Machine Listening for Music and Sound Analysis

Lecture 5 - Environmental Sound Analysis 1

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

https://machinelistening.github.io

Overview

- Introduction
- Sound Event Detection
    - Introduction
    - Challenges & Related Tasks
    - Pipeline
    - Evaluation Metrics & Datasets
    - Data Augmentation
    - Methods
        - Traditional
        - Neural Network Based

Introduction
Motivation

- Sound carries information about our environment
- Challenging attempt to mimic the human‘s abilities

Introduction
Motivation

- Sound carries information about our environment
- Challenging attempt to mimic acoustic scene understanding
  - Environment perception
  - Localization of sound sources
  - Context-awareness

Introduction
Motivation

- Sound carries information about our environment
- Challenging attempt to mimic acoustic scene understanding
  - Environment perception
  - Localization of sound sources
  - Context-awareness
- Complementary sensory path to vision → multimodality
- Related to other content analysis domains (speech, music)

Introduction
Environmental Sounds (Recap)

- Sound sources
  - Nature, climate, humans, machines, etc.

AUD-1

Fig. 1

Fig. 2

Fig. 3

Introduction
Environmental Sounds (Recap)

- Sound sources
  - Nature, climate, humans, machines, etc.
- Sound characteristics
  - Stationary or non-stationary, repetitive or without any predictable nature
- Sound duration
  - Very short (gun shot, door knock, shouts)
  - Very long (running machines, wind, rain)

AUD-1

Fig. 1

Fig. 2

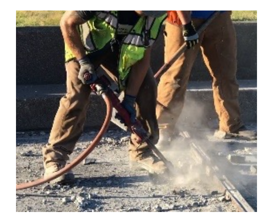


Fig. 3

Introduction
Tasks / Categories

- Sound event detection (SED)
- Acoustic scene classification (ASC)
- Acoustic anomaly detection (AAD)

Sound Event Detection
Introduction

- Sound event detection
  - Segmentation (detection of temporal boundaries)
  - Classification (type of sound)

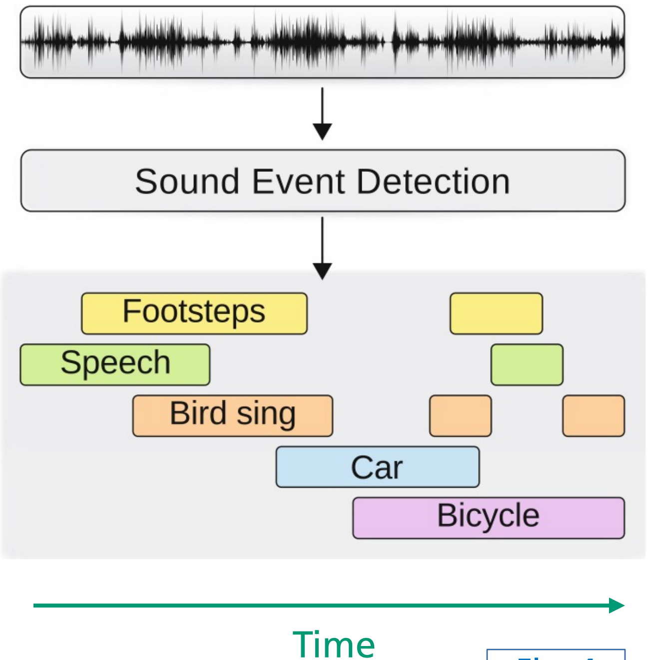


Fig. 4

Sound Event Detection
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

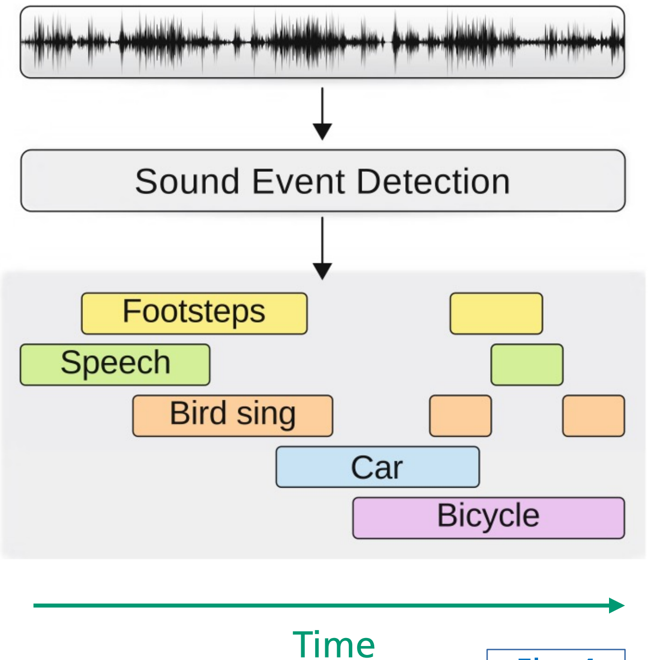
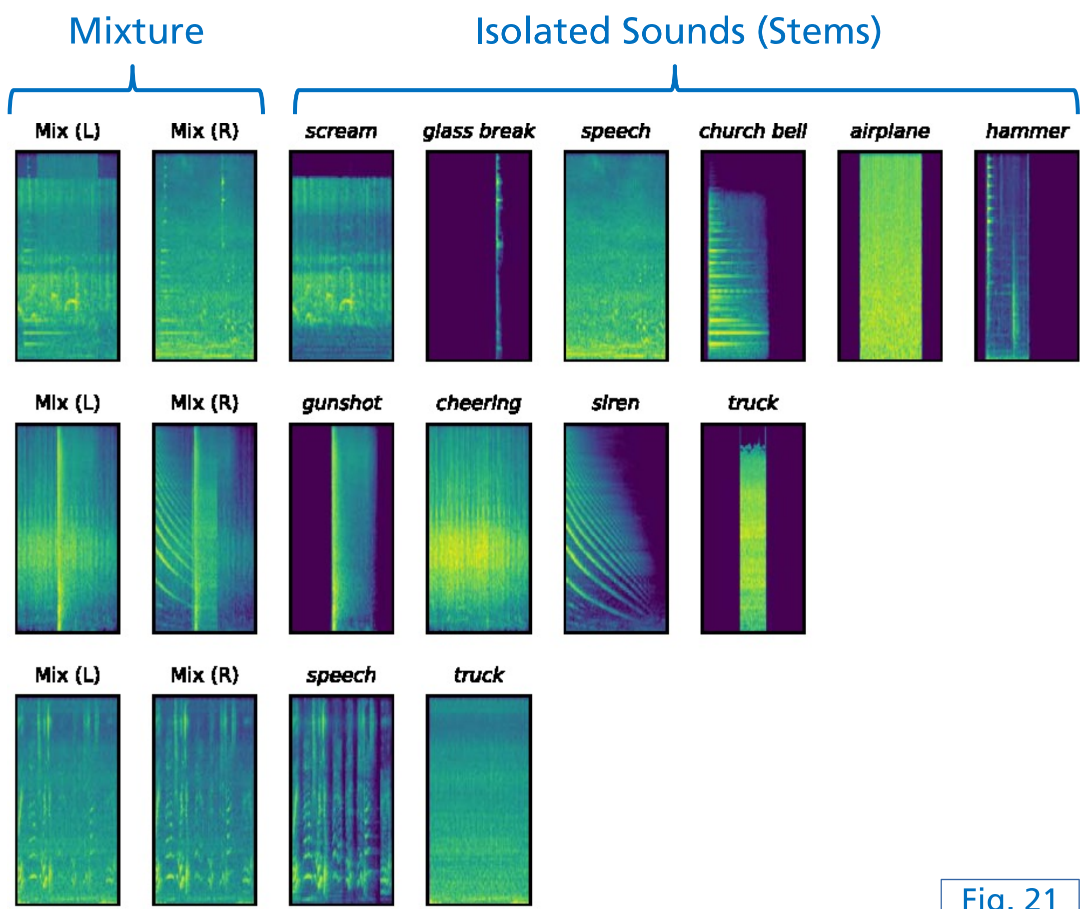


