Machine Listening for Music and Sound Analysis

Lecture 4 – Music Information Retrieval II

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

https://www.machinelistening.de

Fraunhofer

IDMT

Overview

- Pitch Detection

- Instrument Recognition

- Source Separation

Fraunhofer

IDMT

Pitch Detection
Introduction

- Pitch

    - (Subjective) psychoacoustic attribute of sound

    - Allows ordering from low to high in a frequency-related scale

        - Pitch ≠ frequency **!**

- Two subtasks

FMP Notebooks

1) Pitch detection
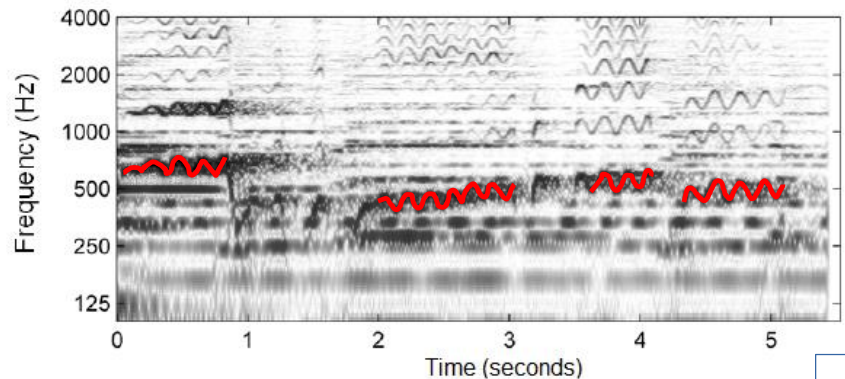
2) Voicing detection

Melody

No melody

Fig. 1

Own

Fraunhofer
IDMT

Pitch Detection
Application Scenarios

- Music Instrument Tuning

- Music Education

- Music Transcription

- Bird Recognition

Fig. 2

Fig. 3

Fig. 4

Pitch Detection
Tasks

- Pitch detection of isolated monophonic instruments

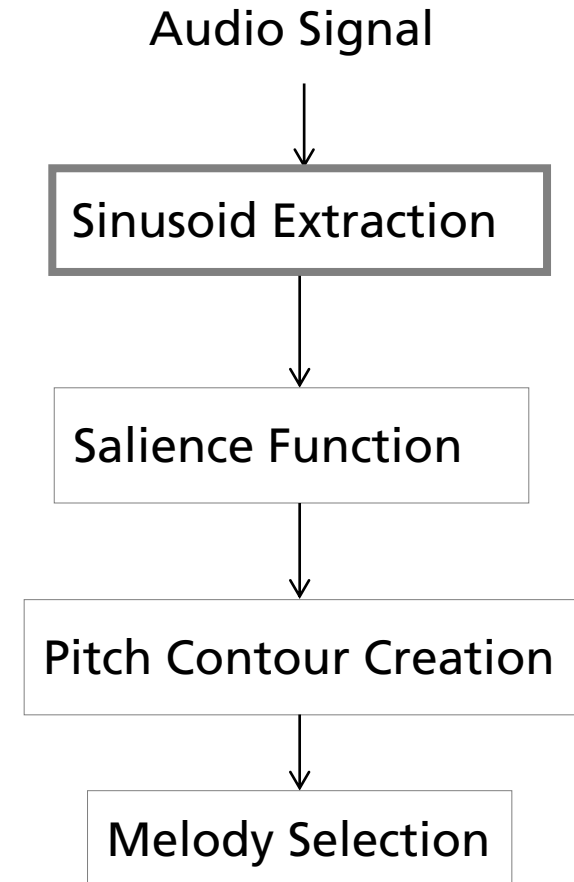- Predominant melody extraction in polyphonic music

- Polyphonic melody extraction

Increasing Difficulty & Complexity

Pitch Detection
Traditional Methods

- MELODIA [Salamon & Gomez, 2012]

  - Melody Extraction from polyphonic audio

- Steps

  - Sinusoid Extraction

    - Equal loudness filter

    - STFT

    - Detection of predominant peaks

    - Frequency refinement via instantaneous frequency (IF)

Audio Signal

↓

Sinusoid Extraction

↓

Salience Function

↓

Pitch Contour Creation

↓

Melody Selection

Fraunhofer
IDMT

Pitch Detection
Traditional Methods

- Salience Function
    - Harmonic summation
        - Sum over possibile harmonic frequencies
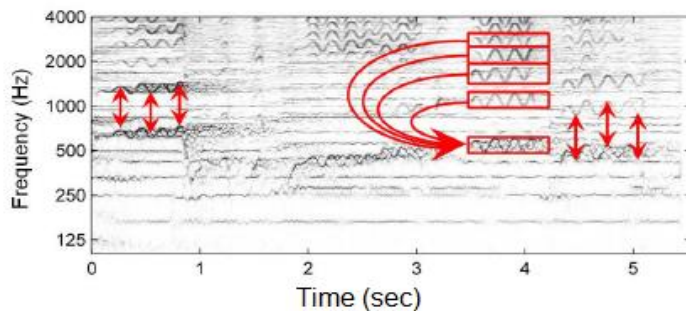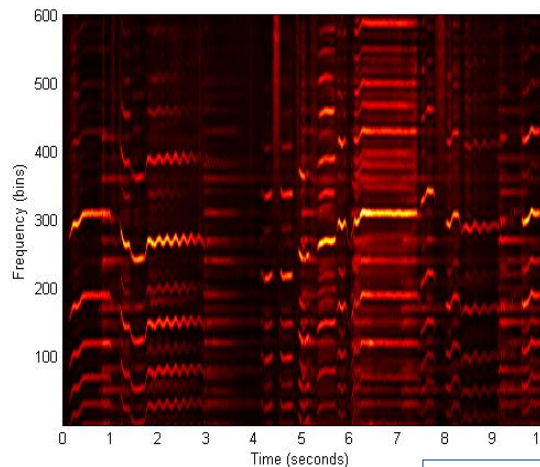   - Frequencies → pitch candidates

Fig. 5

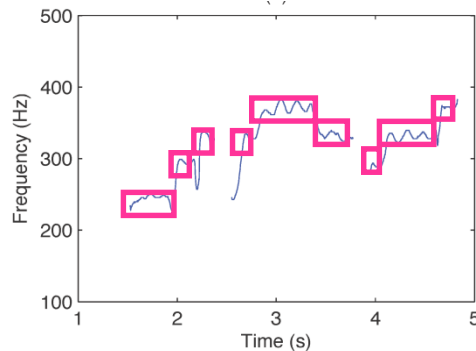Own

Audio Signal

↓

Sinusoid Extraction

↓

Salience Function

↓

Pitch Contour Creation

↓

Melody Selection

Fraunhofer
IDMT

Pitch Detection
Traditional Methods

- Pitch contour creation & melody selection

  - Auditory streaming cues → group peaks to continuous paths (pitch contours)

