PADERBORN UNIVERSITY

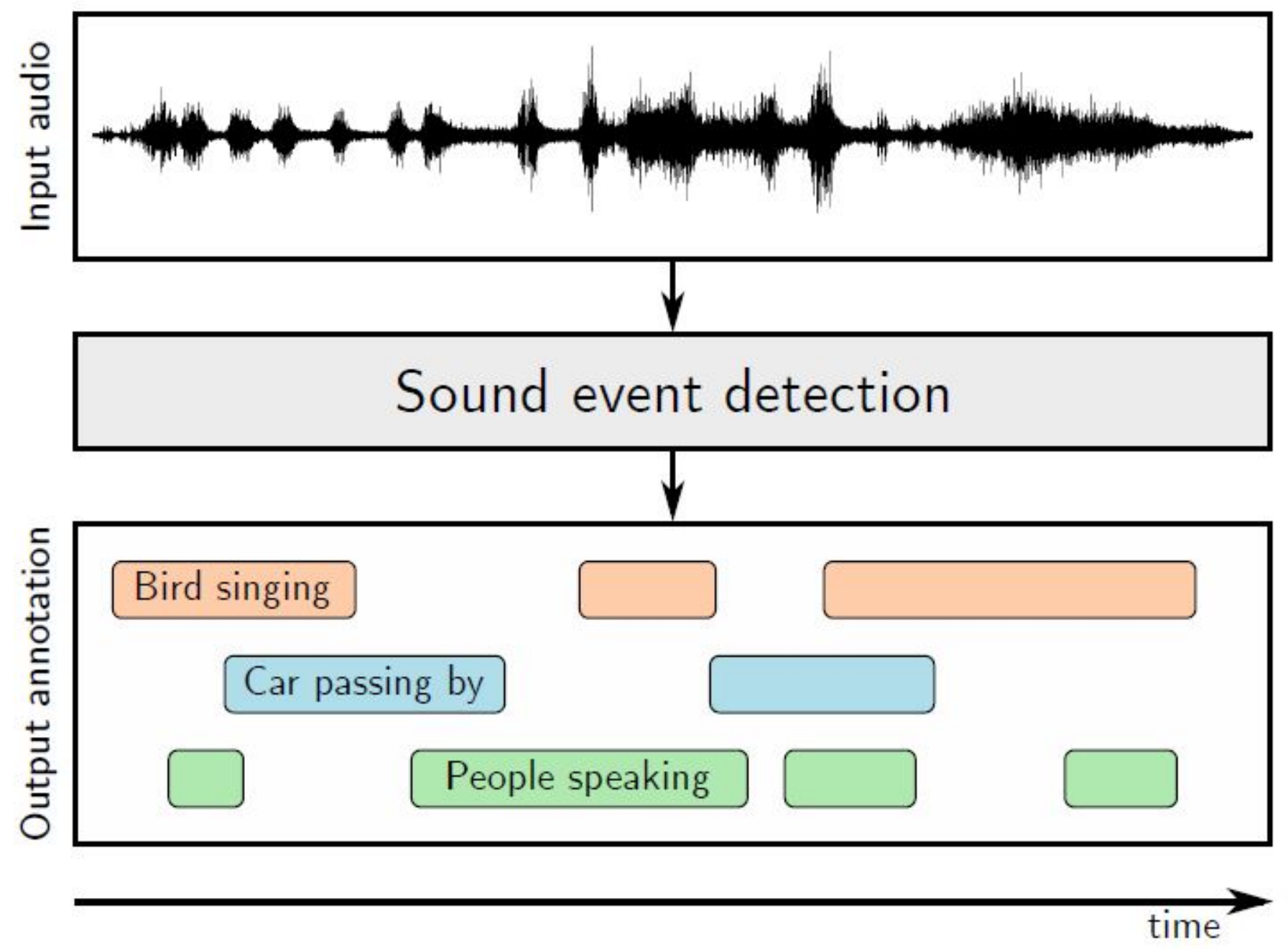
A Study on Transfer Learning for Acoustic Event Detection in a Real Life Scenario

Prerna Arora and Reinhold Haeb-Umbach, Paderborn University, Germany
{arora, haeb}@nt.uni-paderborn.de, http://nt.uni-paderborn.de

