Computational Analysis of Sound and Music

Environmental Sound Analysis – Sound Event Detection 1

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

jakob.abesser@idmt.fraunhofer.de

Sound Event Detection 1

Outline

- Introduction
- Challenges & Related Tasks
- Pipeline
- Evaluation

Sound Event Detection 1

Introduction

- Sound event detection → 2 simultaneous tasks
  - Segmentation (detection of temporal boundaries)
  - Classification (type of sound)
- Sound polyphony
  - Number of simultaneous sounds
  - Depends on the acoustic scene composition & sound sources

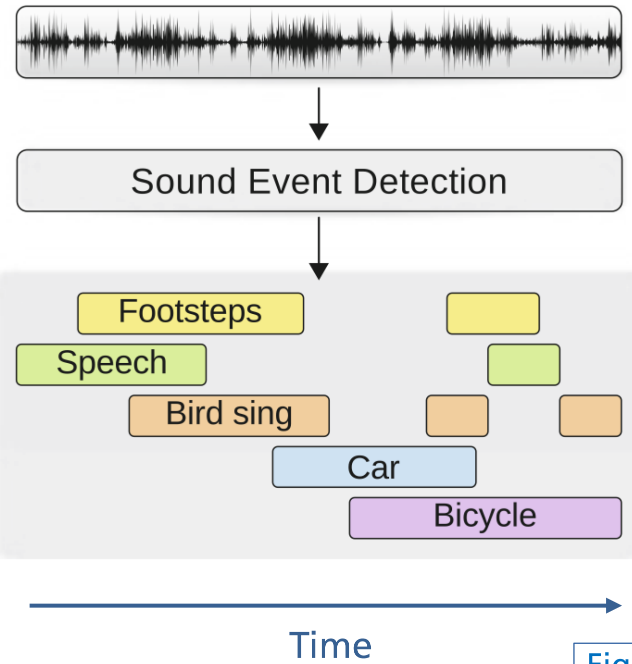


Fig-E1-1

Sound Event Detection 1

Introduction

- USM dataset [Abeßer, 2022]