  - Select melody contours using features (e.g. average pitch / salience, vibrato)

  - Note formation (one pitch value)

Pitch contour(s)

↓

Note events

Own

Audio Signal

↓

Sinusoid Extraction

↓

Salience Function

↓

Pitch Contour Creation

↓

Melody Selection
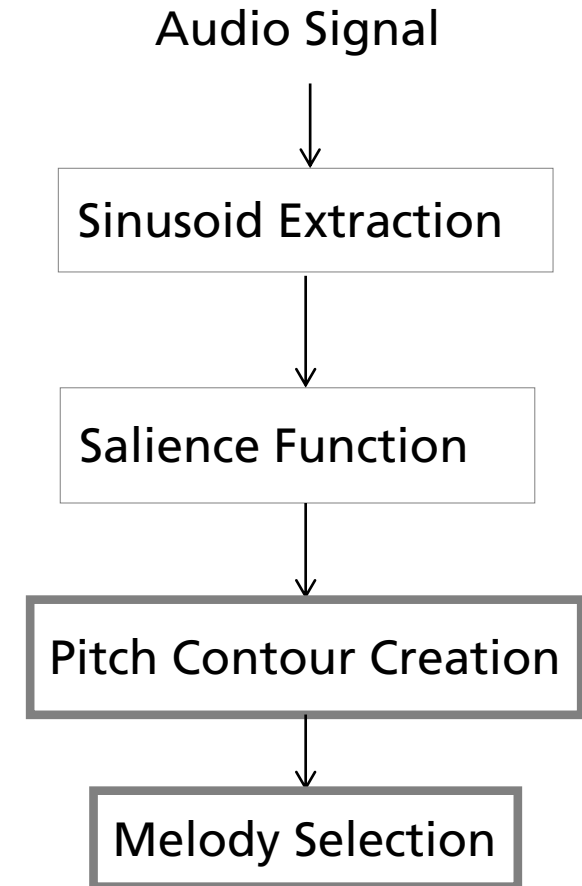
Pitch Detection
Traditional Methods (Melodia)

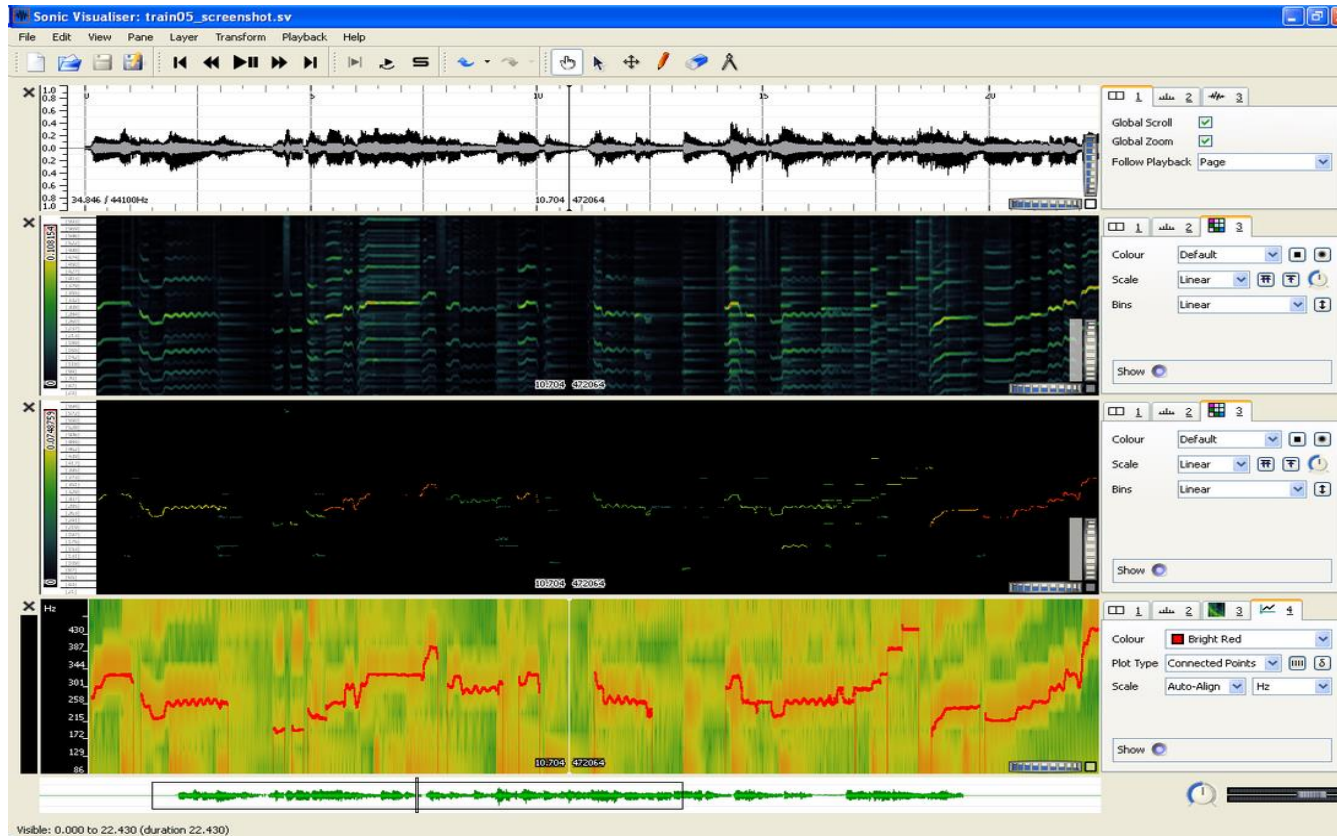- Melodia plugin available for Sonic Visualiser

