INTERNATIONAL AUDIO LABORATORIES ERLANGEN

AUDIO LABS

Lecture

Music Processing

Applications of Music Processing

Christian Dittmar
International Audio Laboratories Erlangen
christian.dittmar@audiolabs-erlangen.de

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

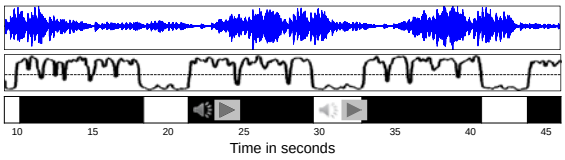
Singing Voice Detection

Important pre-requisite for:

- Music segmentation
- Music thumbnailing (preview version)
- Singing voice transcription
- Singing voice separation
- Lyrics alignment
- Lyrics recognition

Singing Voice Detection

- Detect singing voice activity during course of a recording
- Assumptions:
  - Real-world, polyphonic music recordings are analyzed
  - Singing voice performs dominant melody above accompaniment



Singing Voice Detection

- Challenges:
  - Complex characteristics of singing voice
  - Large diversity of accompaniment music
  - Accompaniment may play same melody as singing
  - Pitch-fluctuating instruments my be similar to singing

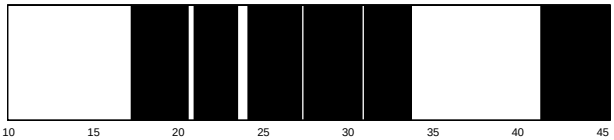
Stable pitch

Fluctuating pitch

Singing Voice Detection

Common approach:

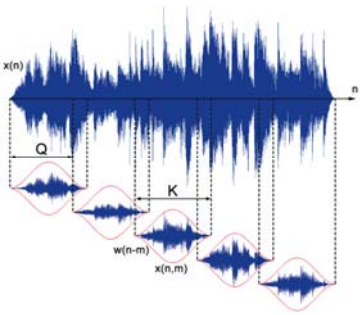
- Frame-wise extraction of audio features
- Classification via machine learning



Audio Feature Extraction

Frame-wise processing:

- Hopsize Q
- Blocksize K
- Window function w(n)
- Signal frame x(n)



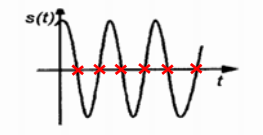
Compute for each analysis frame:

- Time-domain features
- Spectral features
- Cepstral feature
- others …

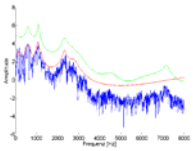
Audio Feature Extraction

Time-domain features:

- Zero Crossing Rate (ZCR)
- High-pitched vs. Low-pitched



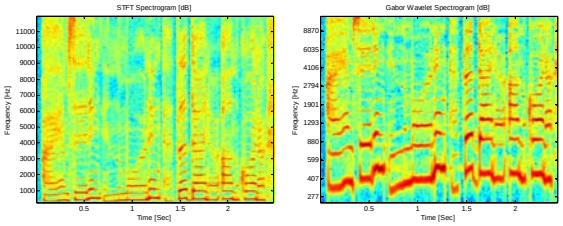
- Linear Prediction Coeff. (LPC)
- Encodes spectral envelope



Audio Feature Extraction

Spectral features:

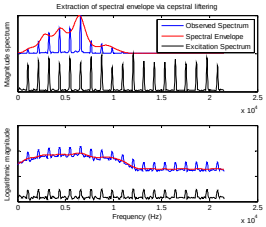
- Spectrogram, linear vs. logarithmic frequency spacing
- Spectral Flatness (SF), Spectral Centroid (SC), and many others …



Audio Feature Extraction

Cepstral features:

- Singing voice as an example
  - Convolutive: **excitation * filter**
  - Excitation: vibration of vocal folds
  - Filter: resonance of the vocal tract
- Magnitude spectrum
  - Multiplicative: **excitation · filter**
- Log-magnitude spectrum
  - Additive: **excitation + filter**
- "Liftering"
  - Separation into smooth spectral envelope and fine-structured excitation



Machine Learning

Application to audio signals:

- Speech recognition
- Speaker recognition
- Singing voice detection
- Genre classification
- Instrument recognition
- Chord recognition
- etc …

Machine Learning

Learning principles:

- Unsupervised learning
  - Find structures in data
- Supervised learning
  - Human observer provides „ground truth"
- Semi-supervised learning
  - Combination of above principles
- Reinforcement learning
  - Feedback of „confident" classifications to the training