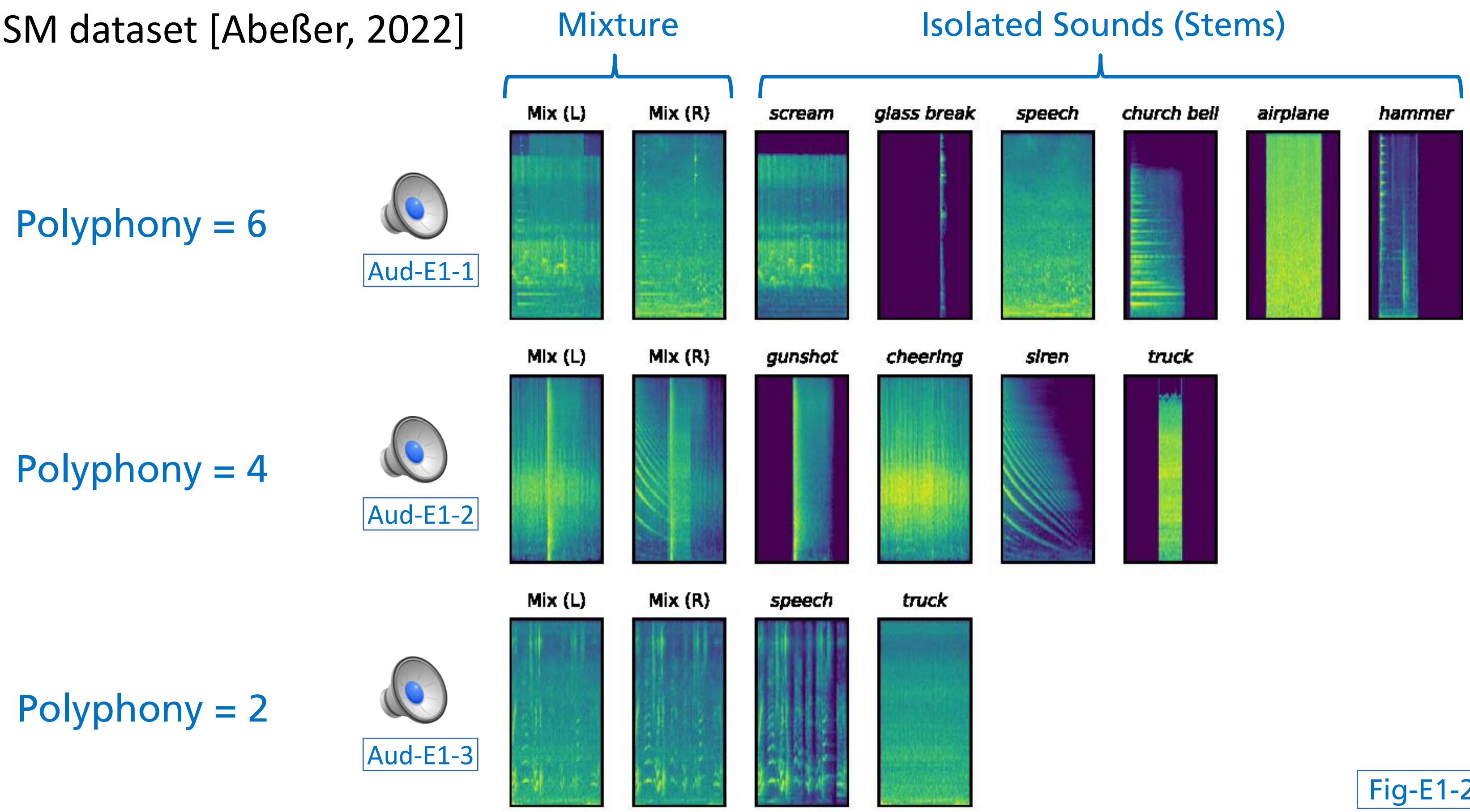


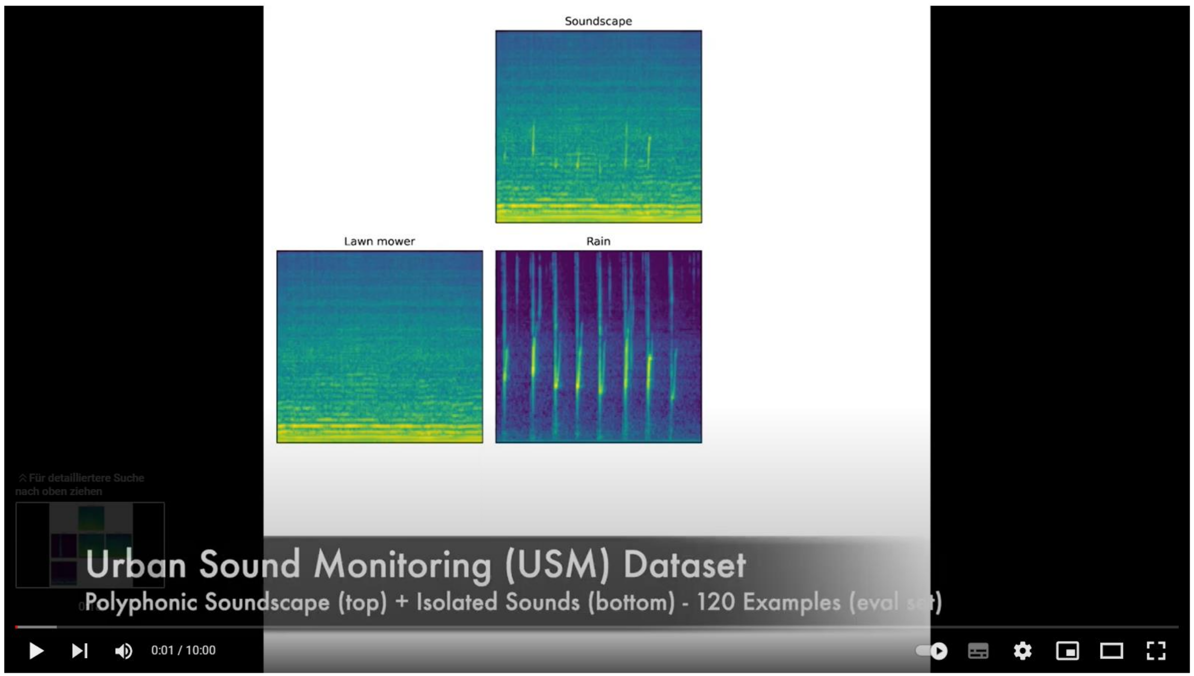
Fig-E1-2

Sound Event Detection 1

Introduction

- USM dataset [Abeßer, 2022]

Demo-Video



Demo of the Urban Sound Monitoring (USM) Dataset for Polyphonic Sound Event Tagging

Sound Event Detection 1

Introduction

- Sound source categories
    - Humans, animals, vehicles, tools, machines, climate, …
- Sound hierarchies
    - Based on regional origin & sound characteristics