Fig. 6

Pitch Detection
Novel Methods

- CREPE (Convolutional Representation for Pitch Estimation) [Kim et al., 2018]
    - Monophonic pitch tracker
    - End-to-end modeling
        - Audio samples → pitch likelihoods
        - 20 cent resolution (5 pitch bins per semitones)
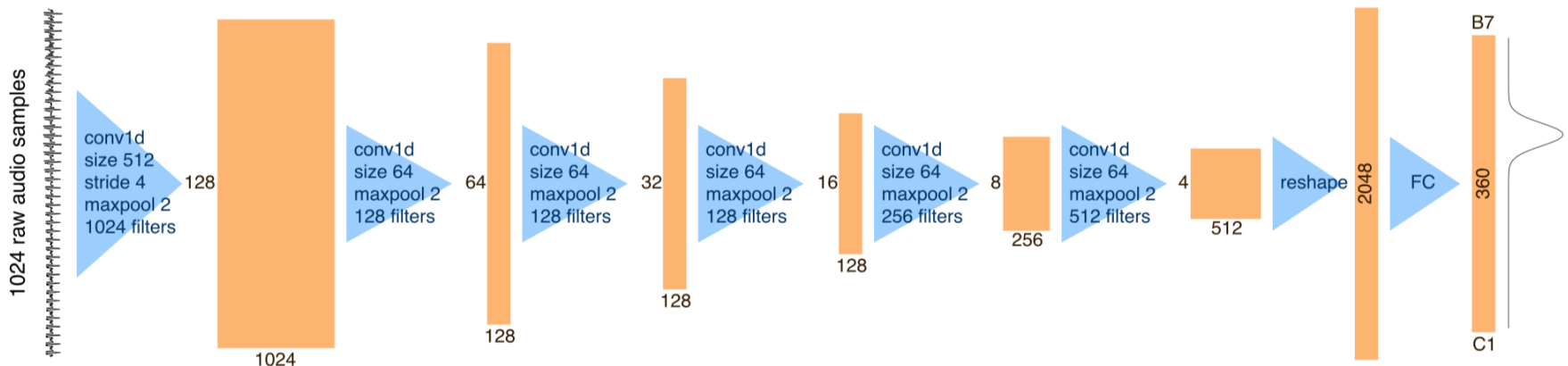

Fig. 7

Pitch Detection
Novel Methods

- Auto-encoder structure (U-Net) [Hsieh et al., 2019]

  - Mapping: multiple time-frequency representations (2D) → pitch saliency map (2D)

  - (Bottleneck) embedding encodes pitch voicing (melody activity)

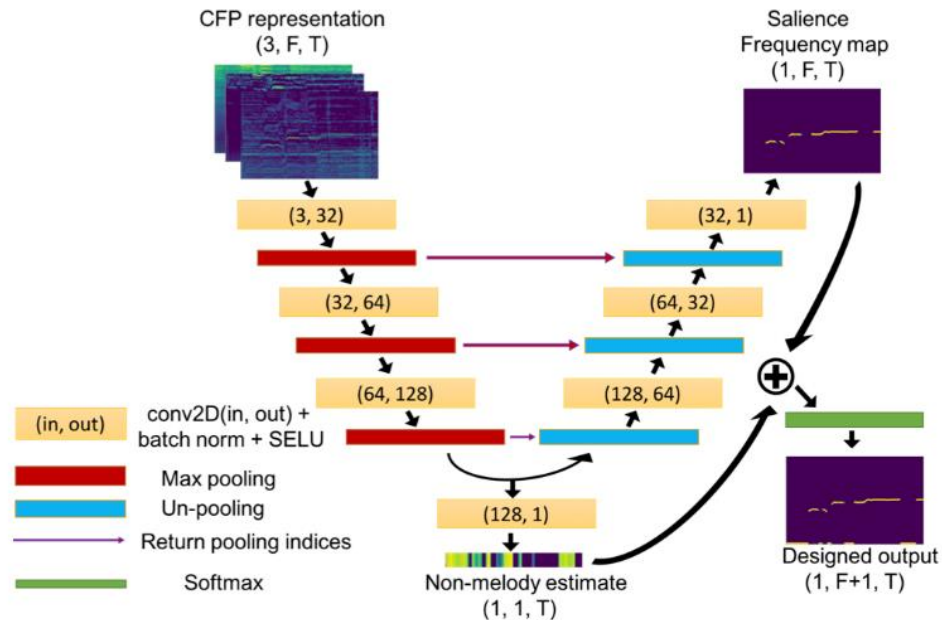
Fig. 8

Instrument Recognition
Introduction

- Music ensembles include multiple instruments

    - Sound production (string / wind / brass / drum instruments)

    - Instrument construction

- Overlapping sound sources (solo recording vs. orchestra)

    - Unison (same pitch)

    - Harmonic intervals (overtone overlap)

    - Rhythmic interconnection (note attacks overlap)

- Classification on different taxonomy levels

    - Woodwind instruments → saxophone → tenor saxophone

Instrument Recognition
Tasks

- Sorted by increasing complexity/difficulty

  - Instrument recognition of isolated note recordings

  - Instrument recognition on isolated instrument tracks

  - Predominant instrument recognition in ensemble recordings

  - Polyphonic instrument recognition (classify all instruments)

Fraunhofer
IDMT

Instrument Recognition
Traditional Methods

- Multiple categories of audio features [Grasis et al., 2014]

    - Frame-level (e.g., spectral flux & flatness)

    - Overtone-level (e.g., modulation rate & frequency)

    - Note-event level (e.g., magnitude ratios of overtones)

- Examples (trumpet / piano)