Fig. 4

Sound Event Detection
Introduction

- USM dataset [Abeßer, 2022]



Polyphony = 6 AUD-6

Polyphony = 4 AUD-7

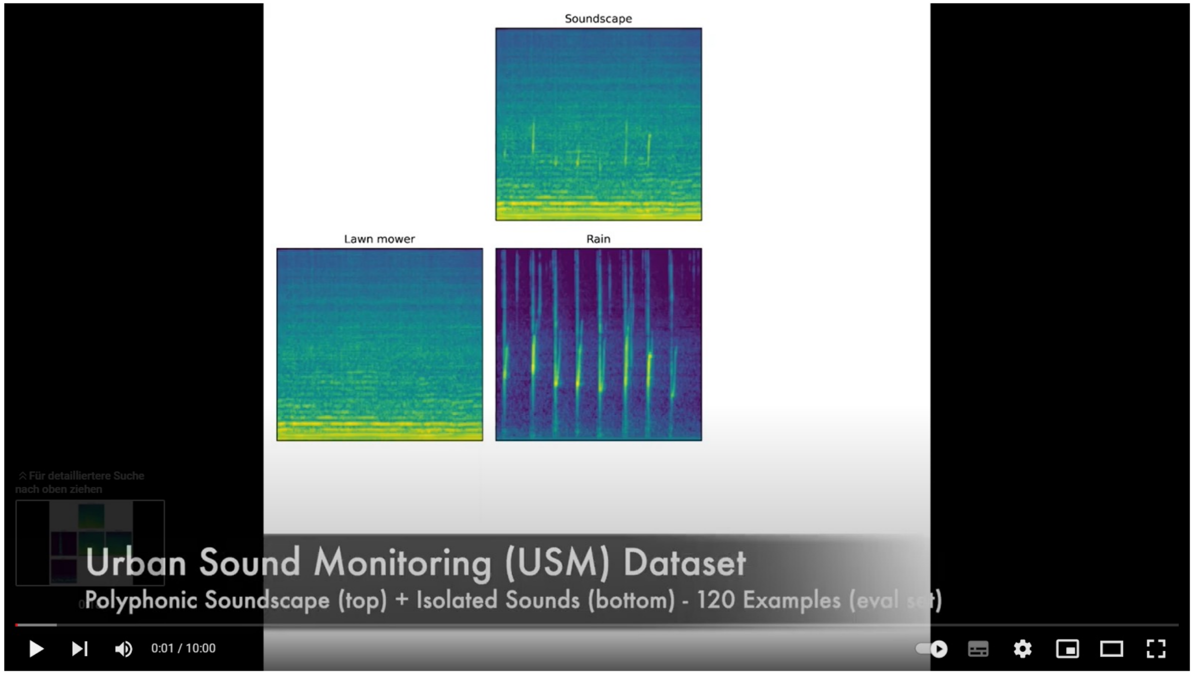
Polyphony = 2 AUD-8

Fig. 21

Sound Event Detection
Introduction

- USM dataset [Abeßer, 2022]

Demo-Video



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

Sound Event Detection
Introduction

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

Sound Event Detection
Introduction

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

AUD-2

Example: Urban Sounds

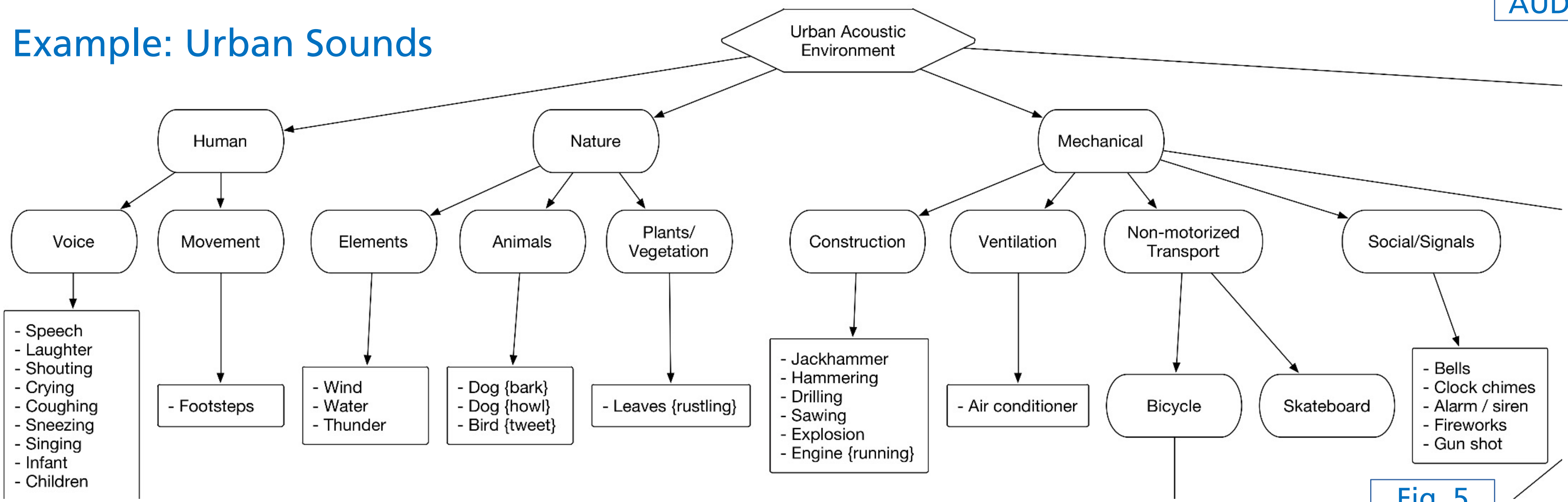


Fig. 5

Sound Event Detection
Challenges

- Sound characteristics
  - Short transients, noise-like signals, harmonic / inharmonic signals

Sound Event Detection
Challenges

- Sound characteristics
  - Short transients, noise-like signals, harmonic / inharmonic signals
- Sound durations
  - Short (gun shot, door knock) → long / stationary (machines, wind)

Sound Event Detection
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

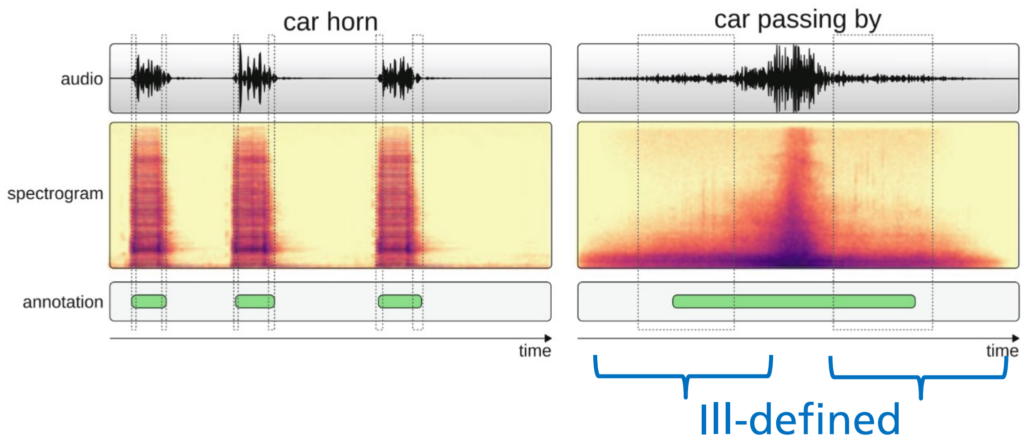


Fig. 6

Sound Event Detection
Challenges

- Sound appear in the foreground & background
  - depending on relative sound source position