Aud-E1-4

Example: Urban Sounds

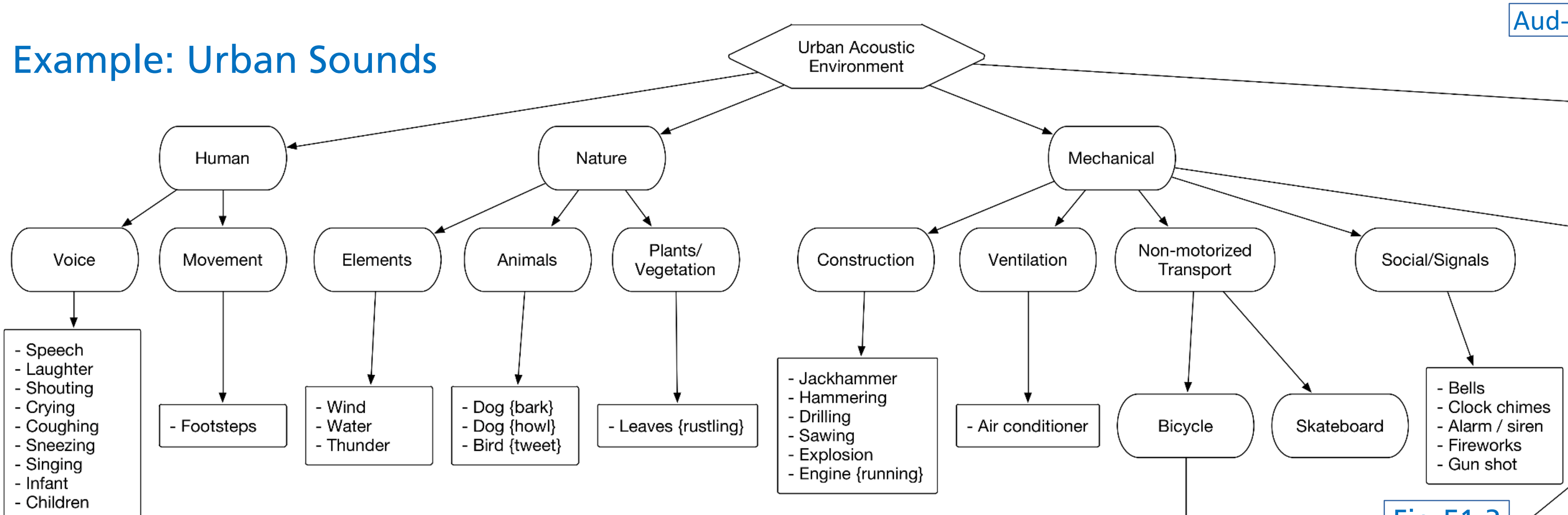


Fig-E1-3

Sound Event Detection 1

Challenges

- Sound characteristics
  - Short transients, noise-like signals, harmonic / inharmonic signals
- Sound durations
  - Short (gun shot, door knock) → long / stationary (machines, wind)
- Ill-defined temporal boundaries
  - Complicates annotation & detection

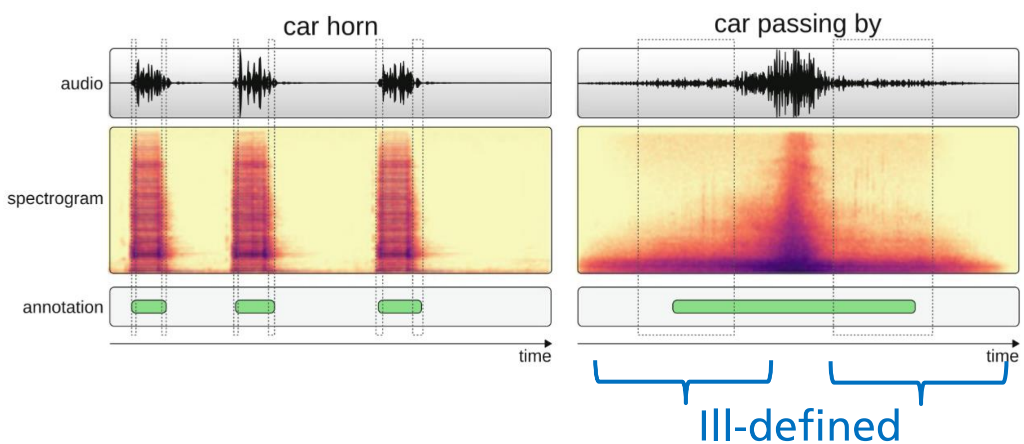


Fig-E1-4

Sound Event Detection 1

Challenges

- Sound appear in the foreground & background
  - depending on relative sound source position
- Non-local / sparse energy distribution
  - Example: fundamental frequency & overtones

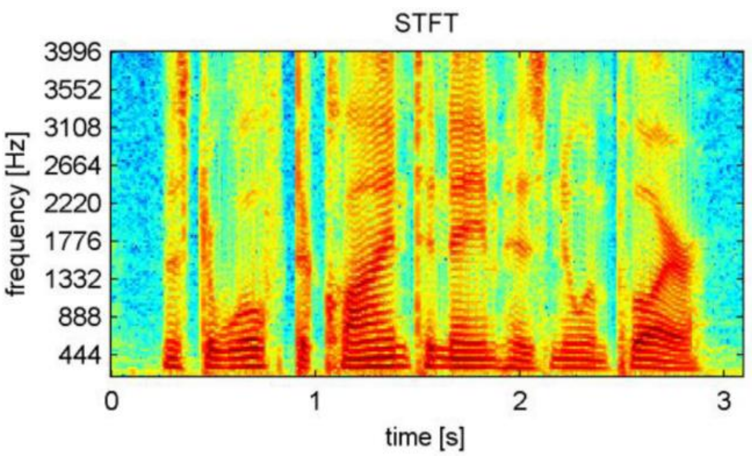


Fig-E1-5

Sound Event Detection 1

Related Tasks

- Sound event localization & tracking
    - Multichannel audio recordings (e.g., first-order ambisonic microphones)
- Estimate direction-of-arrival (DOA) & track source movement

