A-CRNN: A Domain Adaptation Model for Sound Event Detection

Wei Wei, Hongning Zhu, Emmanouil Benetos, and Ye Wang

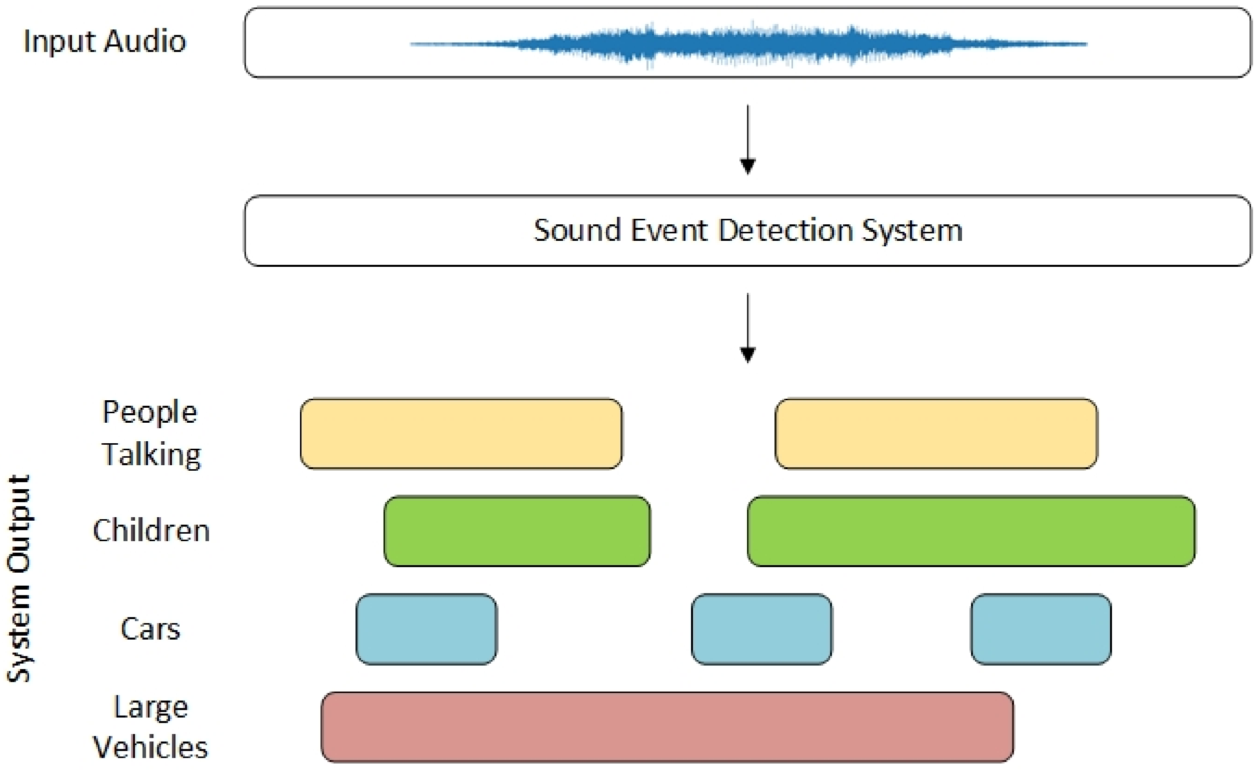
ICASSP 2020

Problem to Address

Domain adaptation for sound event detection

Overview

- Sound event detection
  - Start time
  - End time
  - Label



Overview

- Sound event detection
  - Start time
  - End time
  - Label

- Domain Adaptation
  - Mismatch between datasets

Domain Adaptation

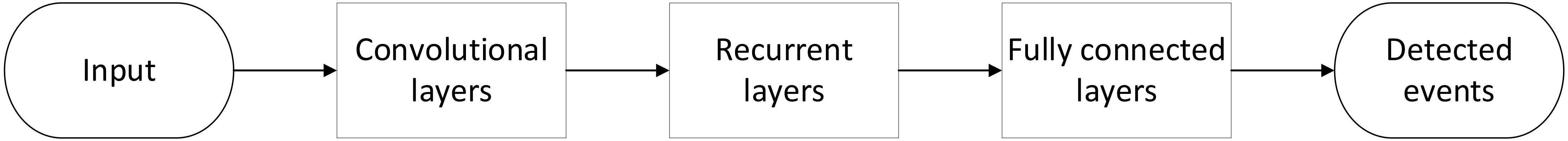
- Source domain
  - Labeled data

Domain Adaptation

- Source domain
  - Labeled data
- Target domain
  - Unlabeled data

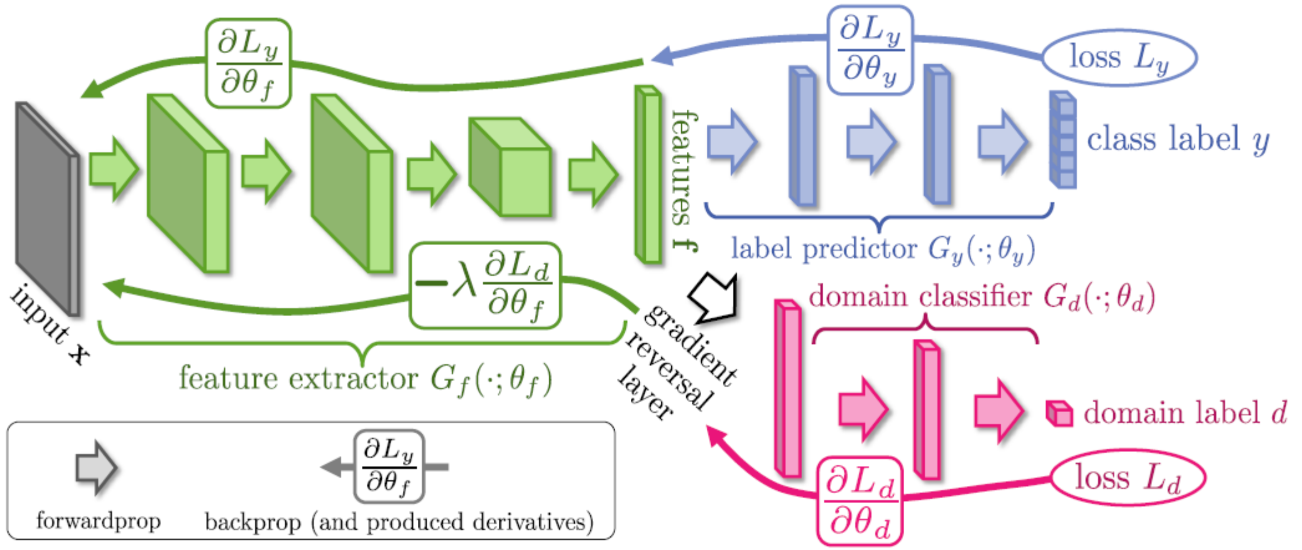
Related Work

- CRNN
  - Convolutional recurrent neural network (Adavanne et al., 2017)
  - State-of-the-art sound event detection system



Related Work

- Adversarial-based domain adaptation models
  - Introducing domain discriminators to perform adversarial training
  - (Ganin et al., 2015)



SG Dataset

- Motivation
  - Most of the datasets are recorded in Europe
  - No existing dataset for sound event detection focuses on the domain adaptation problem

SG Dataset

- Motivation
  - Most of the datasets are recorded in Europe
  - No existing dataset for sound event detection focuses on the domain adaptation problem
- Basic information
  - 3 to 5 minutes each
  - 9 hours in total
  - Collected around university campus in Singapore

SG Dataset

- Event classes
  - car
  - children
  - large vehicle
  - people speaking
  - people walking