Sound Event Detection
Challenges

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

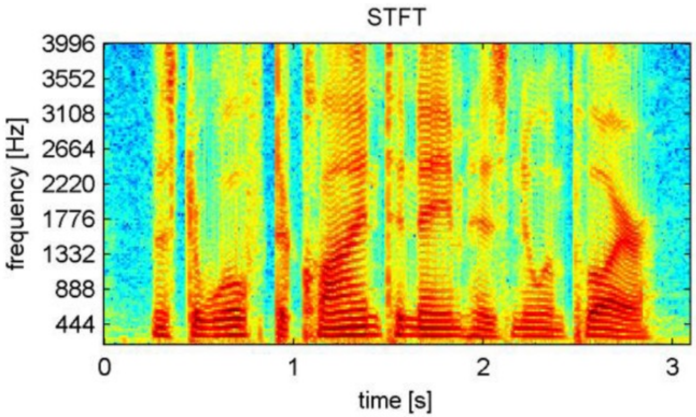


Fig. 7

Sound Event Detection
Challenges

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

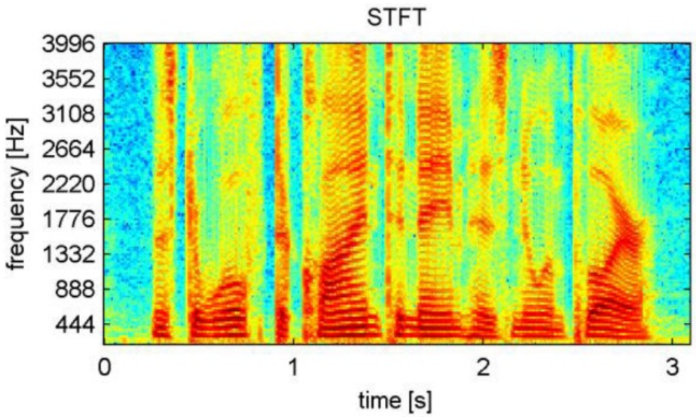


Fig. 7

- Sounds overlap / visual objects occlude
  - Possible phase cancellation

Sound Event Detection
Related tasks

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

Sound Event Detection
Related tasks

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

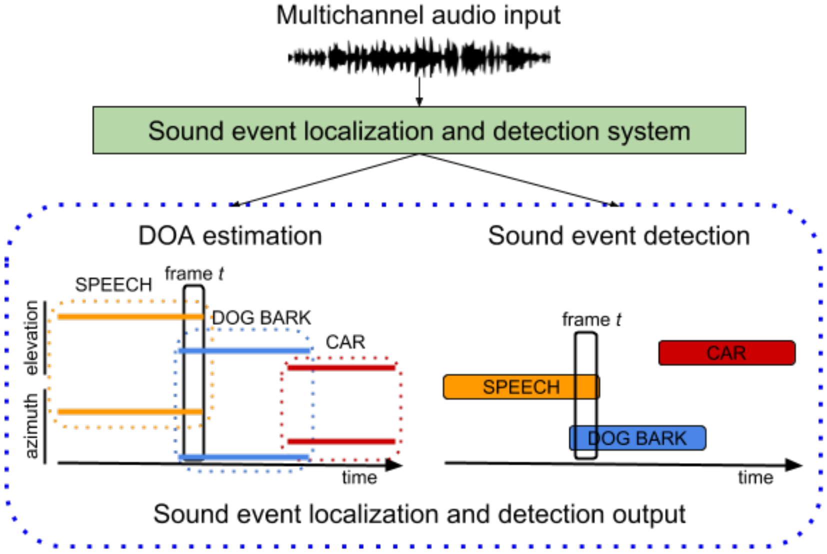


Fig. 14

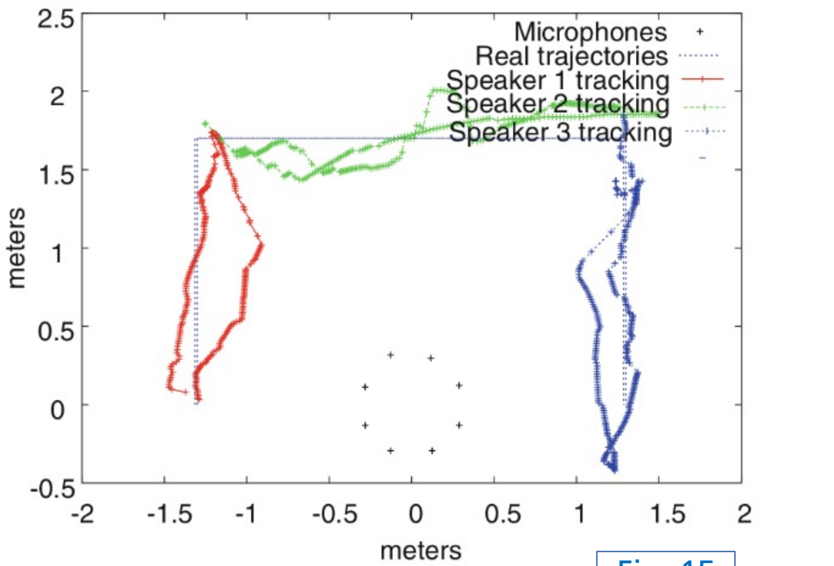


Fig. 15

Sound Event Detection
Related tasks

- Source separation
  - Prior to sound event detection
- Chicken-egg problem
  - Alternative: sound-informed source-separation

