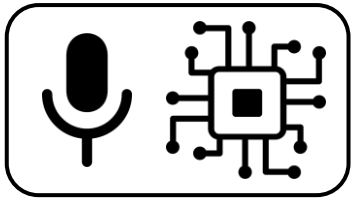
Computational Analysis of Sound and Music

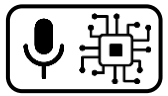


Environmental Sound Analysis – Acoustic Scene Classification

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

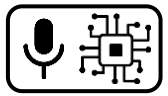
jakob.abesser@idmt.fraunhofer.de



Acoustic Scene Classification

Outline

- Introduction & Application Scenarios
- Traditional Approaches
- Deep Learning-based Approaches
- Current research topics
    - Domain adaptation
    - Efficient models



Acoustic Scene Classification

Introduction

- Acoustic scene classification (ASC)
  - Multi-class (1 of N) classification scenario
  - Summative label (tagging)
- Common Classes
  - Indoor
    - Airport, shopping mall, metro station
  - Outdoor
    - Pedestrian street, urban park, traffic
  - Transportation
    - Travelling by bus / metro / tram

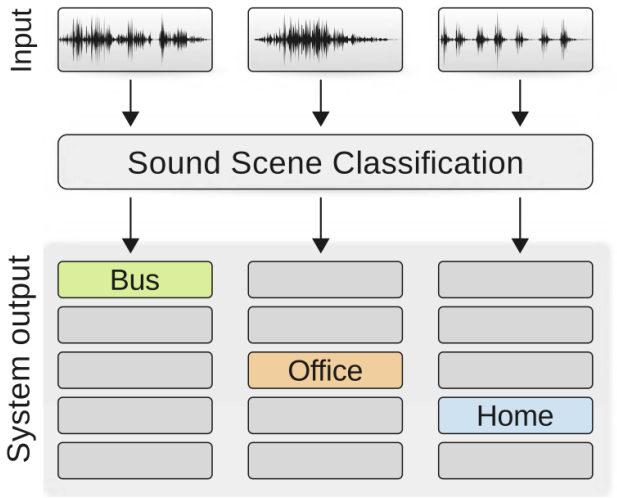


Fig-E3-1

Acoustic Scene Classification

Introduction

- Interdependence between sound events and acoustic scenes
- Acoustic scene
    - Typical set of sounds
    - Example: Office
        - Keyboard clicks
        - Human conversations
        - Printer
        - Air conditioner

