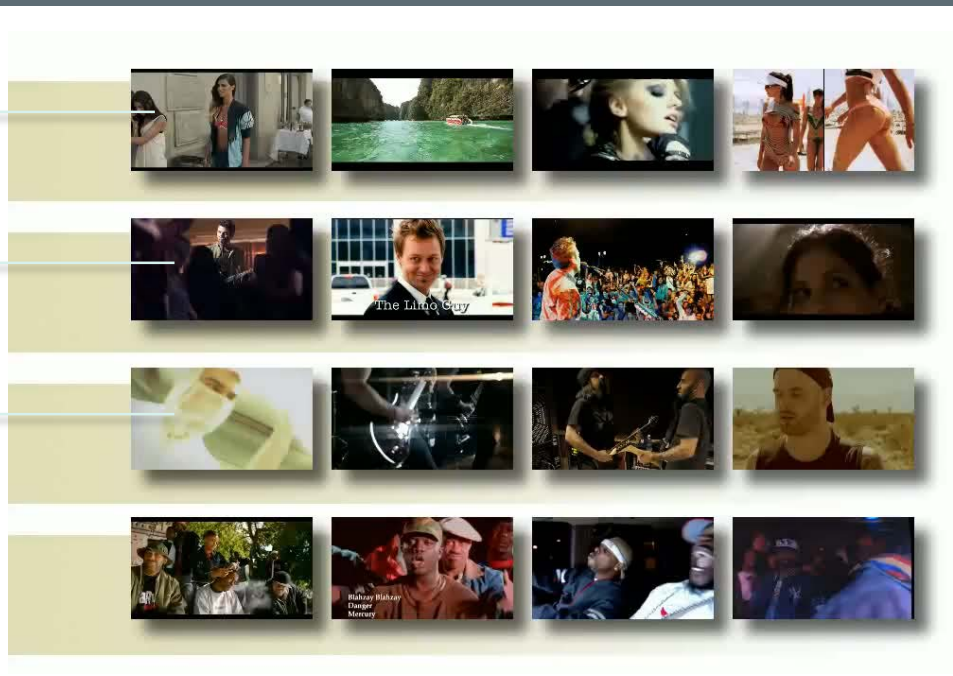
Artificial Intelligence
Multi-media / Multi-modal

Identifying Content
Detecting Objects
Detecting Events

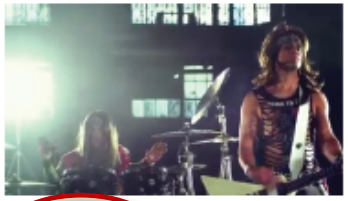
Describing Content
Visual / Acoustic characteristics
Colors, Texture, Timbre, Patterns

Analyzing Content Dependencies
Temporal distribution of detected content / objects
Knowledge modelling

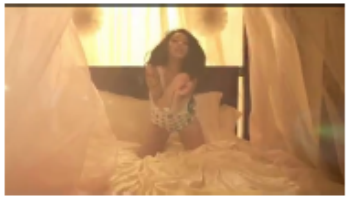
Interpreting Scenes
Action, Location, Intention
Colors, Texture, Timbre, Patterns



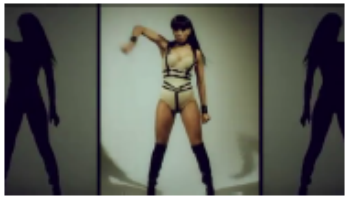
Top concepts of music video frames examples



stage	0.3162
electric guitar	0.1169
bassoon	0.0649
accordion	0.0611
drumstick	0.0386
microphone	0.0313
marimba	0.0276



mosquito net	0.0932
wardrobe	0.0857
brassiere	0.0815
shower curtain	0.0471
candle	0.0400
plastic bag	0.0204
hoopskirt	0.0187



maillot	0.2745
bolo tie	0.0732
Windsor tie	0.0550
letter opener	0.0486
brassiere	0.0390
bikini	0.0384
bassoon	0.0364

lumbermill	0.1925
tow truck	0.1215
harvester	0.1152
thresher	0.0513
jeep	0.0484
half track	0.0473
pickup truck	0.0460

wig	0.4399
neck brace	0.0577
chimpanzee	0.0418
hair spray	0.0375
orangutan	0.0366
cloak	0.0267
Windsor tie	0.0236