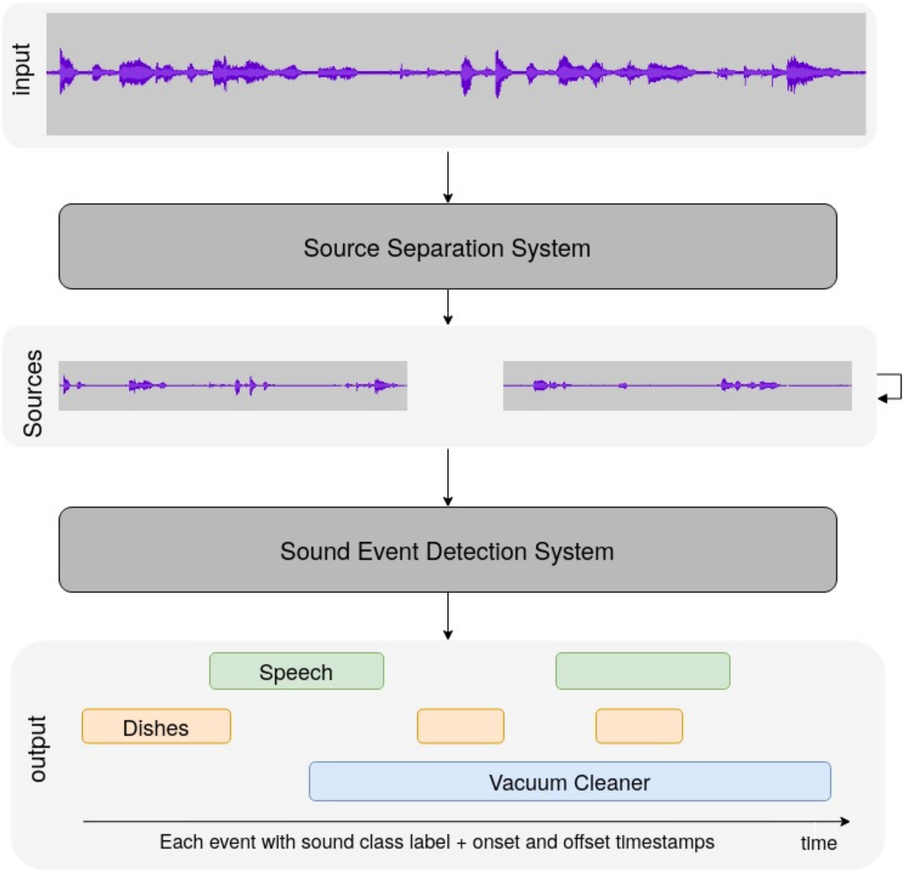


Fig. 16

Sound Event Detection
Pipeline

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

Sound Event Detection
Pipeline

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

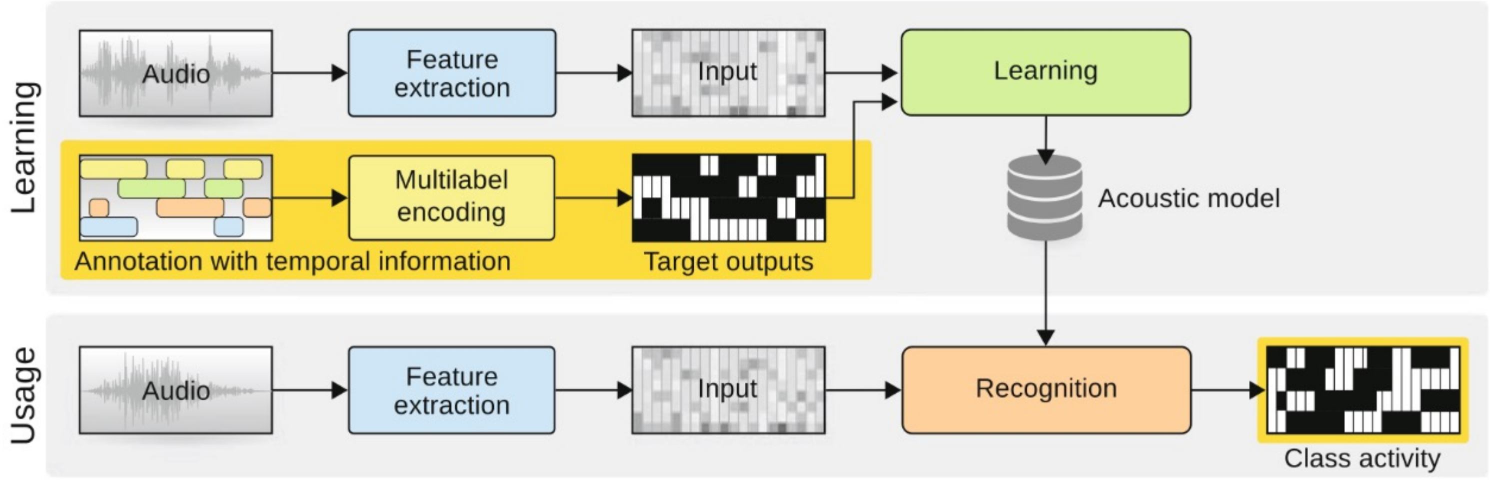


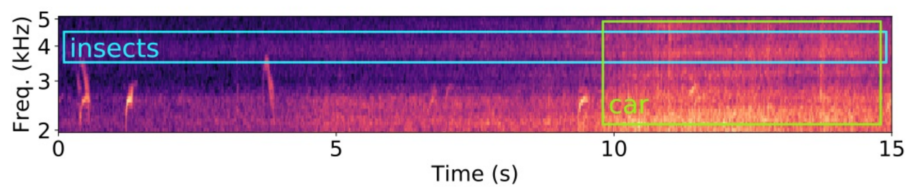
Fig. 8

Sound Event Detection
Pipeline

- Feature extraction
    - 1D features (audio samples) → "end-to-end learning"
    - 2D features (mel-spectrogram, STFT)
- Feature pre-processing
    - Log-magnitude scaling

Sound Event Detection
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. 9

Sound Event Detection
Pipeline

- Annotation
    - Quality of “ground truth”? (limited agreement / reliability)

Sound Event Detection
Pipeline

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

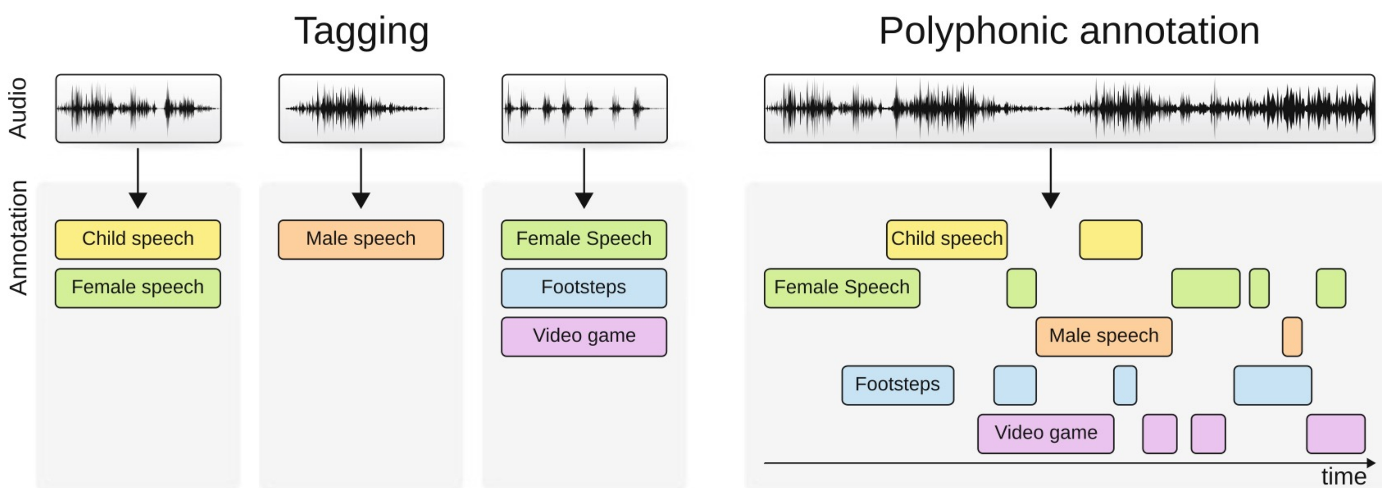


Fig. 10

Sound Event Detection
Pipeline

- Label encoding
    - Binarized sound activity (0/1)
        - Multilabel classification
        - 1 (independent) binary detector per class