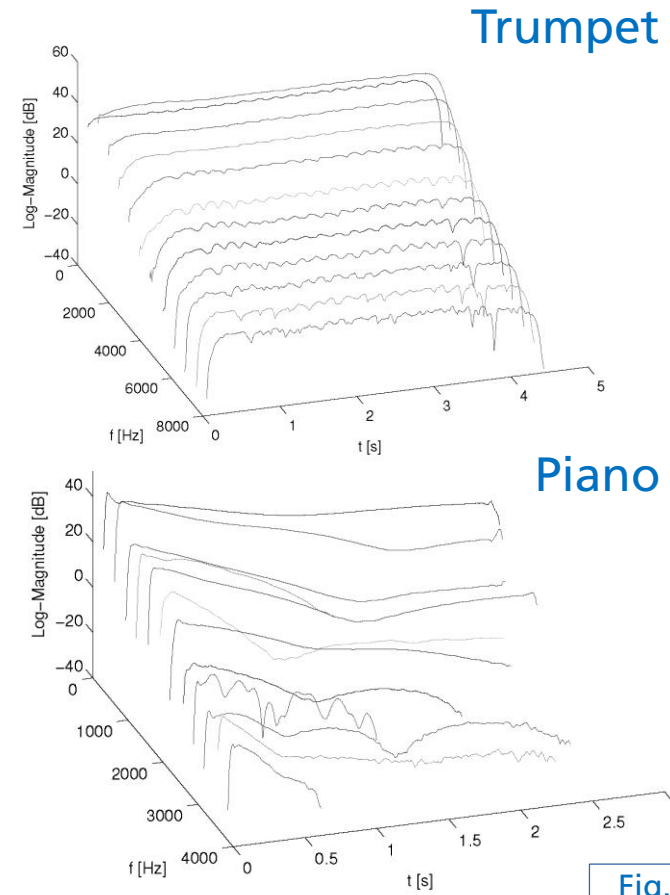   - Partial envelops

    - Observe magnitude decay & modulation

Trumpet

Piano

Fig. 9

Fraunhofer
IDMT

Instrument Recognition
Novel Methods

- Mel spectrogram + CNN model [Han et al., 2017]
    - Mel Spectrogram as input
    - Multiple convolutional layers & pooling operations
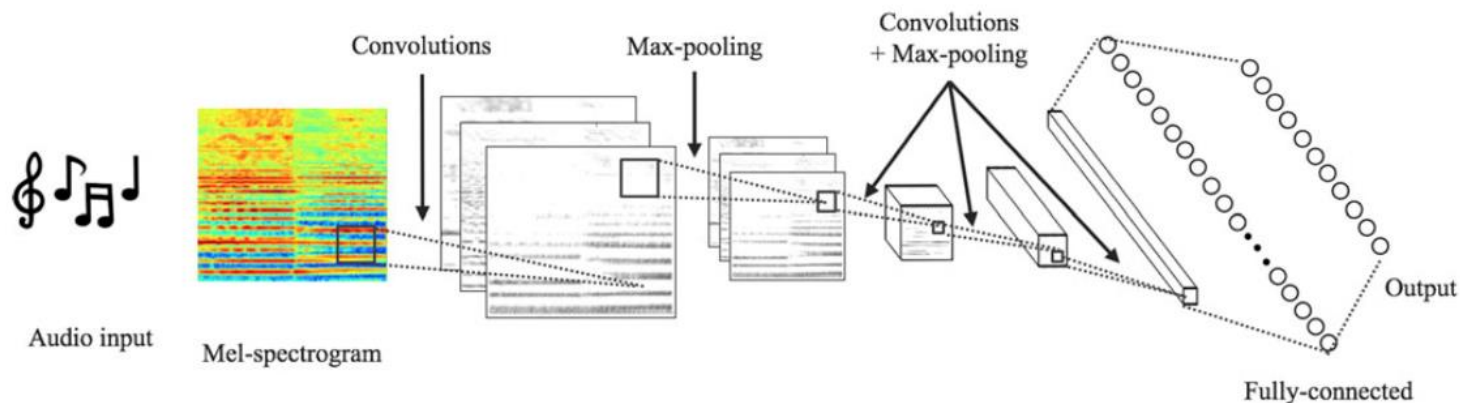   - Final (dense) classification layers

Fig. 10

Instrument Recognition
Novel Methods

- Separability of instrument classes in the feature space

    - Improves for deeper layers

- Example

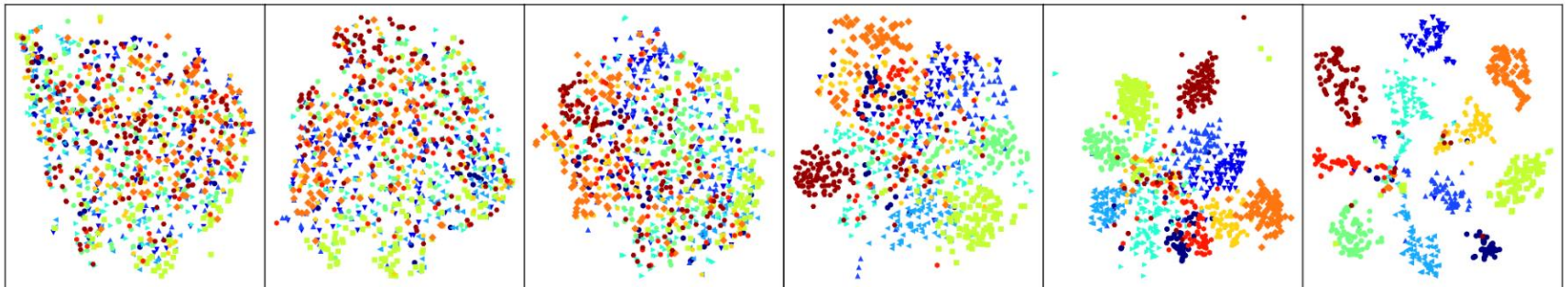   - 2D visualization of multi-dimensional feature space

Fig. 11

Deeper layers

Fraunhofer
IDMT

Instrument Recognition
Novel Methods

- Pitch-Informed Frame-level Instrument Recognition [Hung & Yang, 2018]

  - Combine two input branches

    - Spectral input features (CQT)
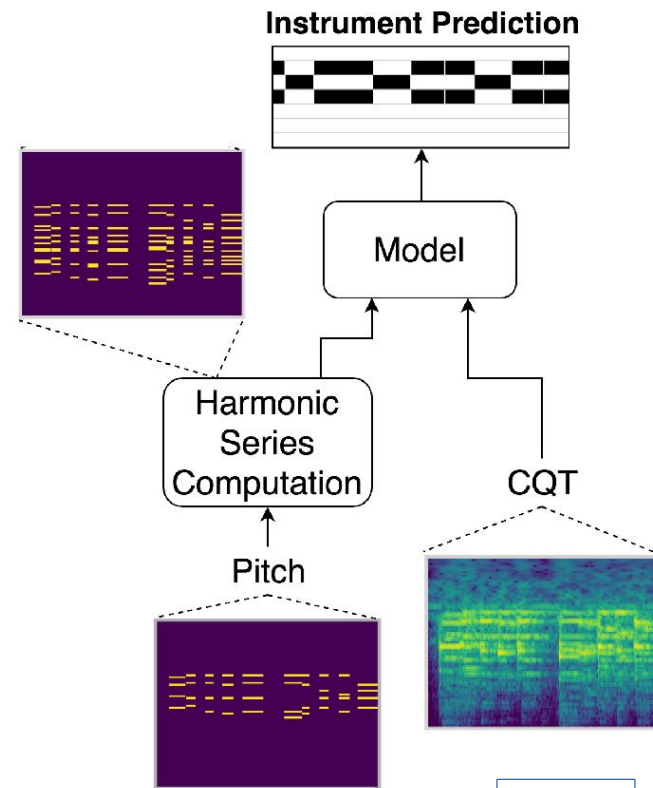
    - Pitch-activity (piano-roll)