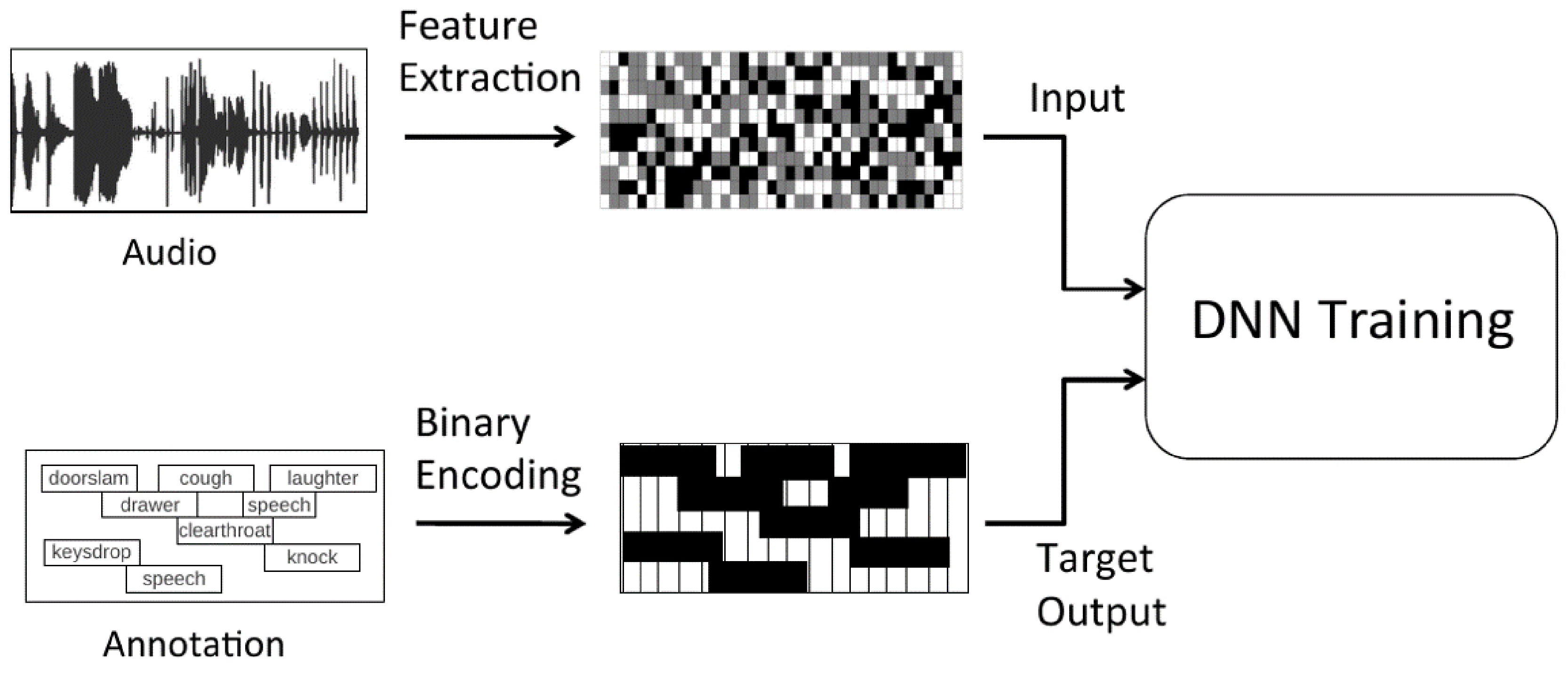
Computer Science, Electrical Engineering and Mathematics
NT Communications Engineering
Prof. Dr.-Ing. Reinhold Häb-Umbach

Introduction

- **Acoustic Event** : Distinct segment of audio that a listener can consistently label
- **Acoustic Event Detection (AED)**: Locate in time (detection) and identify (classification)
- Related but different to **Speech Recognition**
  - Huge variety of sounds and applications
  - Polyphony
  - Lack of labeled training data