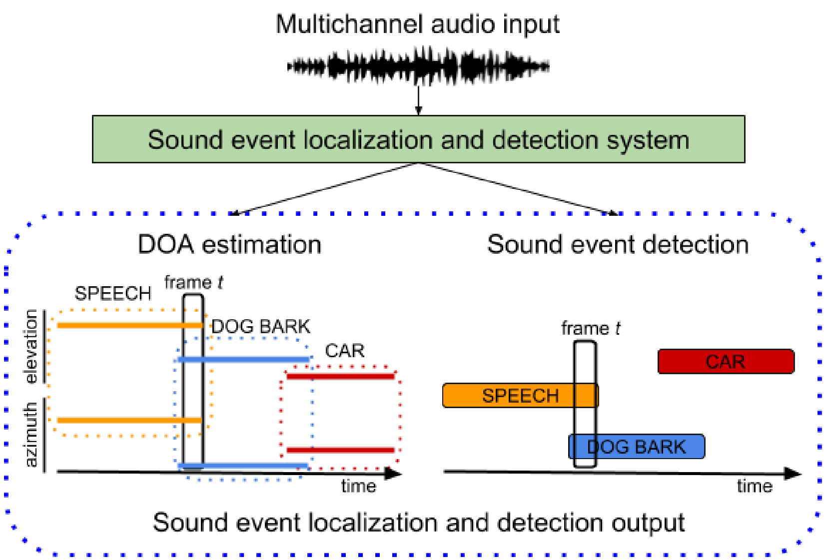


Fig-E1-6

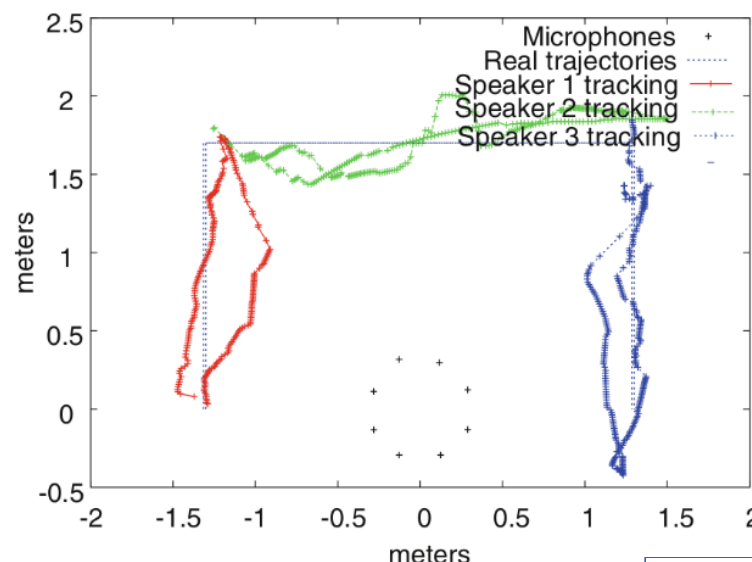


Fig-E1-7

Sound Event Detection 1

Related Tasks

- Source separation
    - Facilitates sound event detection (fewer background sounds)
- Chicken-egg problem
    - Alternative: sound-informed source-separation