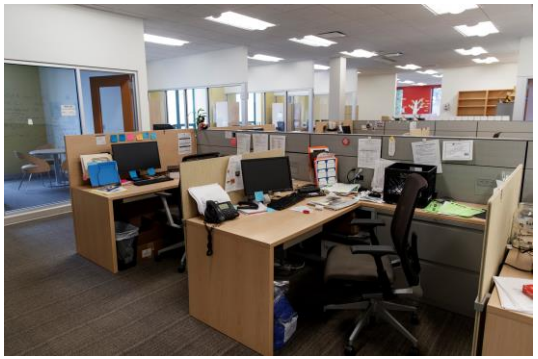


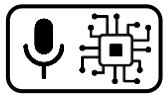
Fig-E3-2

Aud-E3-1

Acoustic Scene Classification

Application Scenarios

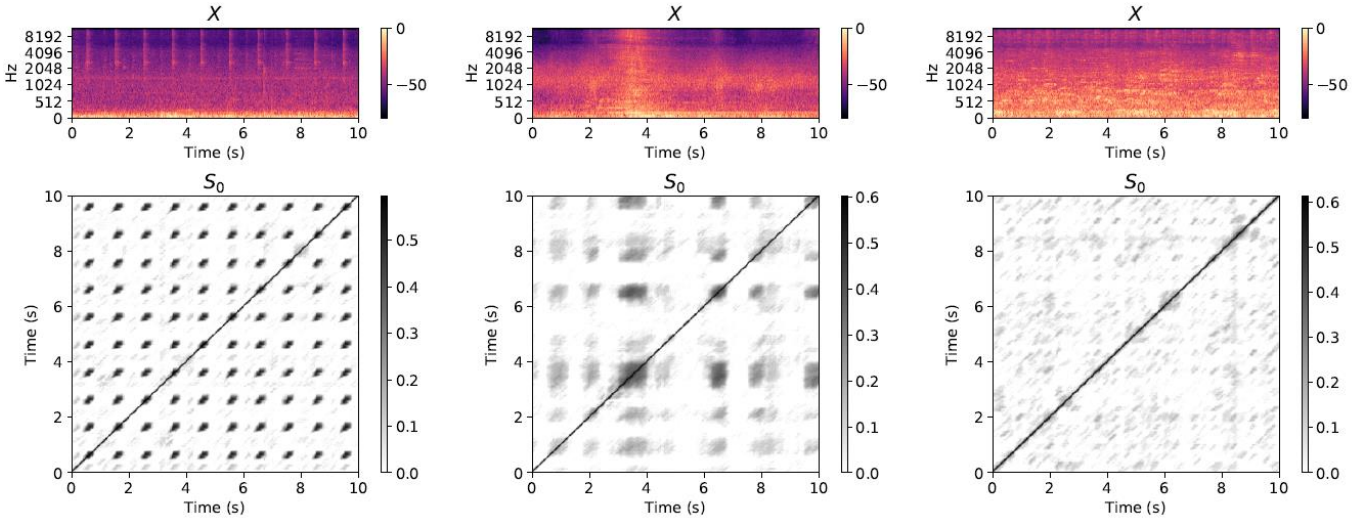
- Context-aware devices (hearables, cell phones)
- Smart cities (improve city planning, traffic management, and public safety)
- Content analysis (indexing and organizing multimedia content)
- Human-computer interaction (natural interaction with devices through voice commands and ambient sound recognition)
- Healthcare (monitoring patients' acoustic environments in hospitals or homes)



Acoustic Scene Classification

Traditional Approaches

- Timbre-related features (MFCC, Mel Spectrogram)
- Classification algorithms (SVM, GMM)
- Recurrence Quantification Analysis (RQA) [Roma et al., 2014]
  - Measure sound repetitively in acoustic scenes



Regular repetitions (beeps from tram doors)

Irregular repetitions (passing cars)

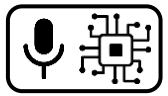
No repetitions (shopping mall)

Fig-E3-3

Acoustic Scene Classification

Deep Learning-based Approaches

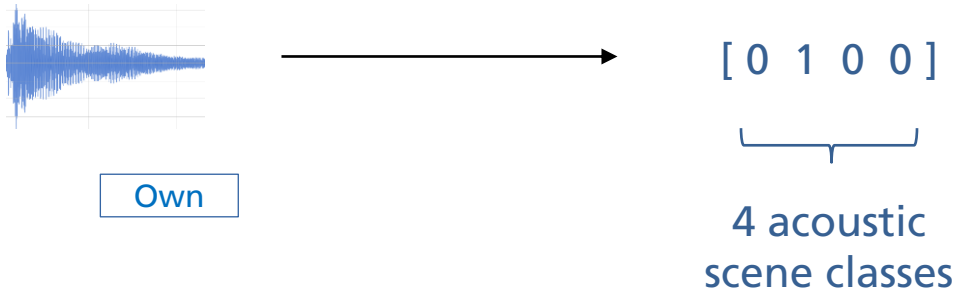
- General design choices
  - CNN & CRNN models (similar to SED)
  - Temporal result aggregation (pooling) within network
  - Final layer: Softmax activation function (multiclass classification)
  - Data Augmentation
    - Mixup
    - SpecAugment
  - Ensemble models



Acoustic Scene Classification

Deep Learning-based Approaches

- Label encoding
  - One-hot-encoded (global) target
- Example
  - 4 scene classes (bus, office, home, forest)
  - Encoding of an office recording



Acoustic Scene Classification

Deep Learning-based Approaches

- Example 1: [McDonnel & Gao, 2020]
  - Based on ResNet architectures
  - Mel spectrogram split into low/high frequencies
  - Late fusion
  - Feature: Mel spectrogram + Δ + Δ Δ