Fig. 12

Source Separation
Introduction

- Music recordings

  - Mixtures of different musical instruments (sources) playing simultaneously

- Sound Separation

  - Reverse engineering the audio mixing process
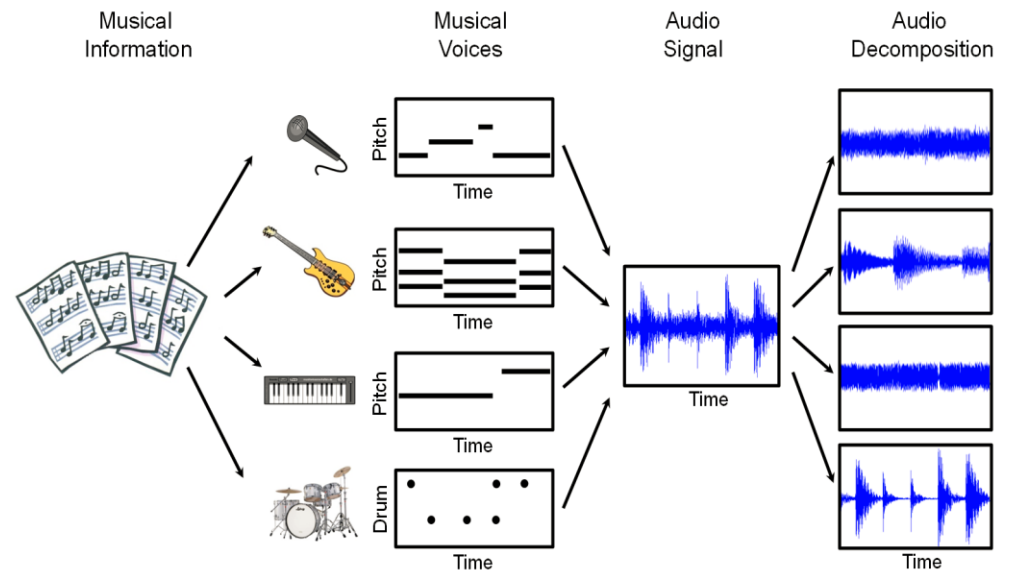
  - Output: 1 stem per instrument

Fig. 18

Source Separation
Introduction

- Audio mix is influenced by

  - Instrument characteristics (timbre, note decay, …)

  - Musical performance (timing, dynamics, playing techniques, …)

  - Recording chain (microphones, room acoustics)

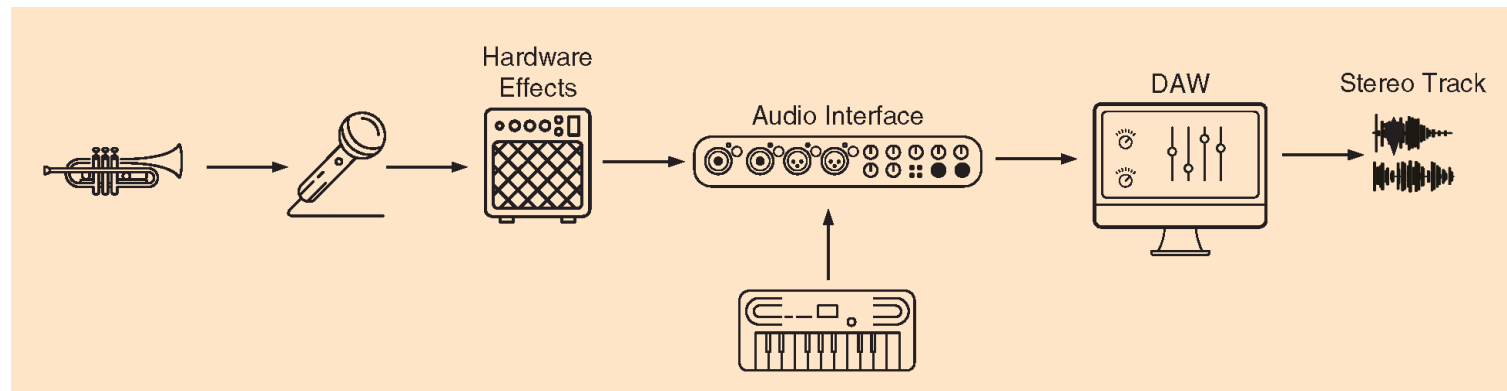 - Post-processing (effects, mastering, DAW mix)

Fig. 13

Source Separation
Application Scenarios

- Audio remixing

- Audio upmixing
    - Mono → stereo
    - Stereo → 5.1

- Music Analysis
    - Transcription, beat tracking, harmony analysis etc.

- Music Education
    - Solo / Backing track generation

Source Separation
Tasks

- Harmonic/percussive separation

  - H → stable harmonic components (fundamental frequency, overtones)

  - P → transient components (drum sounds, note attacks)

- Solo/accompaniment separation

  - S → predominant melody instrument

  - A → accompanying instruments

- Singing voice separation

  - S → singing voice (male / female)

  - A → band

- Separation of all sources

Source Separation
Traditional Approaches

- Harmonic/percussive (H/P) separation

    - Different spectral characteristics of harmonic and percussive signals

spectrogram of saxophon track

Own

spectrogram of drum track
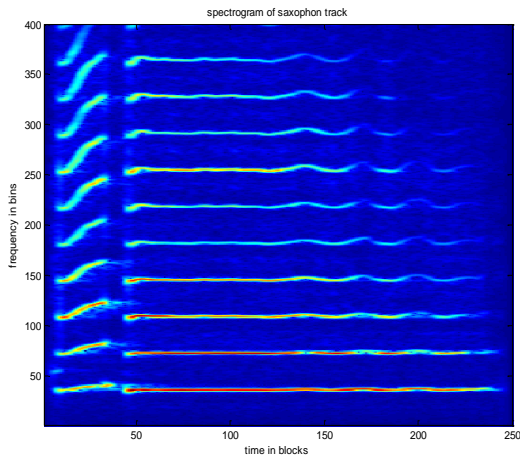
Own

- **Time-continuous (horizontal)**

- **Localized in frequency**
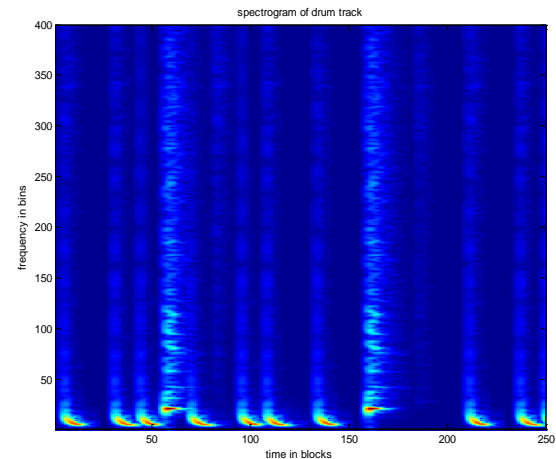
- **Wide-band (vertical)**