The Feature Space

Geometric and algebraic interpretation of ML problems

- Features contain numerical values
  - Concatenation of several features
  - Dimensionality M
- The data set contains N observations
  - Cardinality N
- Illustrative Example → SFM & SCF of 6 complex tones

$$\mathrm{SF} = \frac{\sqrt[K]{\prod_{k=0}^{K-1} s(k)}}{\frac{1}{K}\sum_{k=0}^{K-1} s(k)} \qquad \mathrm{SC} = \frac{\sum_{k=0}^{K-1} f(k)\cdot s(k)}{\sum_{k=0}^{K-1} s(k)}$$

The Feature Space

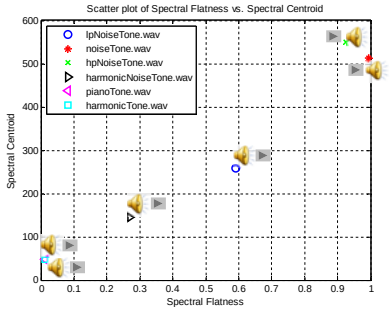
- Each feature has one value → M=2
- Number of observations → N=6

	Spectral Centroid	Spectral Flatness
lpNoiseTone.wav	258.62	0.59
noiseTone.wav	512.73	0.99
hpNoiseTone.wav	550.13	0.92
harmonicNoise.wav	146.50	0.27
pianoTone.wav	47.93	0.01
harmonicTone.wav	43.95	0.01

M (Spectral Centroid, Spectral Flatness) × N (lpNoiseTone.wav … harmonicTone.wav)

The Feature Space

- Each feature has one value → M=2
- Number of observations → N=6
- Mapping of features
  - SC to y-axis
  - SF to x-axis
  - Scatter plot with unnormalized axes



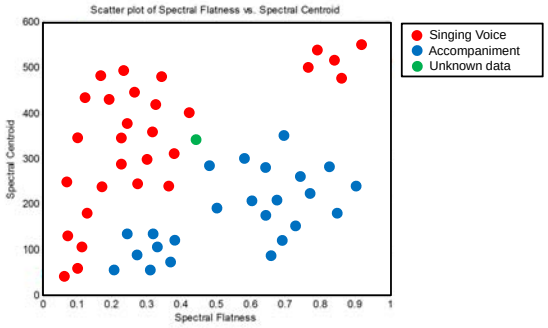
The Feature Space

- Each feature has one value → M=2
- Number of observations → N=6
- Mapping of features
  - SC to y-axis
  - SF to x-axis
  - Scatter plot with unnormalized axes
- Target class labels
  - Provided by manual annotation

	Target Labels		Spectral Centroid	Spectral Flatness
t_1	0	$\mathbf{x}_1$	258.62	0.59
	0		512.73	0.99
	0		550.13	0.92
⋮	1	⋮	146.50	0.27
	1		47.93	0.01
t_N	1	$\mathbf{x}_N$	43.95	0.01

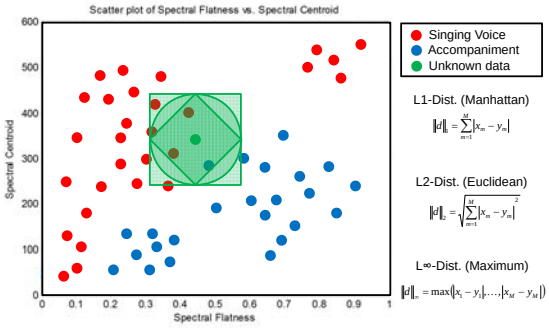
Classification methods

k-Nearest Neighbours (kNN)



Classification methods

k-Nearest Neighbours (kNN)



L1-Dist. (Manhattan)

$$\|d\|_1 = \sum_{m=1}^{M} |x_m - y_m|$$

L2-Dist. (Euclidean)

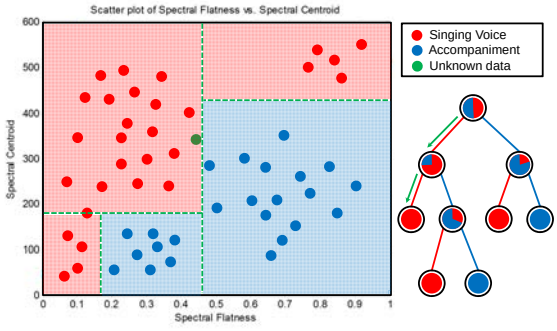
$$\|d\|_2 = \sqrt{\sum_{m=1}^{M} |x_m - y_m|^2}$$