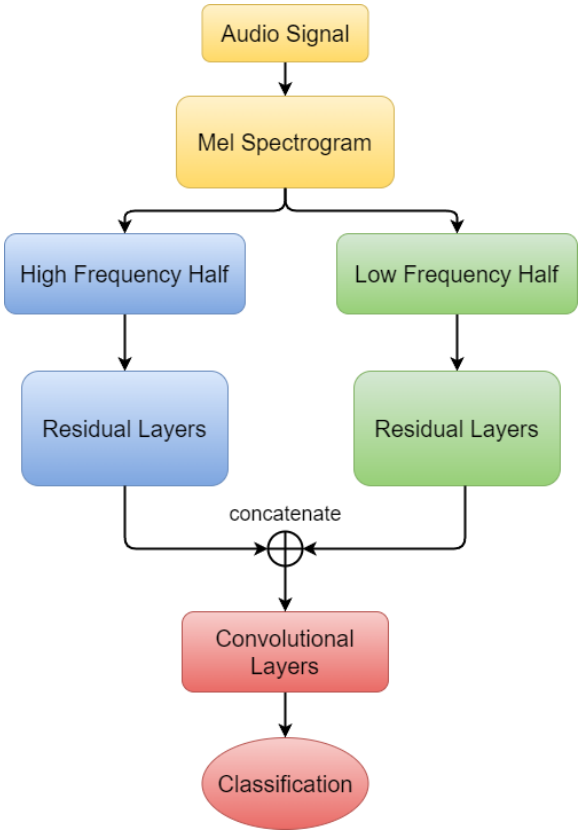


Fig-E3-4

Acoustic Scene Classification

Deep Learning-based Approaches

- Example 2: [Koutini et al. 2019]
  - Modifications of residual block (improved stability and robustness)

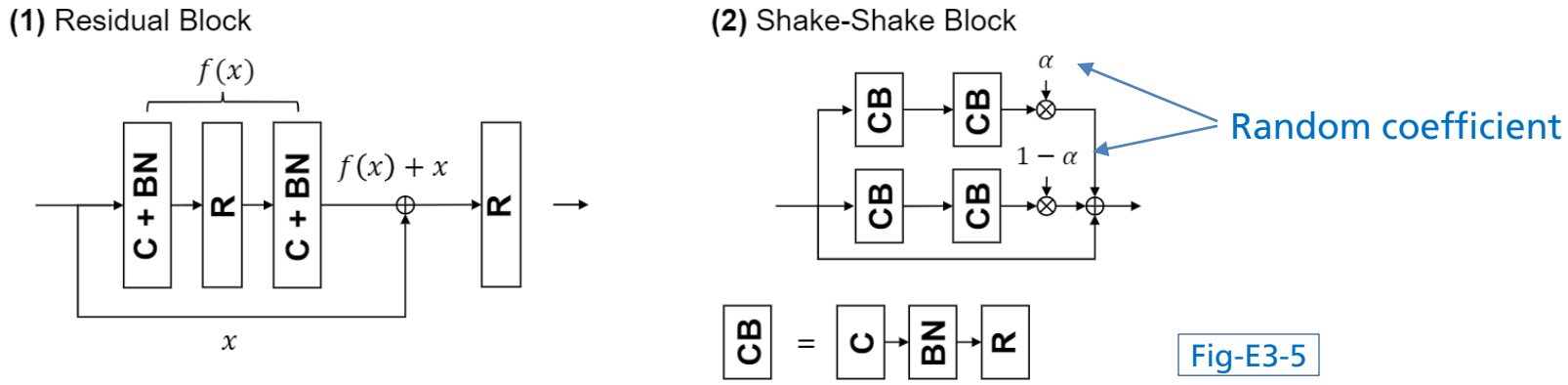


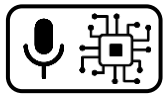
Fig-E3-5

  - Frequency-aware CNN
    - Additional channel with normalized frequency between 0 and 1

Acoustic Scene Classification

Domain Adaptation

- Domain shift differences in data distribution due to
    - Room acoustics (reverb, reflections)
    - Microphone characteristics (frequency response, directionality)
- Domain adaptation
    - Align source and target data distributions
        - Unsupervised: adversarial training [Gharib, 2018]
        - Supervised: transfer learning
- Approaches
    - Data augmentation
    - Data normalization [Johnson, 2020] [Latifi, 2023]



Acoustic Scene Classification

Domain Adaptation

- Domain adaptation (DA)
  - Unsupervised DA via adversarial training [Gharib, 2018]