Sound Event Detection
Pipeline

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

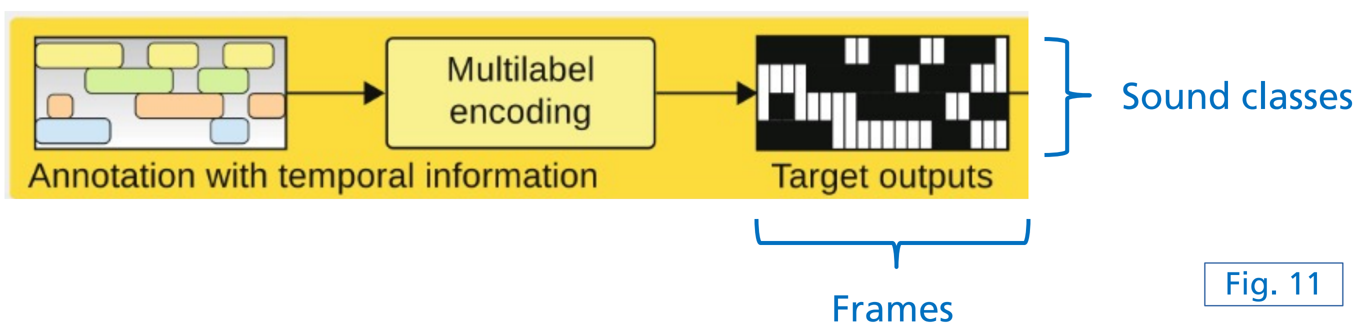
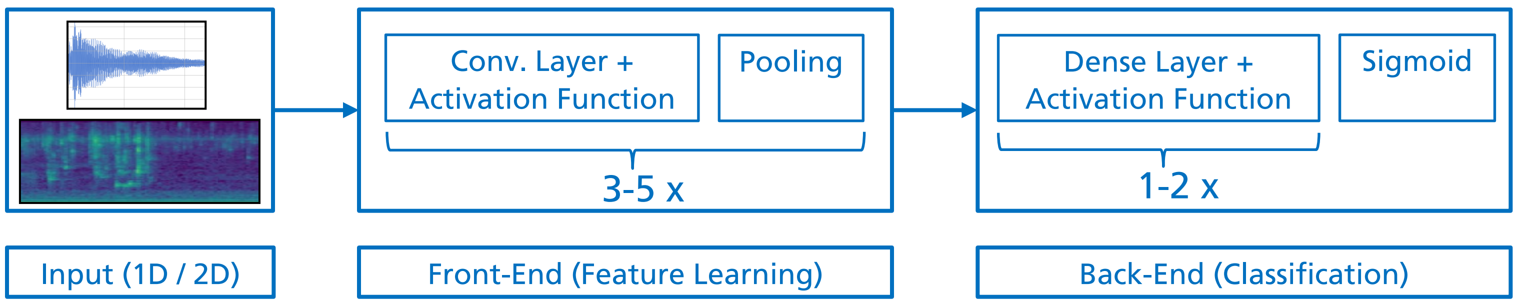


Fig. 11

Sound Event Detection
Pipeline

- Typical neural network architectures
  - CNN

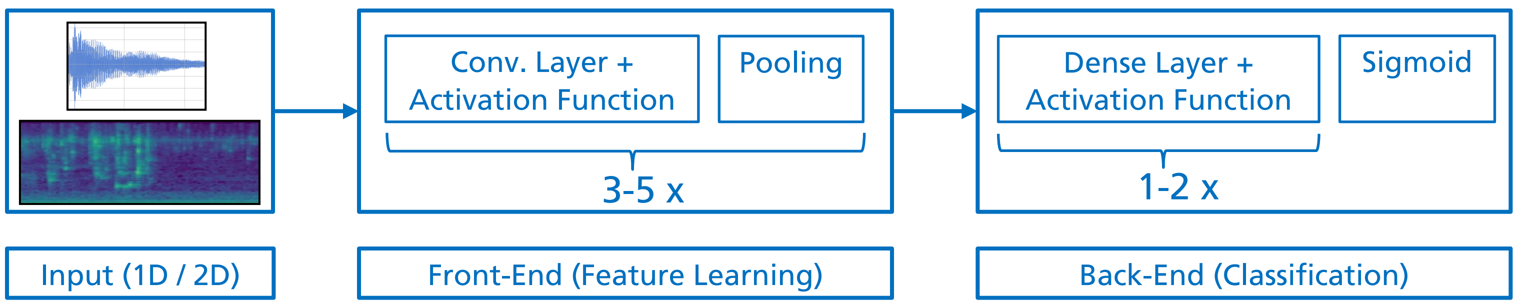
Own



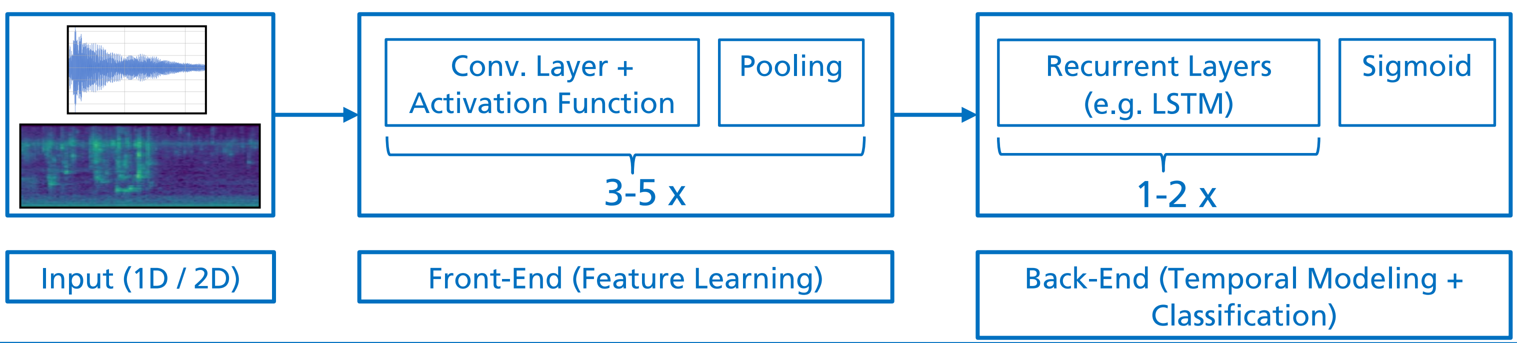
Sound Event Detection
Pipeline

- Typical neural network architectures
  - CNN

Own



  - CRNN



Sound Event Detection
Pipeline

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

Sound Event Detection
Pipeline

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

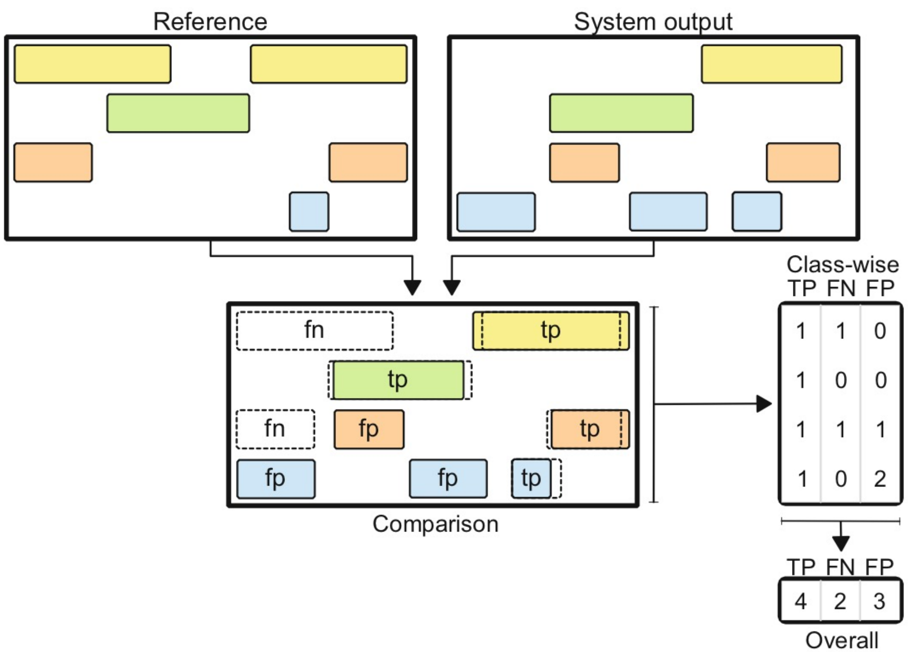


Fig. 13

Sound Event Detection
Evaluation Metrics

- Recap: Binary classification evaluation
    - True/false positives (TP/FP)
    - True/false negatives (TN/FN)