- **Localized in time**

Fraunhofer
IDMT

Source Separation
Traditional Approaches

FMP Notebooks

Fig. 14
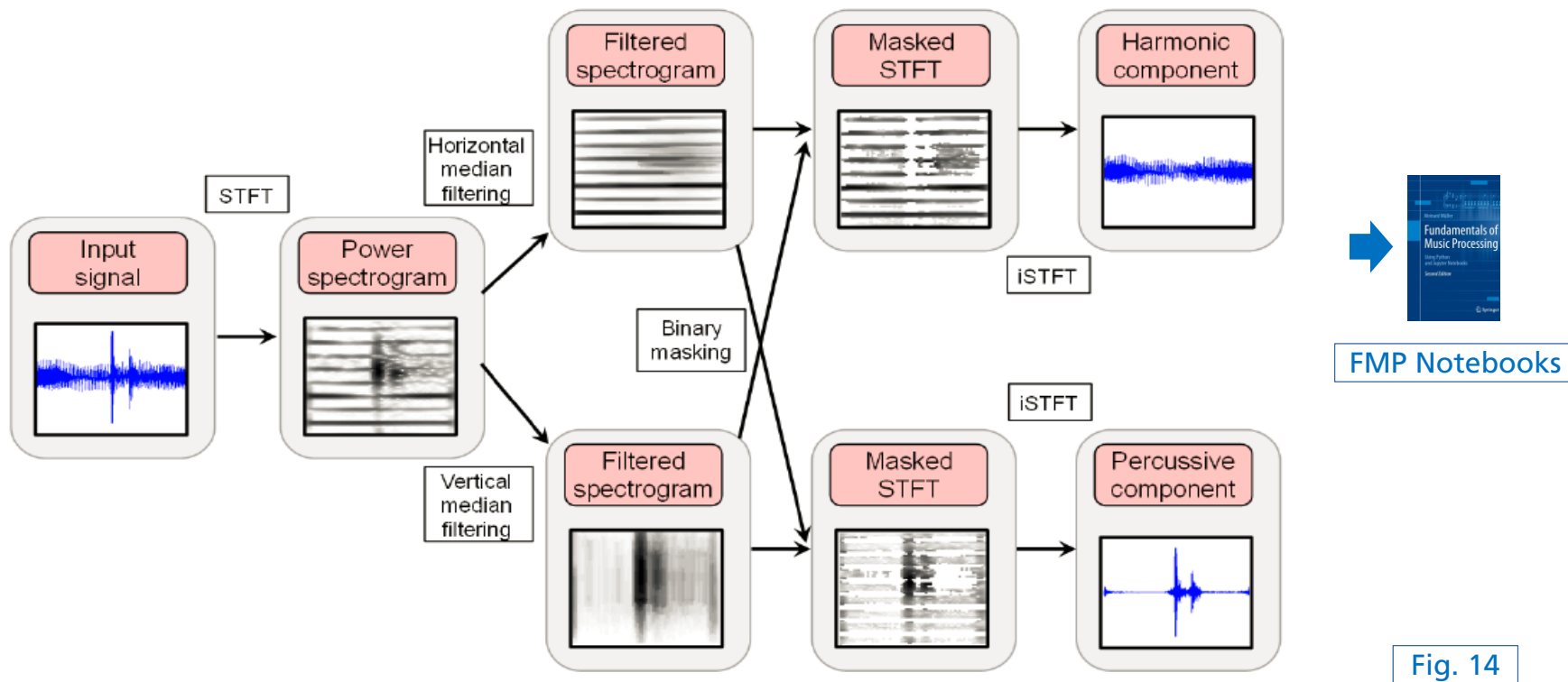
Source Separation
Traditional Approaches

- Phase-based H/P separation

    - Harmonic sources → phase change values are predictable

    - Percussive sources → unpredictable phase (noise-like characteristics)

    - Instantaneous Frequency Distribution (IFD)

        - How does phase change over time?

$$\Phi(k,n) = \frac{1}{2\pi} \frac{d\phi(k,n)}{dn}$$

Unwrapped phase

Instantaeneous Frequency

k – frequency bin

n – time frame

Fraunhofer
IDMT

Source Separation
Traditional Approaches

- Phase-based H/P separation

    - Harmonic mask → phase change within range / predictable?

$$H(k,n) = \begin{cases} 1 & \text{if } \Delta_{k_{Low}} < \Phi(k,n) < \Delta_{k_{High}} \\ 0 & \text{otherwise} \end{cases}$$

    - Percussive mask

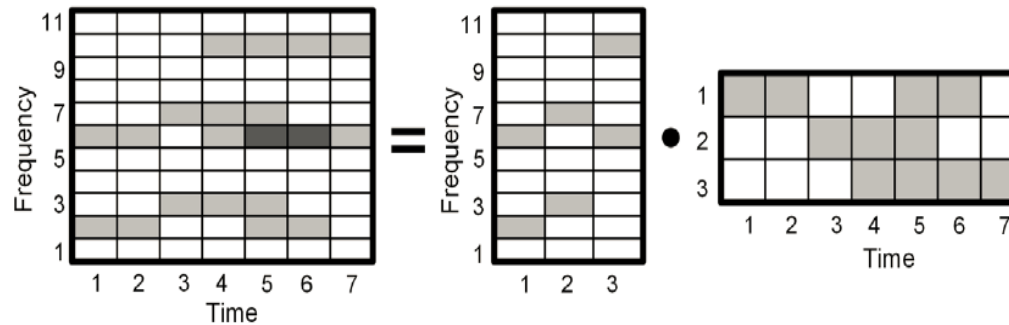$$P(k,n) = 1 - H(k,n)$$

Fraunhofer
IDMT

Source Separation
Traditional Approaches

- Non-Negative Matrix Factorization (NMF)
  - Factorize spectrogram V into set of components: $V \approx WH$

Templates

Activations

$$V \approx W \cdot H$$
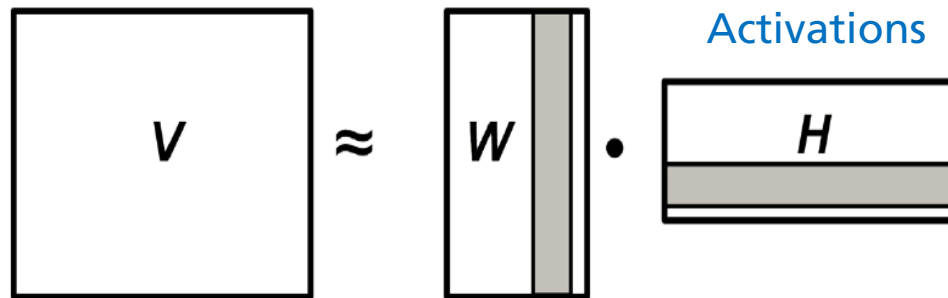
Source Separation
Traditional Approaches

- Non-Negative Matrix Factorization (NMF)
  - Algorithm: $V \approx WH$
    - Randomly initialize W & H
    - Use update rules to alternately update W & H
      - Minimize cost function
    - Cost function examples
      - Euclidean distance

$$||A - B||^2 = \sum_{ij}(A_{ij} - B_{ij})^2$$

      - Kullback-Leibler divergence

$$D(A||B) = \sum_{ij}\left(A_{ij}\log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}\right)$$