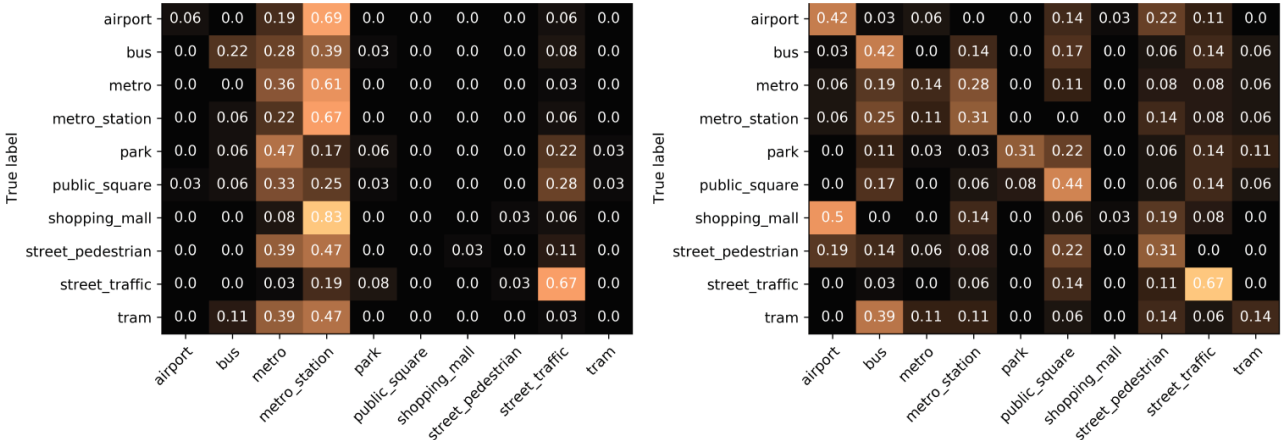


Fig-E3-6

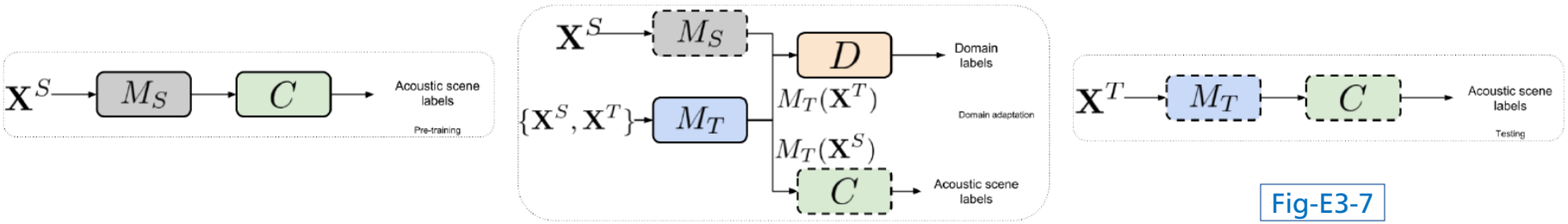


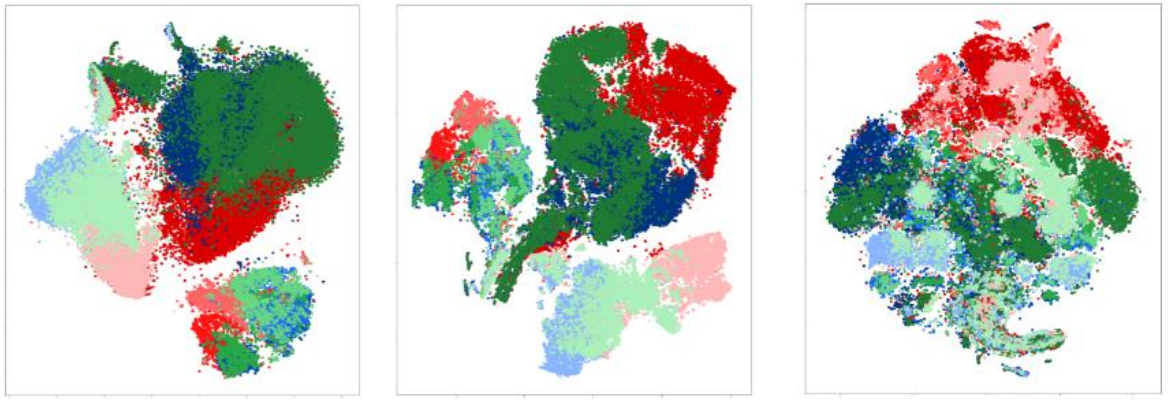
Fig-E3-7

Acoustic Scene Classification

Domain Adaptation

- Data normalization
  - Align source and target data distribution (zero mean & standard deviations) [Johnson, 2020]
  - Reduce domain shift

Metal ball surface classification (colors = classes, shadings = recordings)



(a) No Norm (b) Global Norm (c) Adaptive Norm

Fig-E3-8

Acoustic Scene Classification

Efficient Models

- Goals
  - Reduce model size – fewer parameters, less memory required
  - Reduce latency (inference time) / lower energy consumption

- Approaches ([Wang, 2021])
  - Model pruning
  - Identify & remove redundant connections / neurons