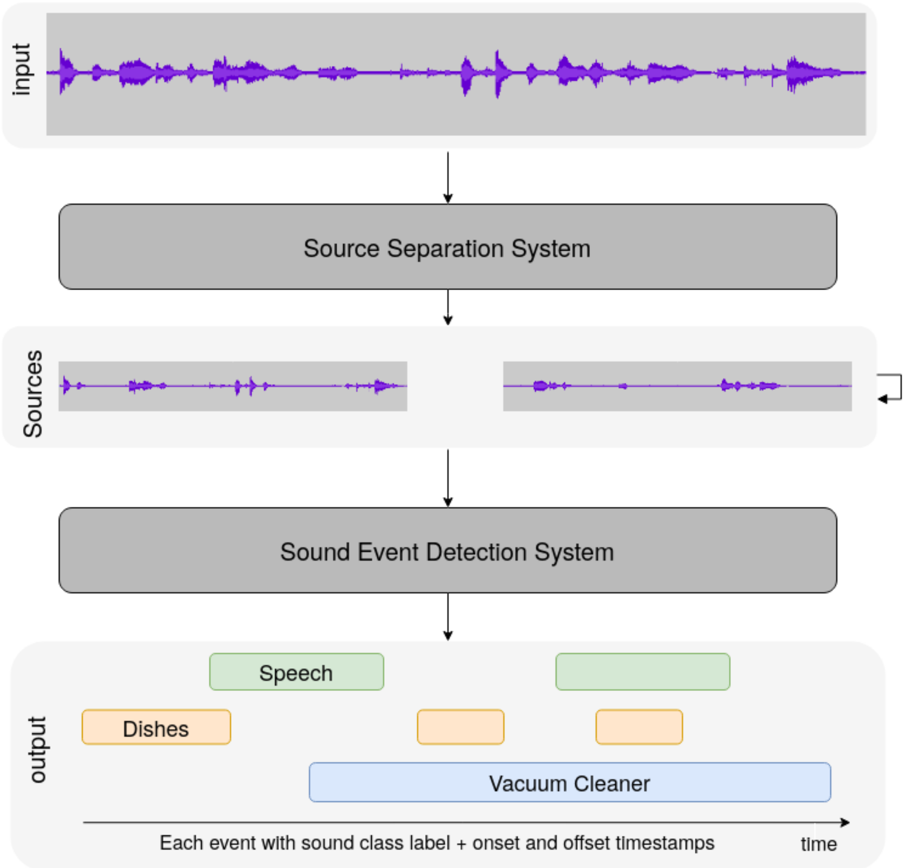


Fig-E1-8

Sound Event Detection 1

Pipeline

- Supervised learning pipeline
  - Feature extraction & pre-processing
  - Label encoding
  - Acoustic modeling

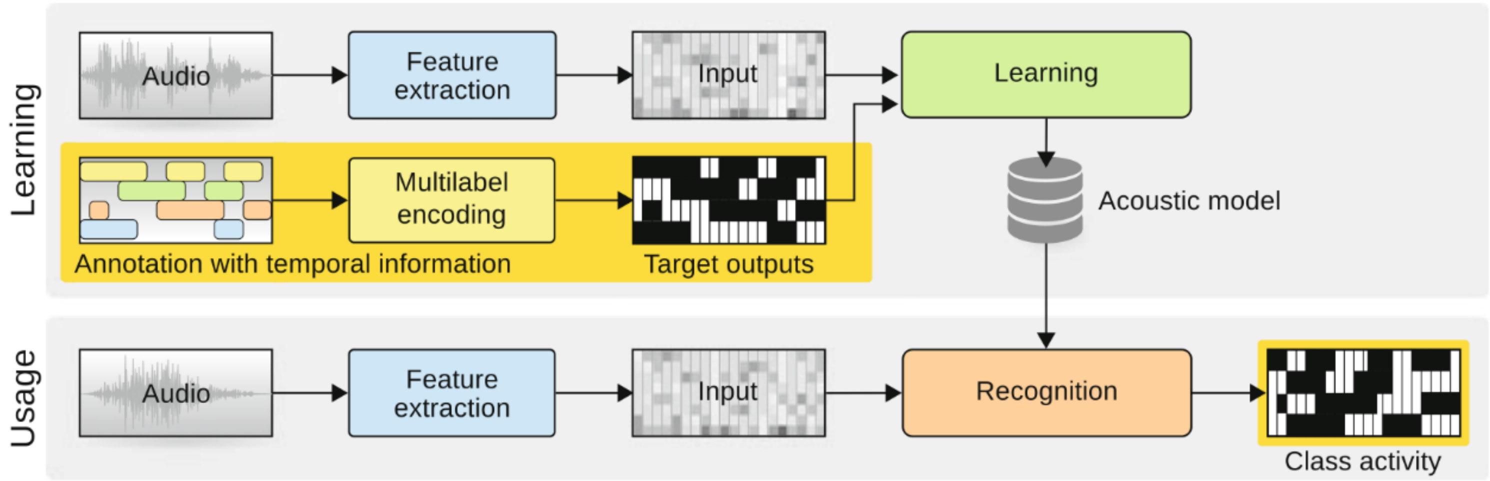
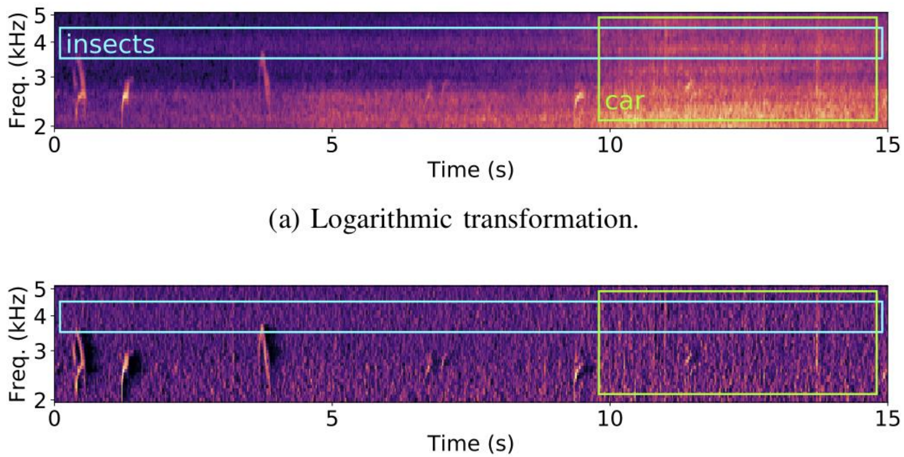


Fig-E1-9

Sound Event Detection 1

Pipeline

- Feature extraction
    - 1D features (audio samples) → “end-to-end learning”
    - 2D features (mel-spectrogram, STFT)
- Feature pre-processing
    - Log-magnitude scaling
    - Per-channel energy (PCEN) [Lostanlen, 2019]
        - Dynamic range compression
        - Adaptive gain control
        - Suppresses stationary (background) noise



(a) Logarithmic transformation.