Fraunhofer
IDMT

Source Separation
Novel Approaches

- U-Net based [Jannson et al., 2017]

  - Input → magnitude spectrogram (mix)

  - Output →
    2 soft masks (voice / others)

- Issue

  - Only magnitude of STFT is modeled

  - Still phase from the mixture is used

Skip Connections

Input 512x128x1

Conv2D 256x64x16

Conv2D 128x32x32

Conv2D 64x16x64

Conv2D 32x8x128

Conv2D 16x4x256

Conv2D 8x2x512

Concat

Concat

Concat

Concat

Concat

Multiply

Output 512x128x1

Mask 512x128x1

Deconv2D 256x16x16

Deconv2D 128x32x32

Deconv2D Dropout 64x16x64

Deconv2D Dropout 32x8x128

Deconv2D Dropout 16x4x256

Encoder

Decoder

Fig. 15
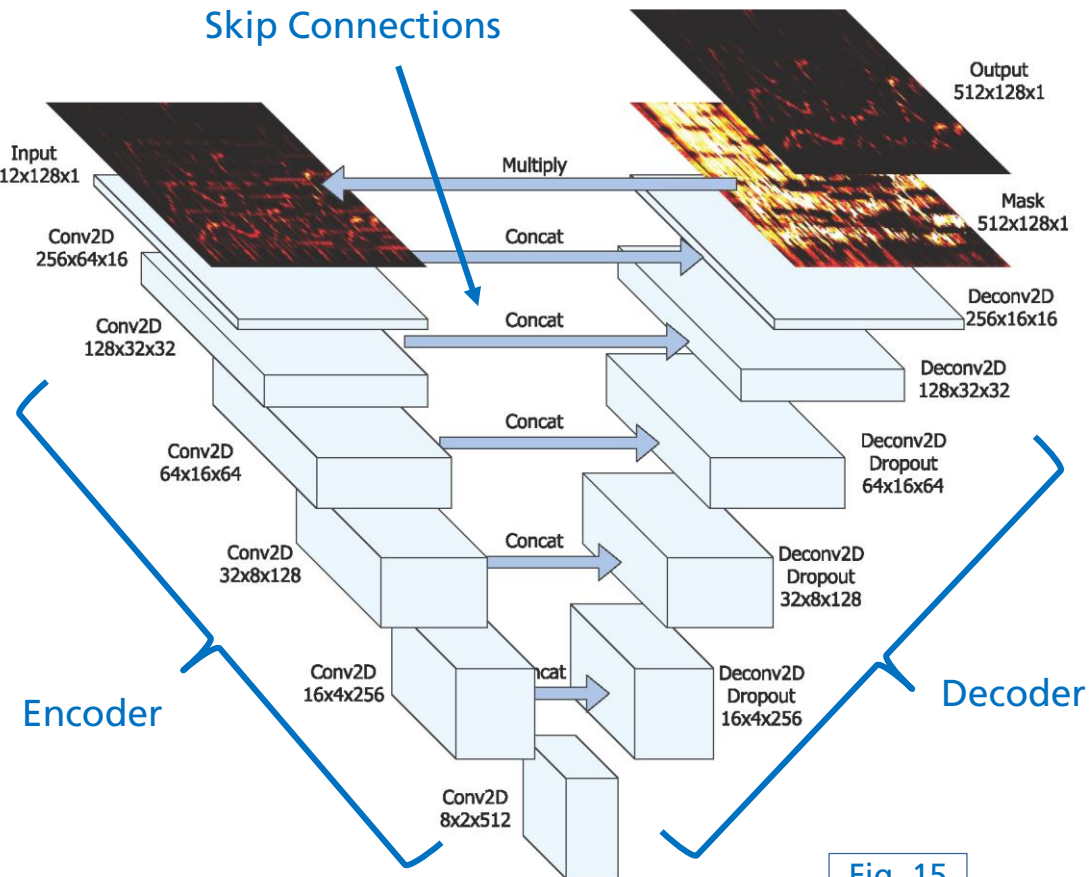
Source Separation
Novel Approaches

- Spleeter [Hennequin et al., 2020]

    - Open-source version for MIR research

    - 3 pre-trained models

        - 2 stems (vocals and accompaniments)

        - 4 stems (vocals, drums, bass, and other)

        - 5 stems (vocals, drums, bass, piano and other)

Spleeter Demo

Fraunhofer
IDMT

Source Separation
Novel Approaches

- Conv-TasNet [Luo & Mesgarani, 2019]

  - Time-domain speech separation network (end-to-end)

  - Encoder → optimized representation for speaker separation

  - Seperation → masks (weighting functions)

  - Decoder → invert to waveforms

  - Temporal convolutional networks (TCN)

    - Stack of 1-D dilated convolutional blocks

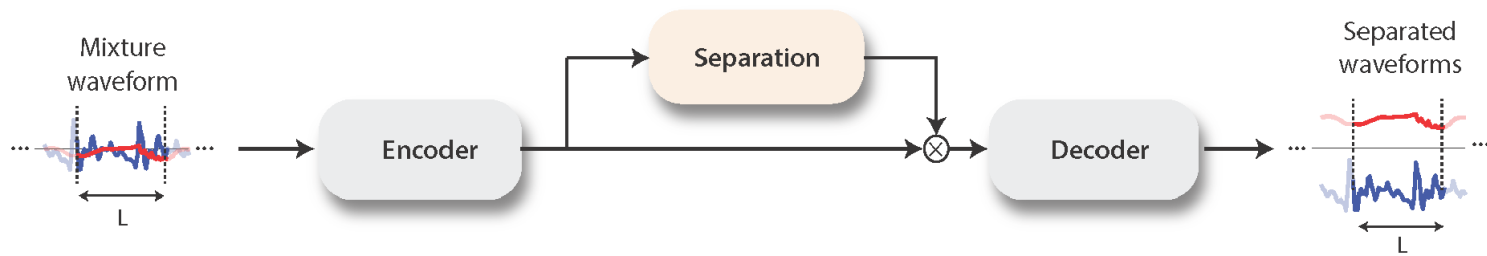   - Large receptive field → model long-term dependencies