Supervised AED

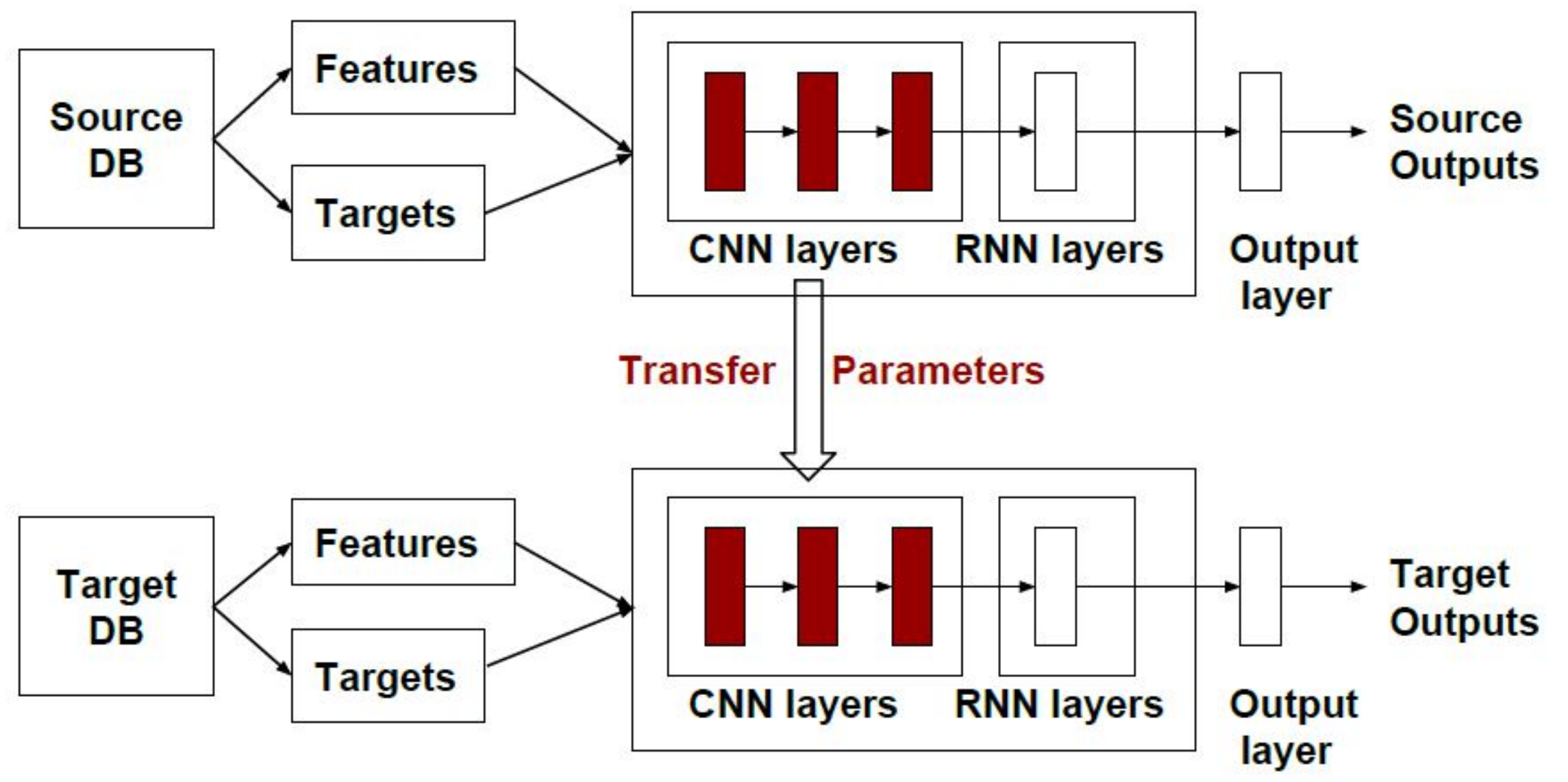


Proposed Approach

- **Transfer Learning (TL)**: Transfer knowledge from **Source** domain to **Target** domain
- **Hypothesis**:
  - Audio Events are made up of acoustic units (AUs) as universal building blocks
  - The order in which they appear distinguishes one event from another
  - Similar to Phonemes in Speech Recognition

    Washing Dishes
    $\rightarrow /RunningWater/ - /Utensils/ - /Scrubbing/$
- **Approach**: Learn the AUs from source events and utilize them to learn the target events which may share some or all of the learned AUs
- **Databases**:
  - **Source**: TUT-SED **Synthetic** 2016 [1], 566 minutes, clean, 16 events $\in$ alarms and sirens, baby crying, bird singing, bus, cat meowing, crowd applause, etc.
  - **Target**: TUT-SED **Real** 2016 [2], 78 minutes, noisy, 17 events $\in$ bird singing, car passing by, cutlery, washing dishes, alarms, mixer, rain, etc.

TL with Convolutional Recurrent NN



- Learn spectral characteristics of events using CNN, while RNN captures time dependencies
- Transfer only CNN layers to the target model, dropping RNN and output layers
- Target model trained in 3 settings:
  - **Frozen All**: all layers frozen
  - **Frozen One**: first layer frozen
  - **Finetune All**: all layers finetuned

References

[1] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, ”Convolutional recurrent neural networks for polyphonic sound event detection”, in arXiv, 2017
[2] A. Mesaros, T. Heittola, and T. Virtanen, ”TUT database for acoustic scene classification and sound event detection”, in EUSIPCO, 2016

Results

- Error rate and F-measure for target real-life database

Model	Sub	Del	Ins	AEER	F-m (%)
State of the Art [1]	-	-	-	0.95	30.3
Baseline (No TL)	0.2	0.43	0.38	1.01	37.8
Frozen All	0.13	0.59	0.22	**0.94**	34.3
Frozen One	0.25	0.32	0.56	1.13	**38.2**
Finetune All	0.22	0.4	0.4	1.02	37.9

- Comparison between specific target events

Event	Model	AEER	F-m (%)
Bird Singing	Baseline (No TL)	1.24	50.7
(large overlap)	Frozen All	**1.02**	54.4
Washing Dishes	Baseline (No TL)	1.51	25.5
(marginal overlap)	Frozen One	1.54	**40.4**
Car Passing by	Baseline (No TL)	**0.98**	**58.0**
(no overlap)	Finetune All	0.991	56.7

Conclusions & Outlook

- Initial hypothesis could only partially be verified
- Probable cause: source DB too small
- Outlook: use Google Audioset as source DB

IEEE Workshop on Multimedia Signal Processing 2017, Luton

FOR 2457 Acoustic Sensor Networks
Work supported by DFG under contract no. <Ha 3455/15-1>
Research Unit FOR2457 “Acoustic Sensor Networks”