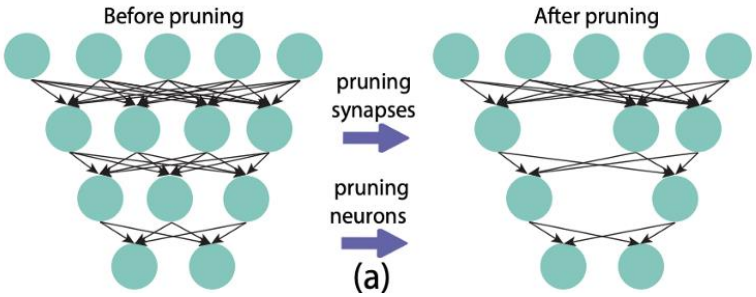


Fig-E3-9

Acoustic Scene Classification

Efficient Models

- Approaches
  - Quantization
    - Reduce numeric precision while minimize information loss
    - Ex.: 32-bit floating point -> 8-bit fixed point (256 values)
  - Reduce memory footprint of network weights
- Low-rank tensor decompositions
  - Replace (many) redundant filters by a linear combination of fewer filters
- Knowledge Distillation
  - Transfer knowledge from complex (teacher) to simpler (student) model

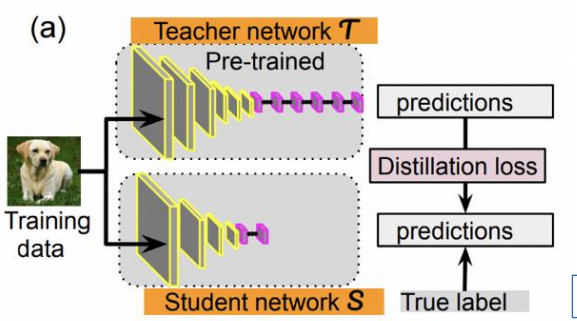


Fig-E3-10

Programming session

Fig-A2-13

References

Images

Fig-E2-1: [Virtanen, 2018], p. 267, Fig. 9.7

Fig-E2-2: https://images.theconversation.com/files/349387/original/file-20200724-15-ldrybi.jpg

Fig-E2-3: [Abeßer, 2024], p. 8, Fig. 2

Fig-E2-4: Zhiwei Liang, Final Presentation, Master Thesis “Self-Similarity Based Representations of Soundscape Recordings for Acoustic Scene Classification”, TU München (2023)

Fig-E2-5: Own

Fig-E2-6: [Gharib, 2018], p. 3., Fig. 2 (a) & (b)

Fig-E2-7: [Gharib, 2018], p. 2., Fig. 1

Fig-E2-8: [Johnson & Grollmisch, 2021], p. 82, Fig. 1

Fig-E2-9: https://miro.medium.com/max/955/1*C3rR1-qzZfgYE_QA7WvLOQ.png

Fig-E2-9: [Wang, 2021], p. 2, fig. 1 (a)

Fig-E2-10: https://dcase.community/images/tasks/challenge2019/task1_acoustic_scene_classification_openset.png

References

References

Roma, G., Nogueira, W., & Herrera, P. (2014). IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events: Recurrence Quantification Analysis Features for Auditory Scene Classification.

Abeßer, J., Liang, Z., & Seeber, B. (2024). Sound Recurrence Analysis for Acoustic Scene Classification. *Under review*

McDonnell, M. D., & Gao, W. (2020). Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 141–145).

Koutini, K., et al. (2019). CP JKU submissions to DCASE'19: Acoustic scene classification and audio tagging with receptive field regularized CNNs. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019).

Bidarouni, A. L. & Abeßer, J. (2023). Unsupervised Feature-Space Domain Adaptation applied for Audio Classification. Proceedings of the International Symposium on the Internet of Sounds (IoS-RN)

Johnson, D. S., & Grollmisch, S. (2021). Techniques improving the robustness of deep learning models for industrial sound analysis. Proceedings of the European Signal Processing Conference (EUSIPCO), 81–85. Amsterdam, The Netherlands.

Gharib, S., Drossos, K., Emre, C., Serdyuk, D., & Virtanen, T. (2018). Unsupervised Adversarial Domain Adaptation for Acoustic Scene Classification. Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE). Surrey, UK.

References

Audio

Aud-E3-1: 16HPanskaTyllova_Terezie - 17-1_atmosphere of office.wav (2019) - CC0 License, https://freesound.org/people/16HPanskaTyllova_Terezie/sounds/497363