L∞-Dist. (Maximum)

$$\|d\|_\infty = \max\left(|x_1 - y_1|, \ldots, |x_M - y_M|\right)$$

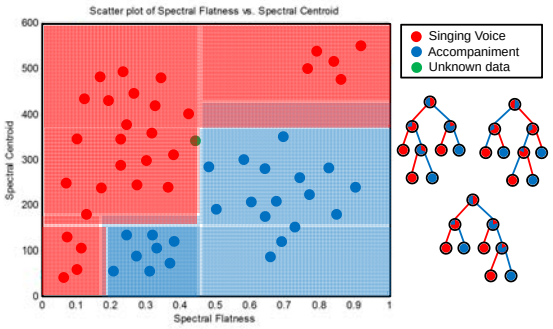
Classification methods

Decision Trees (DT)



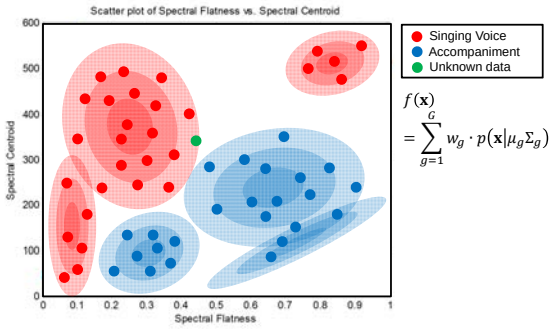
Classification methods

Random Forests (RF)



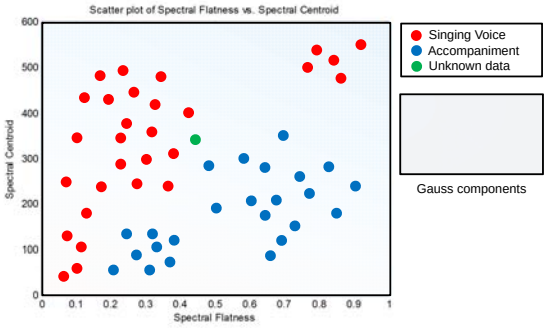
Classification methods

Gaussian Mixture Models (GMM)



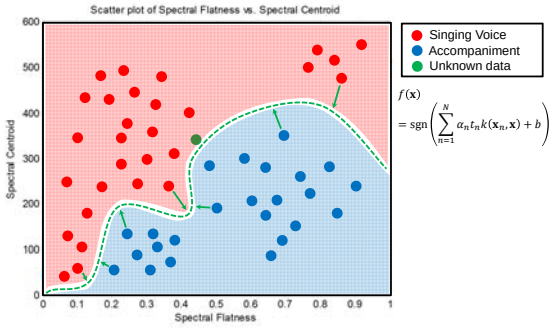
Classification methods

Gaussian Mixture Models (GMM)



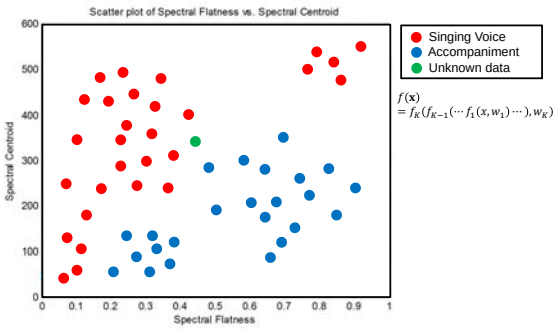
Classification methods

Support Vector Machines (SVM)



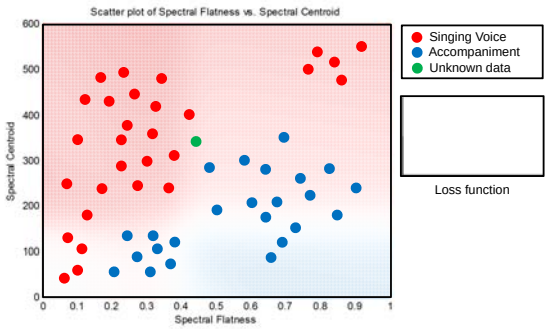
Classification methods

Deep Neural Networks (DNN)



Classification methods

Deep Neural Networks (DNN)