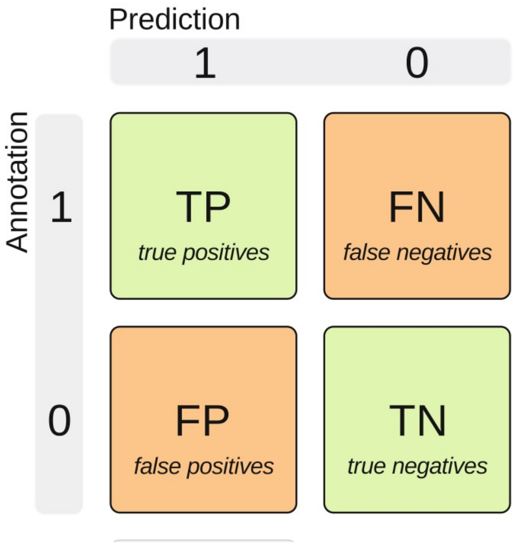


Fig. 12

Sound Event Detection
Evaluation Metrics

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

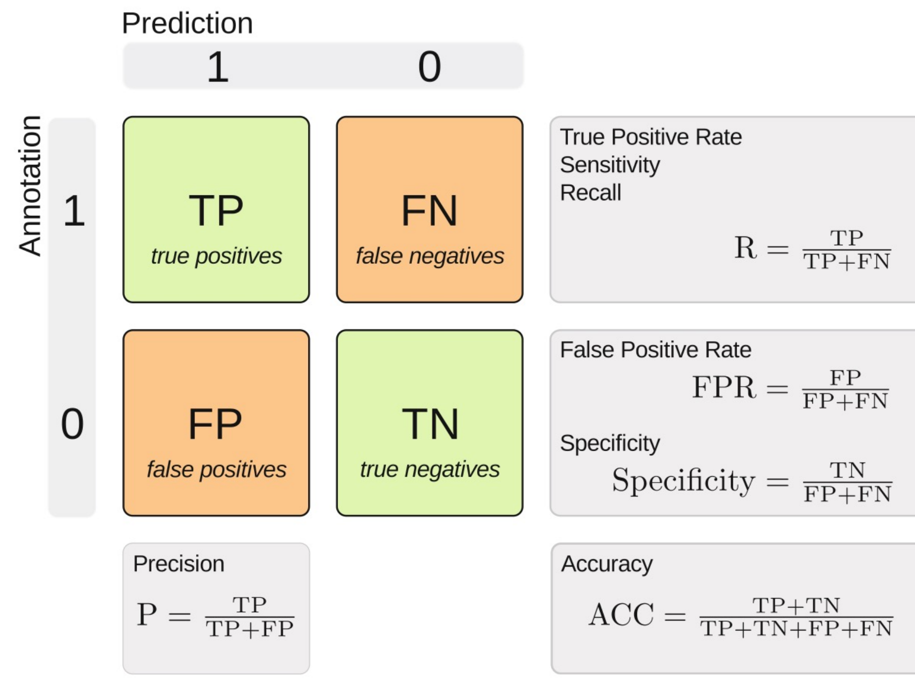


Fig. 12

Sound Event Detection
Data Augmentation

- Data Augmentation
    - Increases amount / variability of training data
    - Improves model generalization towards unseen data

Sound Event Detection
Data Augmentation

- Data Augmentation
    - Increases amount / variability of training data
    - Improves model generalization towards unseen data
- Methods
    - Audio signal transformations
        - Time stretching, pitch shifting, dynamic range compression

Sound Event Detection
Data Augmentation

- Data Augmentation
  - Increases amount / variability of training data
  - Improves model generalization towards unseen data
- Methods
  - Audio signal transformations
    - Time stretching, pitch shifting, dynamic range compression
  - SpecAugment [Park, 2019]
    - Temporal warping (1)
    - Block-wise masking (2)

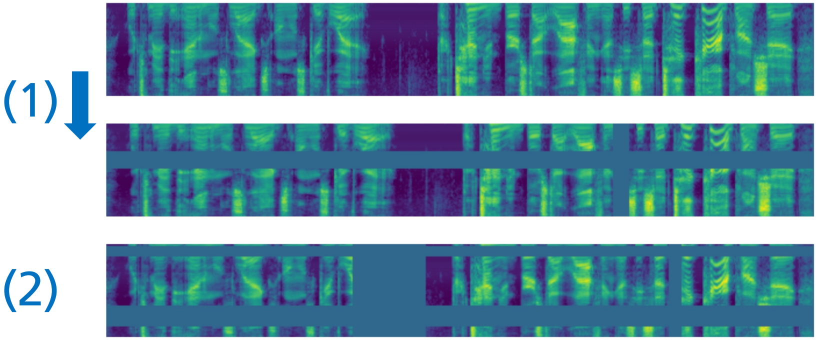


Fig. 19

Sound Event Detection
Data Augmentation

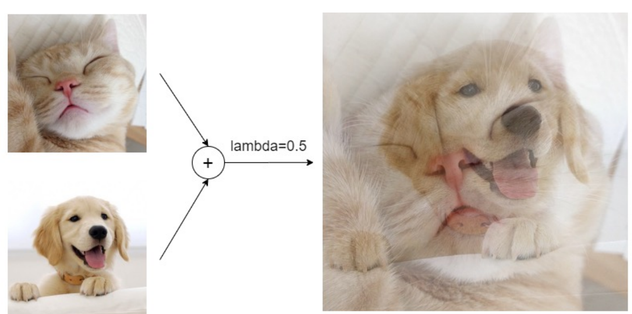
- Methods
  - Mix-up data augmentation [Zhang, 2018]
    - Simulate sound mixtures
    - Mix two data instances with random mixing ratio

$$x = \alpha \cdot x_1 + (1 - \alpha) \cdot x_2$$
$$y = \alpha \cdot y_1 + (1 - \alpha) \cdot y_2$$

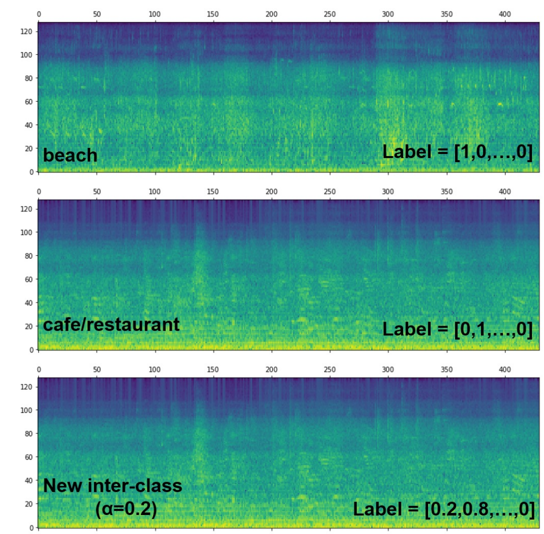
Sound Event Detection
Data Augmentation

- Methods
  - Mix-up data augmentation [Zhang, 2018]
    - Simulate sound mixtures
    - Mix two data instances with random mixing ratio

$$x = \alpha \cdot x_1 + (1 - \alpha) \cdot x_2$$
$$y = \alpha \cdot y_1 + (1 - \alpha) \cdot y_2$$