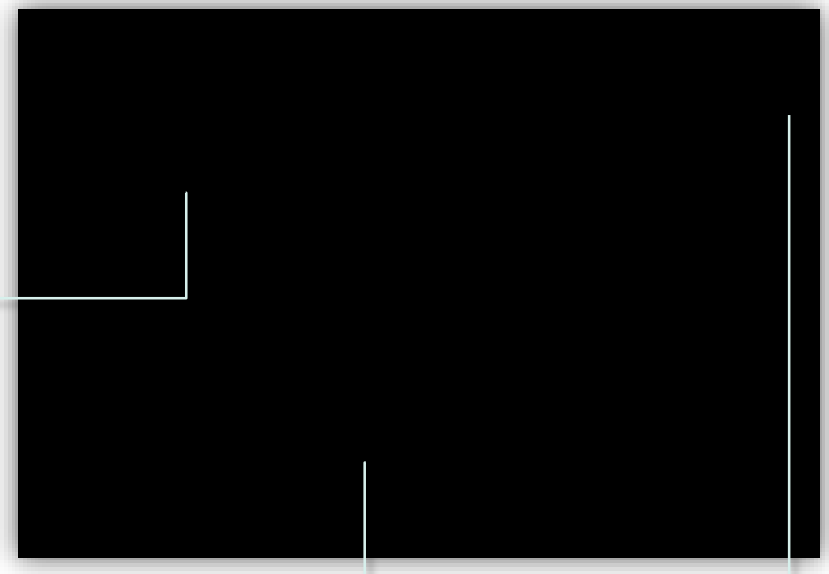
Classification results (visual concepts only)

(c) High-level Visual Concepts											
$v_{in}1$	MEAN	1000	66.86	42.09	53.69	51.26	31.23	37.05	46.87	23.90	**33.07**
$v_{in}2$	STD	1000	69.78	46.76	50.08	51.95	29.99	32.88	48.29	26.83	29.63
$v_{in}3$	MAX	1000	73.15	44.26	46.41	54.60	33.05	31.94	50.07	26.93	27.49
$v_{in}4$	$v_{in}3$+$v_{in}2$	2000	73.61	46.53	**51.21**	55.04	31.48	34.00	51.30	27.03	31.04
$v_{in}5$	$v_{in}3$+$v_{in}1$	2000	**74.36**	47.70	53.65	55.99	**33.70**	**37.83**	**51.58**	**28.88**	33.83

SENTIMENT DETECTION

Sentiment Detection

Artificial Intelligence
Multi-media / Multi-modal



Crowd scene / behaviour analysis

Anomality detection in natural scenes / crowds
Threat / escalation estimation

Audio Analysis

Audio based sentiment detection
Arousal, Pleasure, Dominance

Speaker Sentiment Detection

Sentiment of speaker
Tension, excitement, affection

Color Statistics

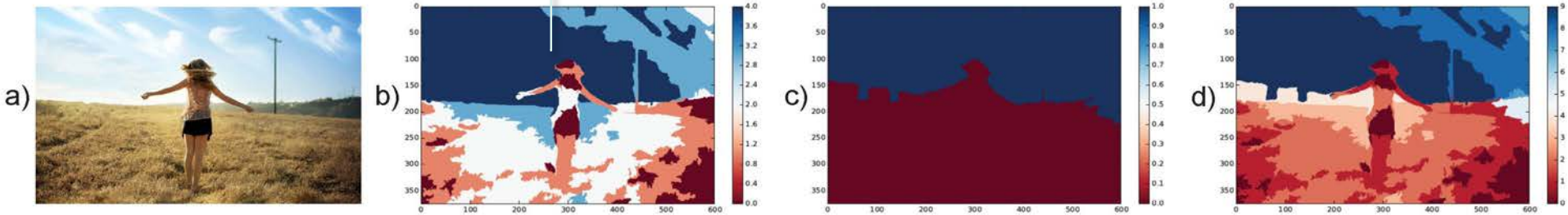
Calculating color distributions
Deriving higher level features

Affective Interpretation

Deduced from Color Information
Arousal, Pleasure, Dominance

Affective Contrasts

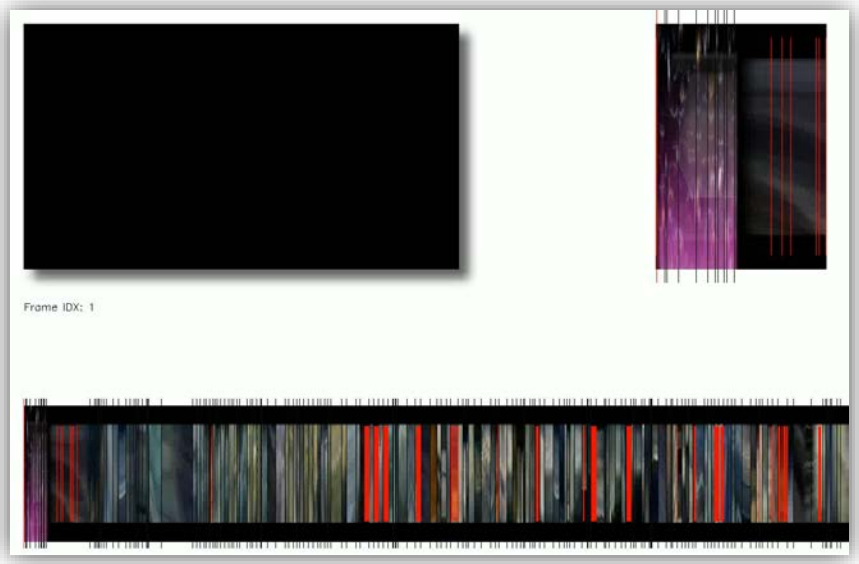
Cold / Warm
Light / Dark



VIDEO SEGMENTATION

Audio-Visual Anomaly Detection

Artificial Intelligence
Multi-media / Multi-modal



- **Anomaly Detection**
  - Identify onsets / changes in progression
  - Categorize types of anomalities
- **Segmentation**
  - Identify coherent segments in multimedia
  - Identfiy semantically labelled sections (e.g. speech, riots, music)
- **Synchronisation**
  - Time-metadata may be missing or not reliable
  - Synchronize multi-media files according their content
  - Audio-synchronization