(b) Per-channel energy normalization (PCEN).

Fig-E1-10

Sound Event Detection 1

Pipeline

- Annotation
  - Quality of "ground truth"? (limited agreement / reliability)
  - Different granularities
    - Tagging / Global level ("weak" labels) → cheap
    - Event-level ("strong" labels) → expensive

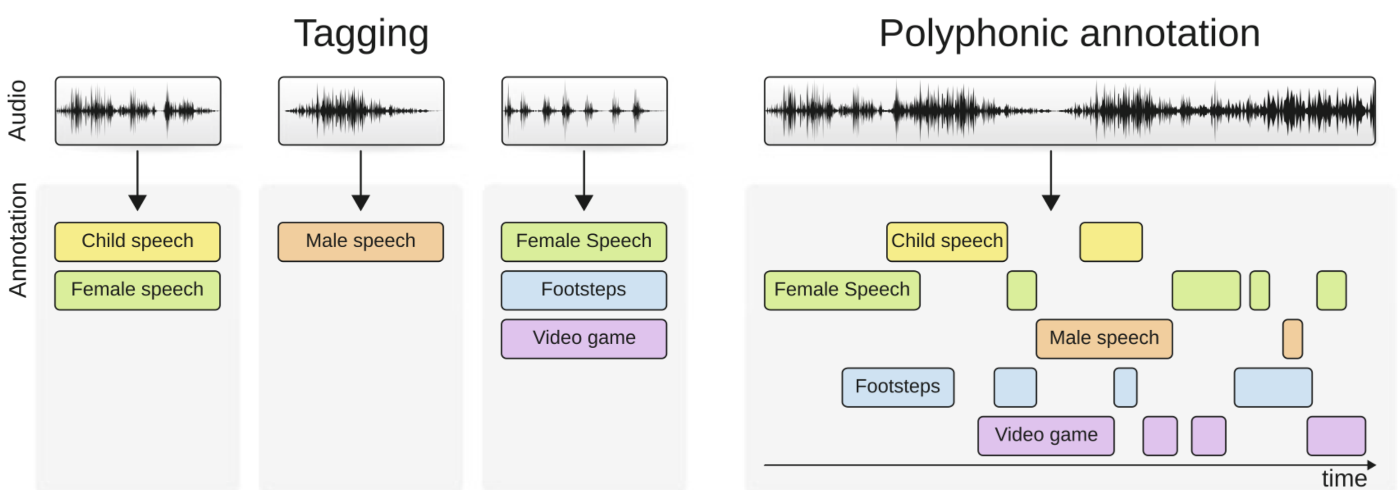


Fig-E1-11

Sound Event Detection 1

Pipeline

- Label encoding
    - Binarized sound activity (0/1)
        - Multilabel classification
        - 1 (independent) binary detector per class
    - Temporal resolution (duration of each annotated time frame)