Computer Vision Fig. 17

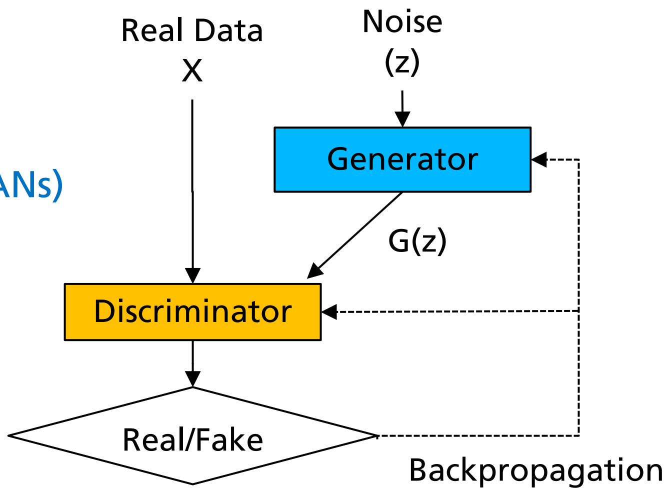


Machine Listening Fig. 18

Sound Event Detection
Data Augmentation

- Methods
  - Data Synthesis
    - Example: WaveGAN [Donahue, 2019]
      - Synthesize waveforms with Generative Adversarial Networks (GAN)

Training (GANs)



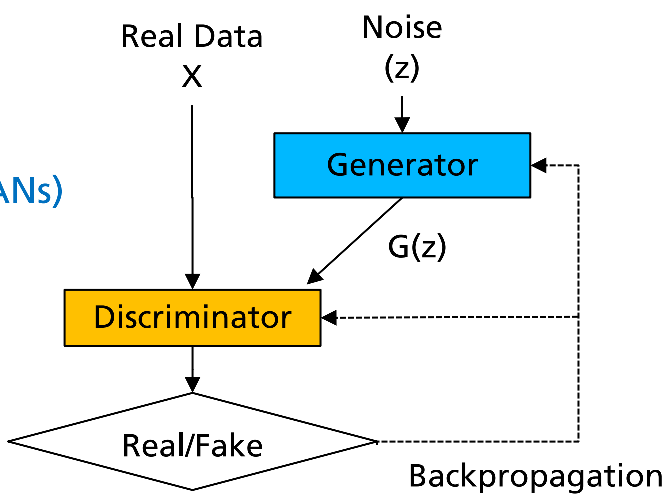
Own

Fig. 20

Sound Event Detection
Data Augmentation

- Methods
  - Data Synthesis
    - Example: WaveGAN [Donahue, 2019]
      - Synthesize waveforms with Generative Adversarial Networks (GAN)

Training (GANs)



Own

Synthesis examples (Spectrograms)

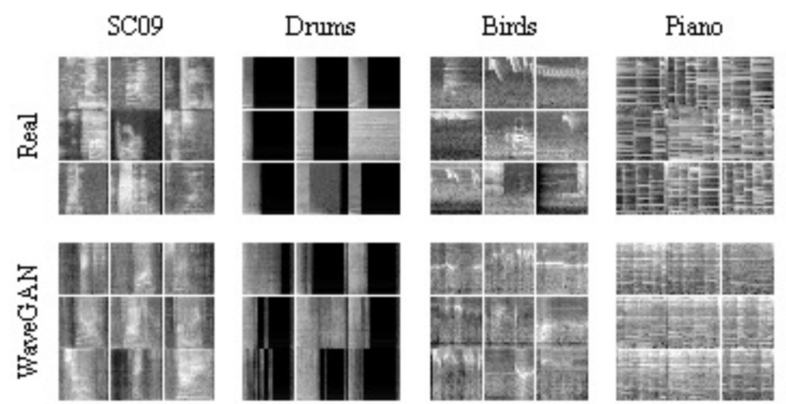


Fig. 20

Sound Event Detection
Novel Methods

- VGG-style CNN [Sakashita, 2018]
    - Main Idea
        - Pairs of convolutional layers + non-linearity before max pooling
    - Effect
        - Smaller kernel shapes
        - More non-linearities → model is more expressive

Sound Event Detection
Novel Methods

- CRNN [Adavanne, 2017]

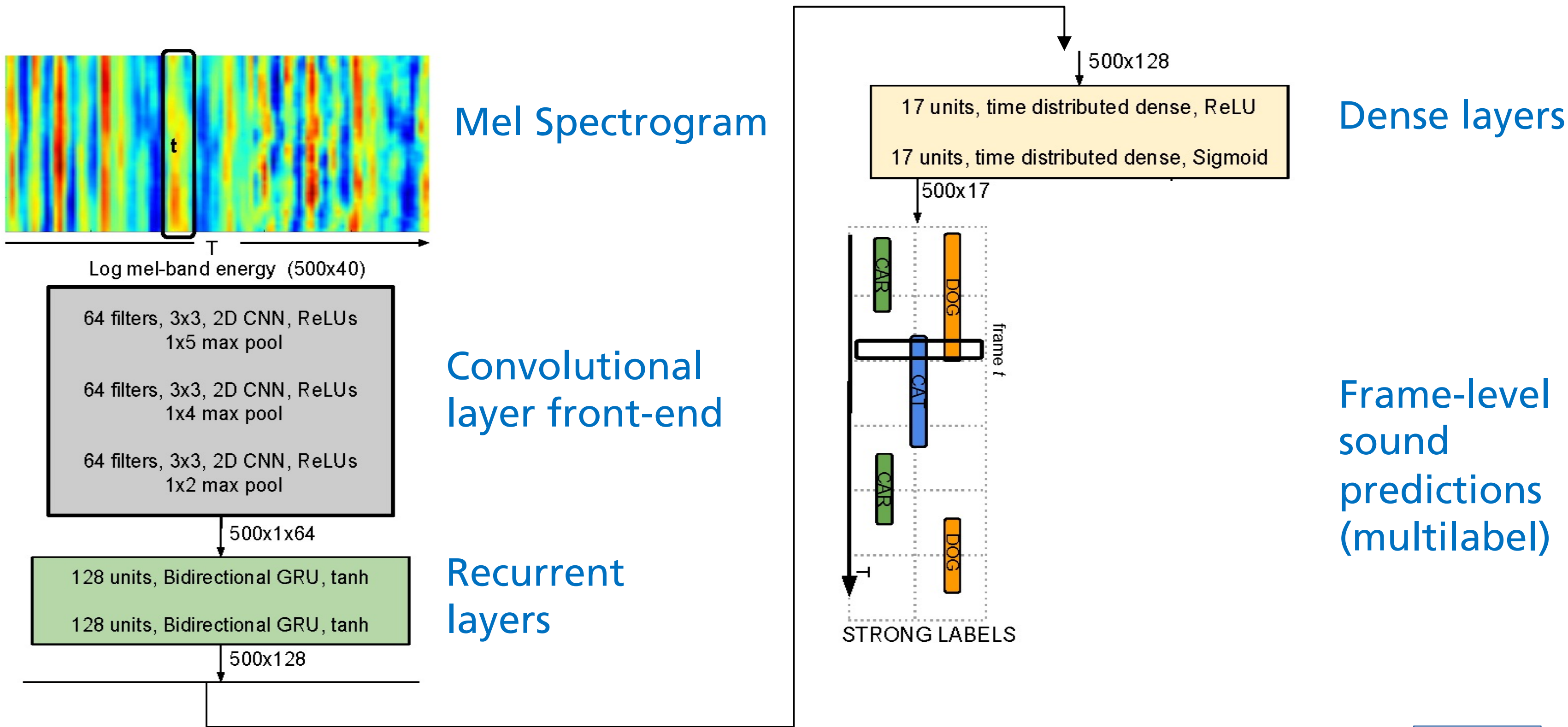
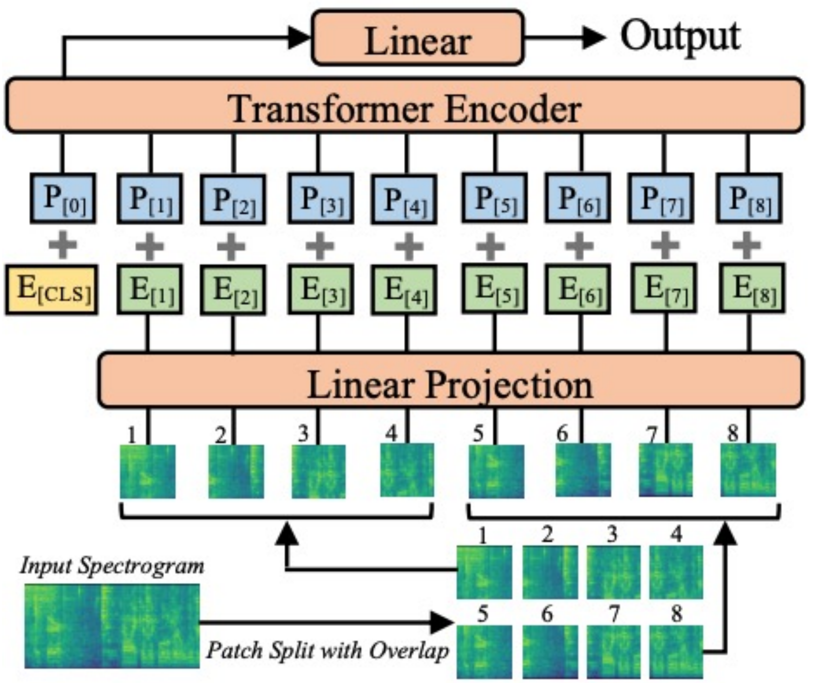