Fig. 16

Source Separation
Novel Approaches

■ Conv-TasNet [Luo & Mesgarani, 2019]
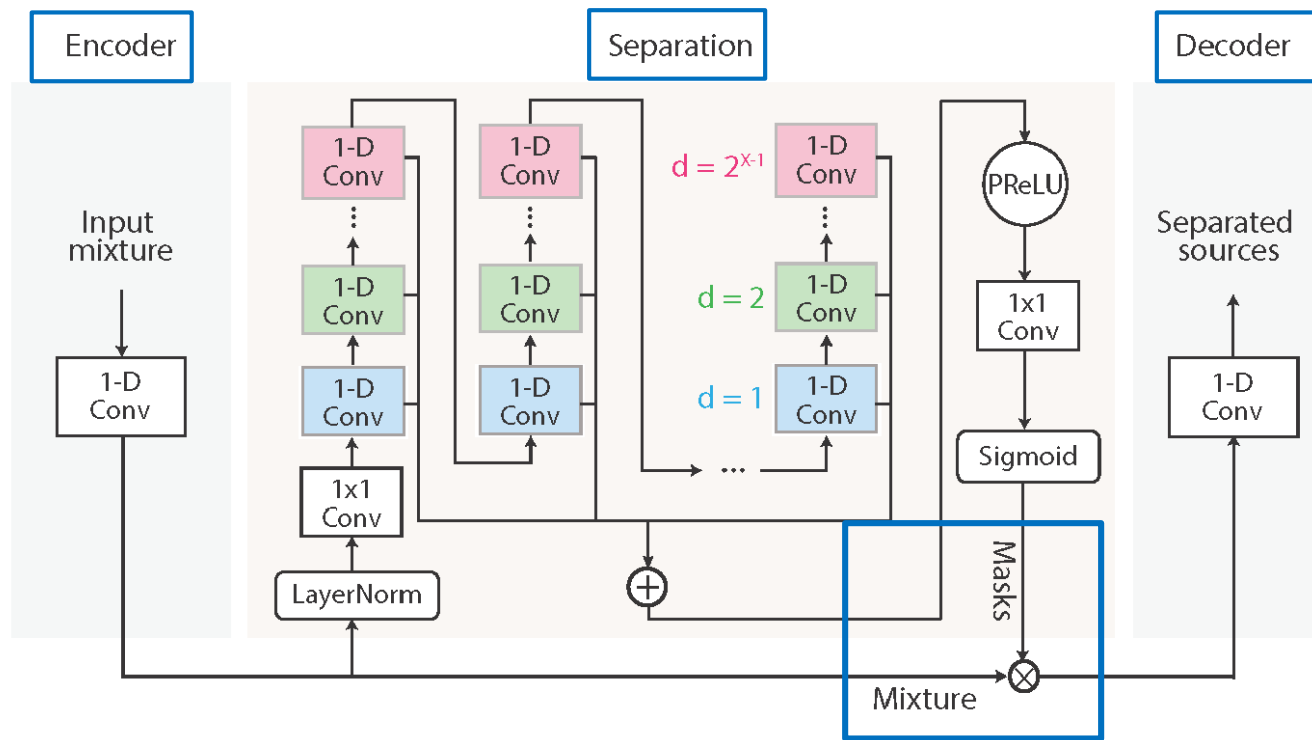

Fig. 17

Summary

- Case Studies
  - Pitch Detection
  - Instrument Recognition
  - Source Separation

Fraunhofer
IDMT

References

Cano, E., Fitzgerald, D., Liutkus, A., Plumbley, M. D., & Stoter, F. R. (2019). Musical Source Separation: An Introduction. *IEEE Signal Processing Magazine*, *36*(1), 31–40.

Grasis, M., Abeßer, J., Dittmar, C., & Lukashevich, H. (2014). A Multiple-Expert Framework for Instrument Recognition. *Lecture Notes in Computer Science 8905*, 619–634.

Han, Y., Kim, J., & Lee, K. (2017). Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *25*(1), 208–221.

Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, *5*(50), 2154.

Hsieh, T. H., Su, L., & Yang, Y. H. (2019). A Streamlined Encoder/Decoder Architecture for Melody Extraction. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 156–160. Brighton, UK.

Hung, Y.-N., & Yang, Y.-H. (2018). Frame-Level Instrument Recognition by Timbre and Pitch. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 135–142. Paris, France.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). Singing Voice Separation with Deep U-Net Convolutional Networks. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 745–751. Suzhou, China.

Fraunhofer

IDMT

References

Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018). Crepe: A Convolutional Representation for Pitch Estimation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 161–165. New Orleans, USA.

Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(8), 1256–1266.

Müller, M. (2021). *Fundamentals of Music Processing - Using Python and Jupyter Notebooks* (2nd ed.). Springer.

Salamon, J., & Gomez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech and Language Processing*, *20*(6), 1759–1770.

Images

Fig. 1: [Müller, 2021], p. 449, Fig. 8.15(b)

Fig. 2: http://www.guitaradventures.com/wp-content/uploads/Tuning-your-guitar.jpg

Fig. 3: https://cdn2.whatoplay.com/screenshots/2631slide-4.jpg

Fig. 4: https://cdn.androidcommunity.com/wp-content/uploads/2010/11/500x_angrybirdsdarwin.jpg

Fig. 5: [Müller, 2021], p. 449, Fig. 8.15(a)

Fig. 6: Sonic Visualiser: http://www.sonicvisualiser.org/ , Melodia plugin: http://mtg.upf.edu/technologies/melodia

Fig. 7: [Kim et al., 2018], p. 2, Fig. 1

Fig. 8: [Hsieh et al., 2019], p. 2, Fig. 2

Fig. 9: [Grasis et al., 2014], p. 6, Fig. 3

Fig. 10: [Han et al., 2017], p. 3, Fig. 1

Fig. 11: [Han et al., 2017], p. 9, Fig. 6

Fig. 12: [Hung & Yang, 2018], p. 4, Fig. 1

Fig. 13: [Cano et al., 2019], p. 3, Fig. 3

Fig. 14: [Müller, 2021], p. 425, Fig. 8.3

Images

Fig. 15: [Jansson, 2017], p. 3, Fig. 1

Fig. 16: [Luo & Mesgarani, 2019], p. 3, Fig. 1(A)

Fig. 17: [Luo & Mesgarani, 2019], p. 3, Fig. 1(B)

Fig. 18: [Müller, 2021], p. 422, Fig. 8.1

Sounds

AUD-1: Aislinn – Capclear (2013), https://freemusicarchive.org/music/Aislinn/Aislinn/10_-_Aislinn_-_Capclear

AUD-2: Aislinn – Fourteen Days (2013), https://freemusicarchive.org/music/Aislinn/Aislinn/11_-_Aislinn_-_Fourteen_days

AUD-3: Anonymous Choir – Amicus Meus (2009), https://freemusicarchive.org/music/Anonymous_Choir/Toms_Luis_de_Victorias_Amicus_Meus/Amicus_Meus

Fraunhofer

IDMT

Thank you!

- Any questions?

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

https://www.machinelistening.de

Fraunhofer

IDMT