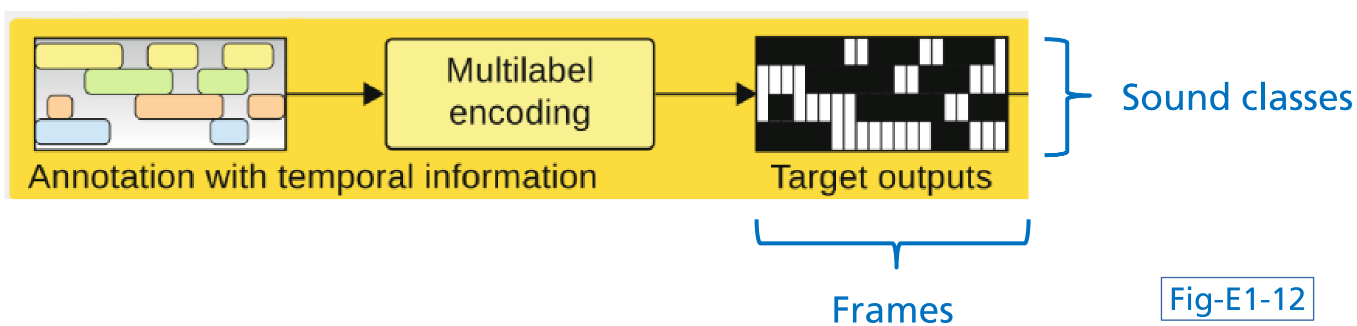


Fig-E1-12

Sound Event Detection 1

Pipeline

- Typical neural network architectures
- CNN

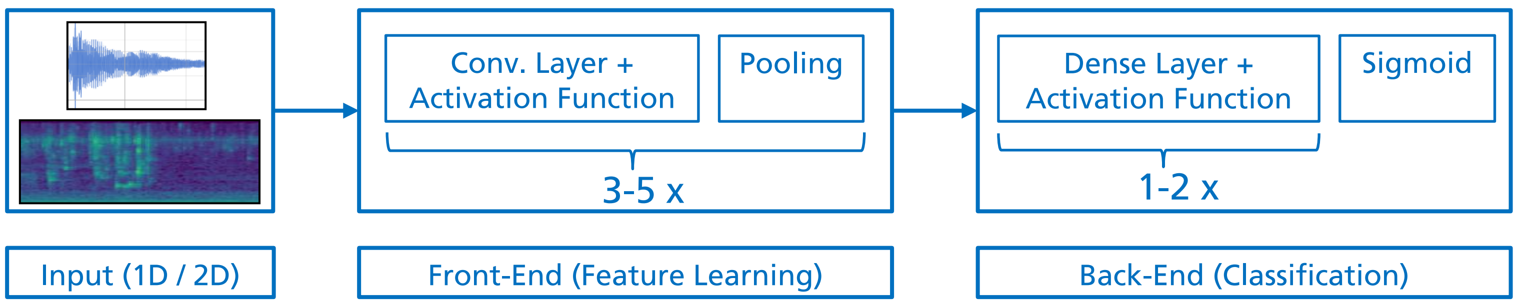


Fig-E1-13

- CRNN

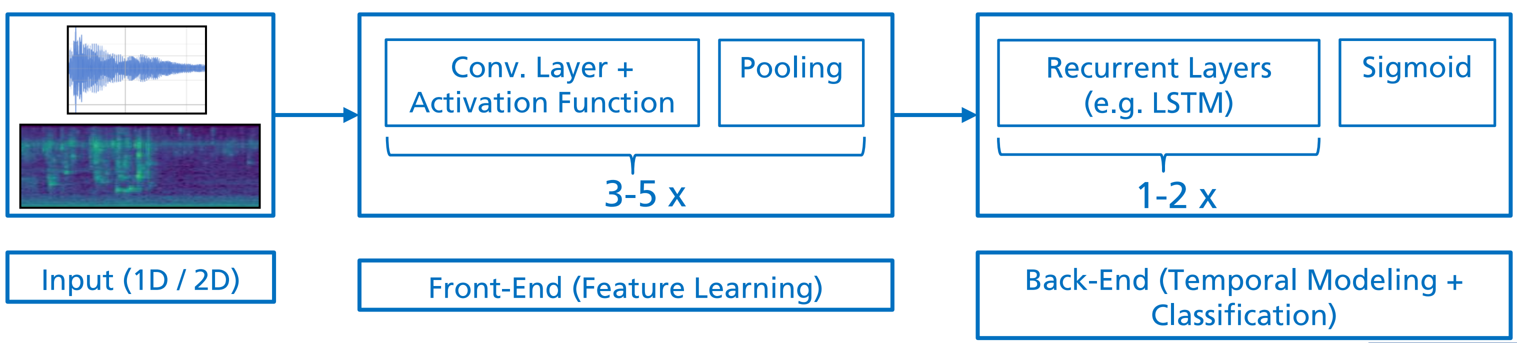


Fig-E1-14

Sound Event Detection 1

Evaluation

- Evaluate SED → binary classification results on a frame-level
- Compare reference with predictions
- Count TP/FN/FP → aggregate over time → compute metrics

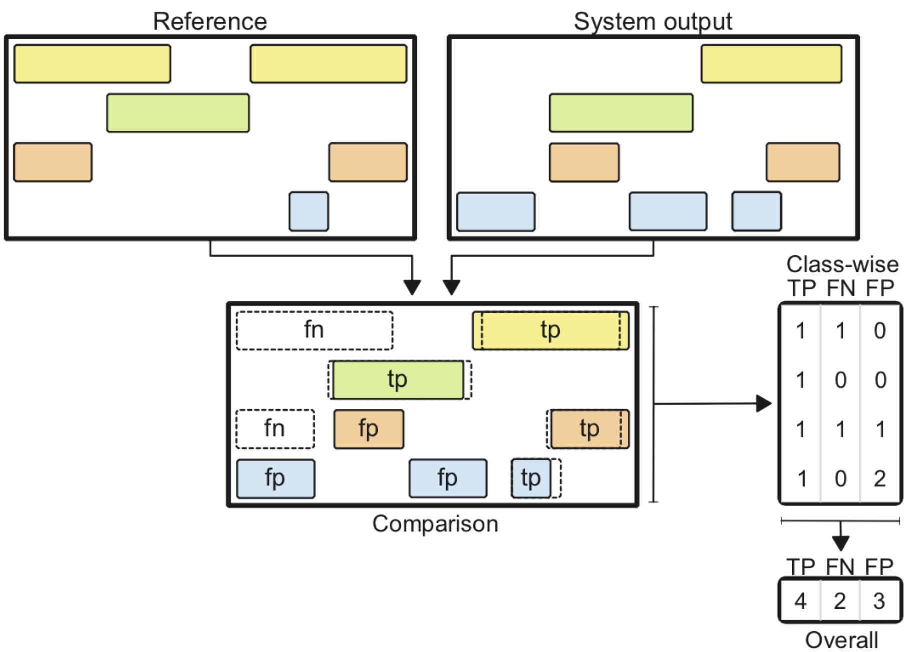


Fig-E1-15

Sound Event Detection 1

Evaluation

- Binary classification evaluation
    - True/false positives (TP/FP)
    - True/false negatives (TN/FN)
- Metrics
    - Precision
    - Recall
    - Accuracy
    - F-score