Fig. 23

Sound Event Detection
Novel Methods

- Audio Spectrogram Transformer (AST) [Gong, 2021]
  - Spectrogram patches mapped to embedding sequence
  - Self-attention (model longer time dependencies)
- State-of-the-art on sound event tagging



Summary

- Introduction
- Sound Event Detection
    - Introduction
    - Challenges & Related Tasks
    - Pipeline
    - Evaluation Metrics & Datasets
    - Data Augmentation
    - Methods
        - Traditional
        - Neural Network Based

References

Abeßer, J. (2021). USM-SED - A Dataset for Polyphonic Sound Event Detection in Urban Sound Monitoring Scenarios. *Submitted to the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*. Barcelona, Spain.

Adavanne, S., & Virtanen, T. (2017). Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network. *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*. Munich, Germany.

Donahue, C., Mcauley, J., & Puckette, M. (2019). Adversarial Audio Synthesis. *Proceedings of the International Conference on Learning Representations (ICLR)*, 1–16. New Orleans, LA, USA.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. Las Vegas, NV, USA.

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.

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2613–2617. Graz, Austria.

Sakashita, Y., & Aono, M. (2018). Acoustic scene classification by ensemble of spectrograms based in adaptive temporal division. In *Detection and Classification of Acoustic Scenes and Events (DCASE)*.

References

Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio Spectrogram Transformer. arXiv:2104.01778

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

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

Xu, K., Feng, D., Mi, H., Zhu, B., Wang, D., Zhang, L., … Liu, S. (2018). Mixup-Based Acoustic Scene Classification Using Multi-Channel Convolutional Neural Network. *Proceedings of the Pacific Rim Conference on Multimedia (PCM)*, 14–23. Hefei, China.

Xu, Y., Kong, Q., Wang, W., & Plumbley, M. D. (2018). Large-Scale Weakly Supervised Audio Classification Using Gated Convolutional Neural Network. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 121–125. Calgary, AB, Canada.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond Empirical Risk Minimization. *Proceedings of the International Conference on Learning Representations (ICLR)*. Vancouver, Canada.

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2017). Random Erasing Data Augmentation. *ArXiv Preprint ArXiv:1708.04896*.

Images

Fig. 1: https://ccsearch-dev.creativecommons.org/photos/39451123-ee45-4ec3-ad8d-b42d856bca06

Fig. 2: https://ccsearch-dev.creativecommons.org/photos/c69d3b07-76bd-43e2-a44e-8742edc8447a

Fig. 3: https://ccsearch-dev.creativecommons.org/photos/ab3062ab-fe0f-420d-b93d-7451db166b4e

Fig. 4: [Virtanen, 2018], p. 15, Fig. 2.1

Fig. 5: https://urbansounddataset.weebly.com/uploads/4/3/9/4/4394963/3427002_orig.png

Fig. 6: [Virtanen, 2018], p. 157, Fig. 6.3

Fig. 7: https://towardsdatascience.com/whats-wrong-with-spectrograms-and-cnns-for-audio-processing-311377d7ccd

Fig. 8: Virtanen et al., Computational Analysis of Sound Scenes and Events, p. 31, Fig. 2.11

Fig. 9: [Lostanlen, 2019], p. 1, Fig. 1

Fig. 10: [Virtanen, 2018], p. 154, Fig. 6.2

Fig. 11: [Virtanen, 2018], p. 31, Fig. 2.11 (excerpt)

Fig. 12: [Virtanen, 2018], p. 170, Fig. 6.7

Fig. 13: [Virtanen, 2018], p. 169, Fig. 6.6

Fig. 14: http://dcase.community/challenge2019/task-sound-event-localization-and-detection, Fig. 1

Images

Fig. 15: [Virtanen, 2018] , p. 267, Fig. 9.7

Fig. 16: http://dcase.community/challenge2020/task-sound-event-detection-and-separation-in-domestic-environments, Fig. 2

Fig. 17: https://miro.medium.com/max/955/1*XqyD5OE47AdqeR6KeMg9FQ.png

Fig. 18: [Xu, Feng, et al., 2018], p. 17, Fig. 2

Fig. 19: [Park, 2019], p. 2614, Fig. 2

Fig. 20: [Donahue, 2019], p. 5, Fig. 4

Fig. 21: [Abeßer, 2021], p. 3, Fig. 2

Fig. 23: [Adavanne, 2017], p. 2, Fig. 1

Fig. 24: [Xu, Kong, et al., 2018], p. 2, Fig. 1

Fig. 24: [He, 2015], p. 2, Fig. 2

Fig. 25: https://miro.medium.com/max/1400/1*Voah8cvrs7gnTDf6acRvDw.png

Sounds

AUD-1: https://freesound.org/people/{InspectorJ/sounds/416529, prometheus888/sounds/458461, MrAuralization/sounds/317361}

AUD-2: https://freesound.org/people/G_M_D_THREE/sounds/424404/

AUD-3: https://freesound.org/people/IFartInUrGeneralDirection/sounds/96195/

AUD-4: https://freesound.org/people/InspectorJ/sounds/400860/

AUD-5: https://freesound.org/people/Simon%20Spiers/sounds/516876/

AUD-6: USM-SED dataset [Abeßer, 2021], Evaluation Set, Sound ID 2417

AUD-7: USM-SED dataset [Abeßer, 2021], Evaluation Set, Sound ID 1930

AUD-8: USM-SED dataset [Abeßer, 2021], Evaluation Set, Sound ID 339

Thank you!

- Any questions?

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

https://www.machinelistening.de