SPEAKER / SINGER IDENTIFICATION

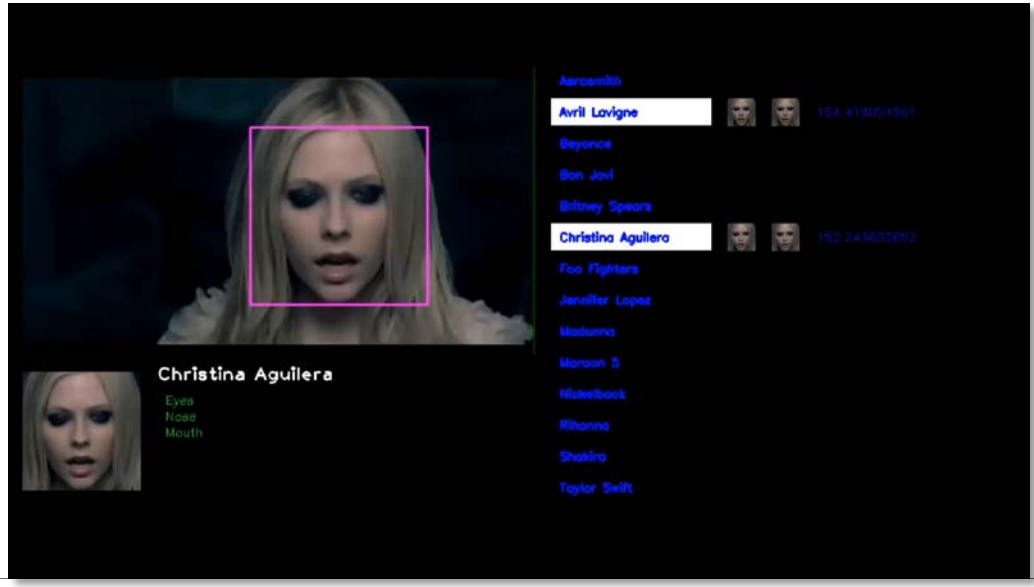
Audio-Visual Signer Identification

Artificial Intelligence
Multi-media / Multi-modal

- **Identify Speaker / Singer in Audio/Video Sequence**
  - Based on segmentation (identified vocalized segments)
  - Supervised identification (models trained on declared persons)
  - Un-supervised identification (relative similarity estimation)

- **Audio-Identification**
  - Based on acoustic models
  - Distinguis between speakers
  - Transcription of spoken words

- **Audio-Visual Identification**
  - Based on acoustic and visual models
  - More accurate
  - Visual segmentation / boxes
  - Visual identification / tracking

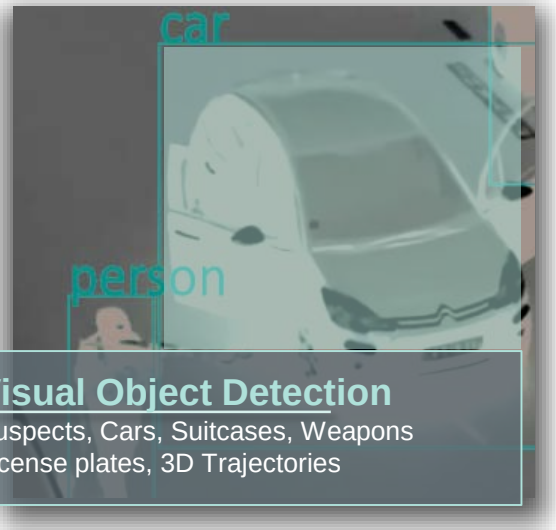


AUDIO-VISUAL SINGER IDENTIFICATION

AI SOLUTION EXAMPLE

Example: Forensic Analytics in Massive Video Content

Artificial Intelligence
Scalable Computing



Visualization
Data Aggregation & Visualization
Interactive Processing

Large Scale Computing
Distributed Computing Clusters
Cloud Computing Big Data Architectures

Visual Object Detection
Suspects, Cars, Suitcases, Weapons
License plates, 3D Trajectories

Scalable Plattform
Apache Hadoop

GPU Plattform
NVIDIA, Cuda

Audio Event Detection
Gunshots, Explosions, Screams
Spoken Words, Transcription

Massive Video Data
e.g. after an terroristic attack
CCTV, mobile phones

GPU Scale Computing
Deep Learning, Deep Neural Networks
Artificial Intelligence Modules

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

THANK YOU!