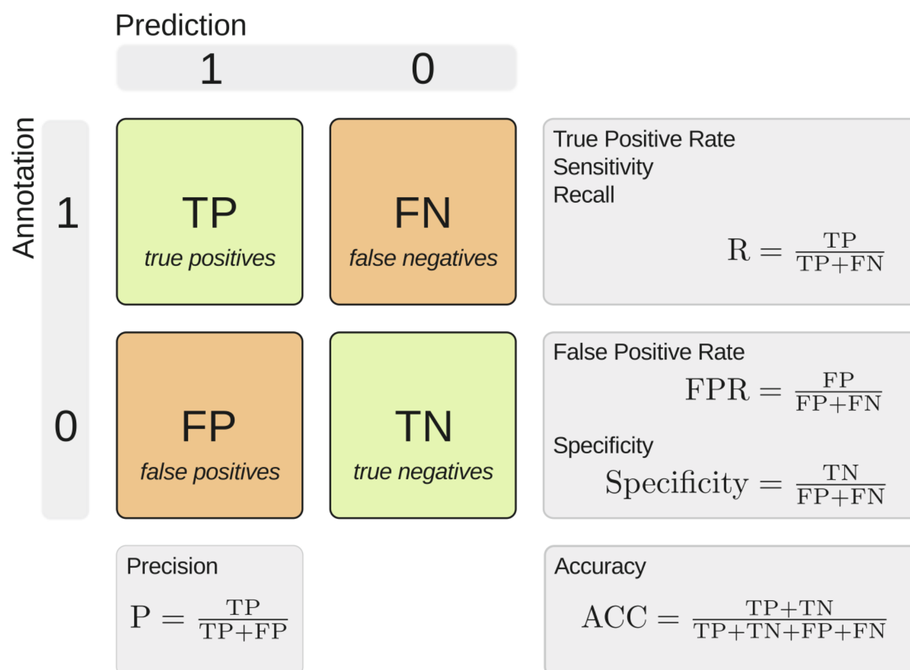


Fig-E1-16

Programming session

Fig-A2-13

References

Images

Fig-E1-1: [Virtanen, 2018], p. 15, Fig. 2.1

Fig-E1-2: [Own]

Fig-E1-3: https://urbansounddataset.weebly.com/uploads/4/3/9/4/4394963/3427002_orig.png

Fig-E1-4: [Virtanen, 2018], p. 157, Fig. 6.3

Fig-E1-5: https://towardsdatascience.com/whats-wrong-with-spectrograms-and-cnns-for-audio-processing-311377d7ccd

Fig-E1-6: http://dcase.community/challenge2019/task-sound-event-localization-and-detection, Fig. 1

Fig-E1-7: [Virtanen, 2018] , p. 267, Fig. 9.7

Fig-E1-8: http://dcase.community/challenge2020/task-sound-event-detection-and-separation-in-domestic-environments, Fig. 2

Fig-E1-9: [Virtanen, 2018], p. 31, Fig. 2.11

Fig-E1-10: [Lostanlen, 2019], p. 1, Fig. 1

Fig-E1-11: [Virtanen, 2018], p. 154, Fig. 6.2

Fig-E1-12: [Virtanen, 2018], p. 31, Fig. 2.11 (excerpt)

Fig-E1-13 & 14: [Own]

Fig-E1-15: [Virtanen, 2018], p. 169, Fig. 6.6

Fig-E1-16: [Virtanen, 2018], p. 170, Fig. 6.7

References

Audio

Aud-E1-1: USM v2 dataset, Evaluation Set, Sound ID 2417

Aud-E1-2: USM v2 dataset, Evaluation Set, Sound ID 1930

Aud-E1-3: USM v2 dataset, Evaluation Set, Sound ID 339

Aud-E1-4: G_M_D_THREE - CANAL_STREET_NEW_YORK_A011.wav (2018) - CC0 License, https://freesound.org/people/G_M_D_THREE/sounds/424404

References

References

Abeßer, J. (2022). Classifying Sounds in Polyphonic Urban Sound Scenes. Proceedings of the 152nd AES convention, online

Virtanen, T., Plumbley, M. D., & Ellis, D. (Eds.). (2018). Computational Analysis of Sound Scenes and Events. Cham, Switzerland: Springer International Publishing.

Lostanlen, V., Salamon, J., Cartwright, M., McFee, B., Farnsworth, A., Kelling, S., & Bello, J. P. (2019). Per-Channel Energy Normalization: Why and How. IEEE Signal Processing Letters, 26(1), 39–43.