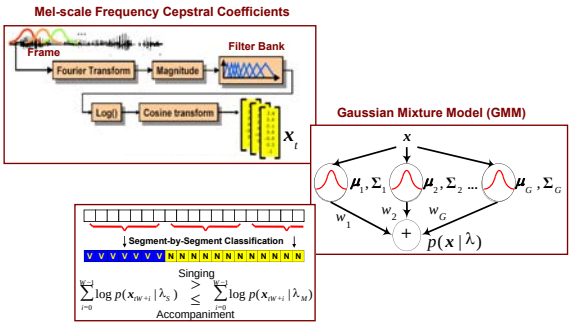


Classification methods

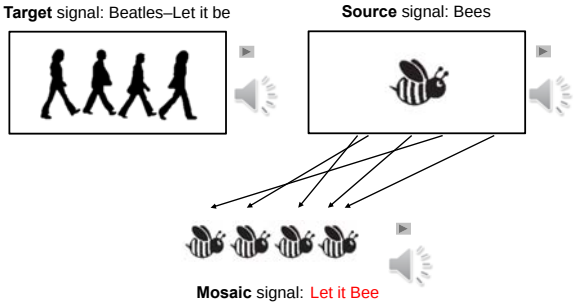
Further methods:

- Hidden Markov Models
  - Transition probabilities between GMMs
- Sparse Representation Classifier
  - Sparse linear combination of training data
- Boosting
  - Combine many weak classifiers
- Convolutional Neural Networks
- Recurrent Neural Networks
- Multiple Kernel Learning
- others …

Singing Voice Detection

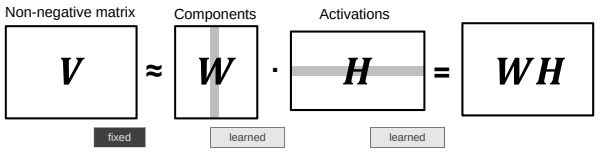


Audio Mosaicing

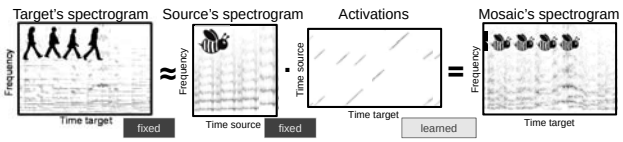


NMF-Inspired Audio Mosaicing

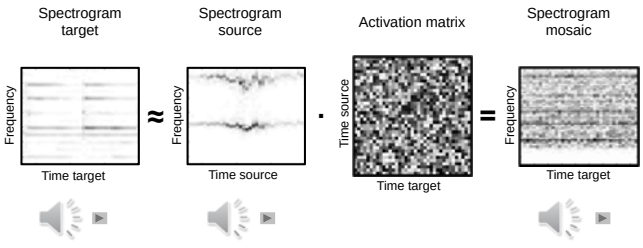
Non-negative matrix factorization (NMF) [Driedger et al. ISMIR 2015]



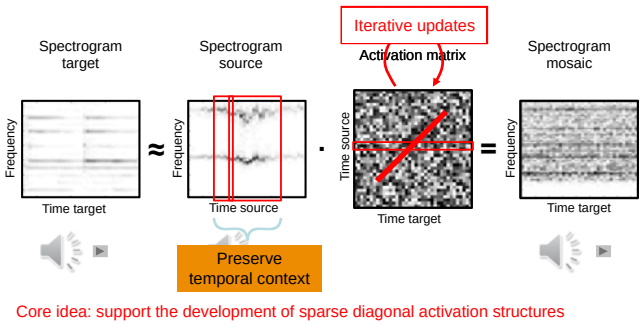
Proposed audio mosaicing approach



Basic NMF-Inspired Audio Mosaicing

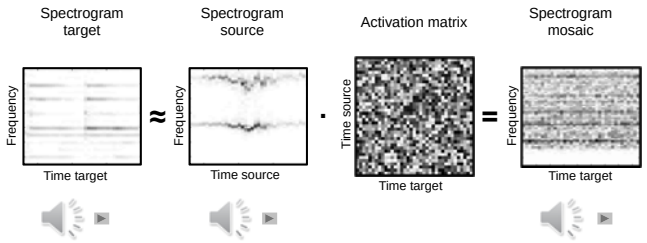


Basic NMF-Inspired Audio Mosaicing

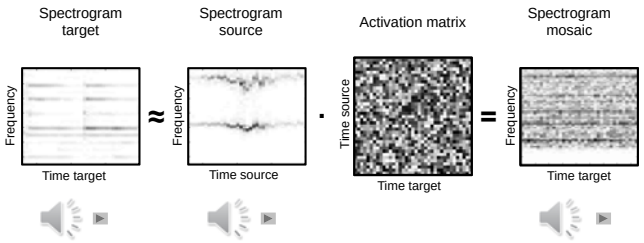


Core idea: support the development of sparse diagonal activation structures

Basic NMF-Inspired Audio Mosaicing



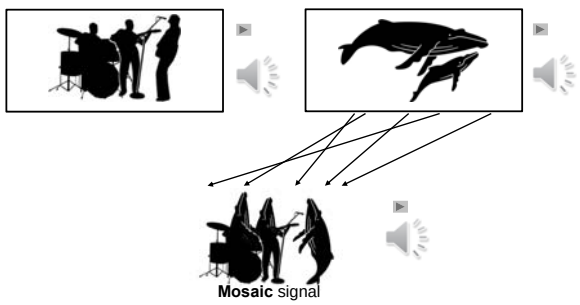
Basic NMF-Inspired Audio Mosaicing



Audio Mosaicing

Target signal: Chic–Good times

Source signal: Whales



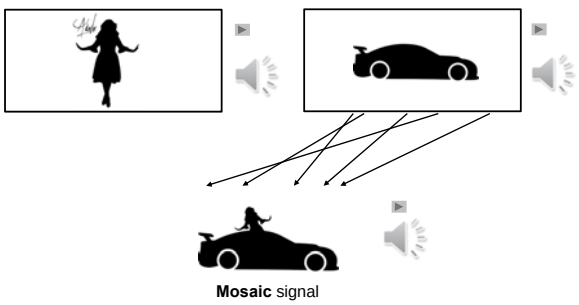
Mosaic signal

https://www.audiolabs-erlangen.de/resources/MIR/2015-ISMIR-LetItBee

Audio Mosaicing

Target signal: Adele–Rolling in the Deep

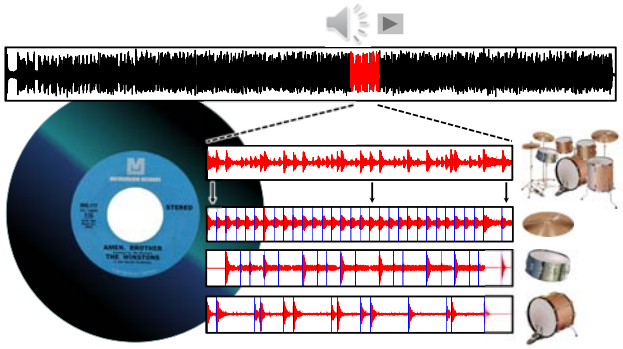
Source signal: Race car



Mosaic signal

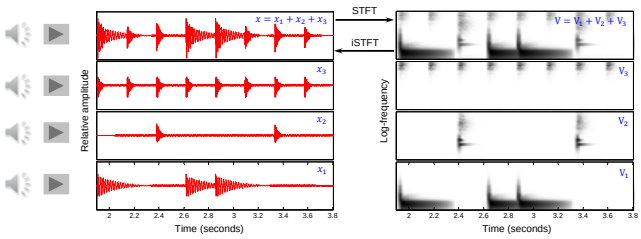
https://www.audiolabs-erlangen.de/resources/MIR/2015-ISMIR-LetItBee

Drum Source Separation



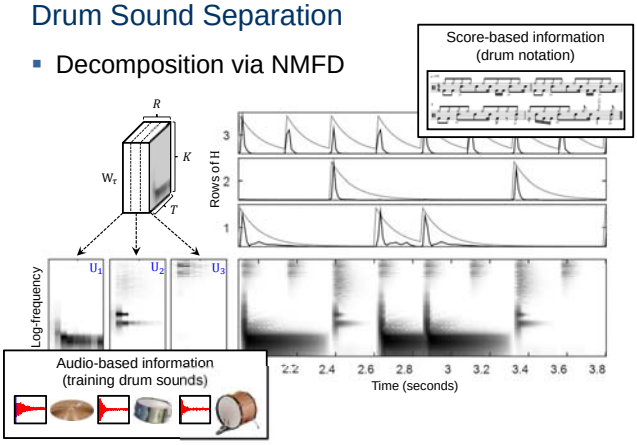
Drum Source Separation

- Signal Model

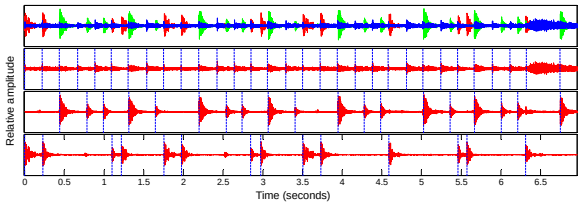


Drum Sound Separation

- Decomposition via NMFD



Drum Sound Separation



https://www.audiolabs-erlangen.de/resources/MIR/2016-IEEE-TASLP-DrumSeparation