- Same as the DCASE dataset
  - Task 3 of DCASE 2017 challenge

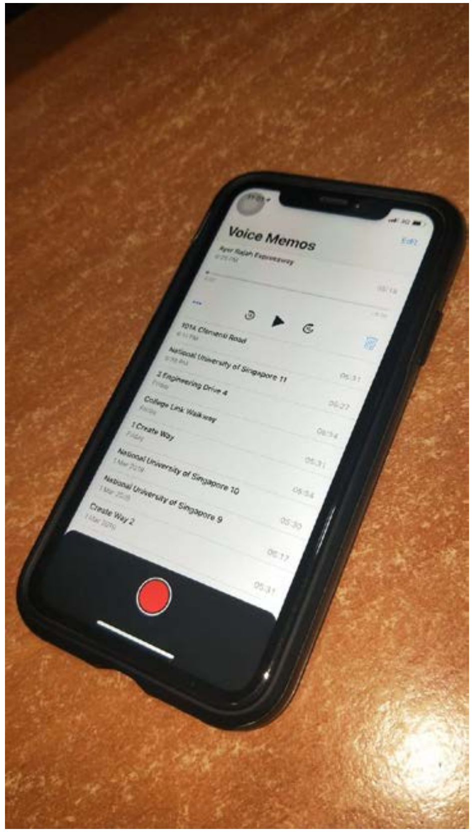
SG Dataset

- Recording equipment - high quality
  - Roland CS-10EM

  - Zoom H5

SG Dataset

- Recording equipment - poor quality
  - iPhone XS

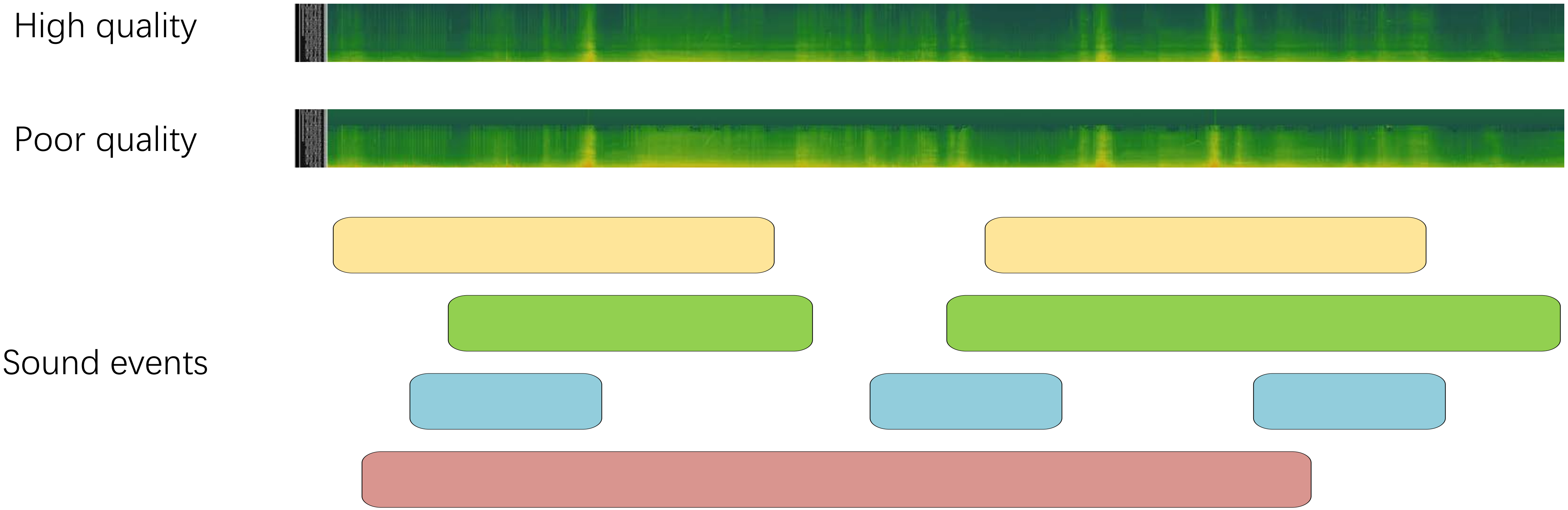


SG Dataset

- Recording equipment - annotation
  - Action camera

SG Dataset

- Post-processing
  - Alignment



SG Dataset

- Domain adaptation
  - Mismatch of event characteristics and acoustic environment

SG Dataset

- Domain adaptation
  - Mismatch of event characteristics and acoustic environment
  - Mismatch of recording conditions

SG Dataset

- Domain adaptation
  - Mismatch of event characteristics and acoustic environment
  - Mismatch of recording conditions
  - Mismatch of background noise

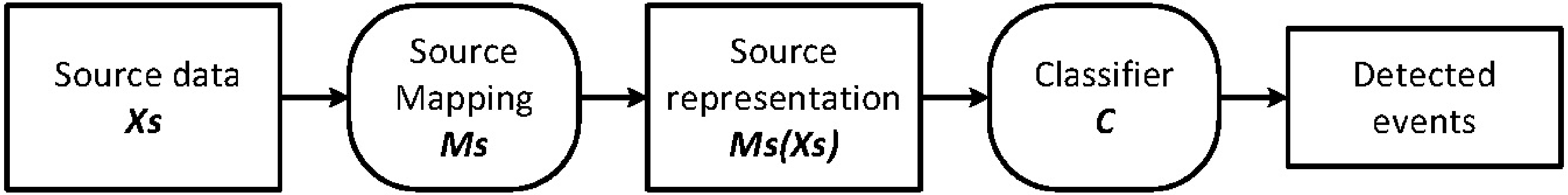
A-CRNN

- An unsupervised adversarial-based domain adaptation model
- Based on a domain adaptation model for acoustic scene classification [1]
- Three steps
  - Pre-training
  - Adversarial training
  - Testing

[1] Shayan Gharib, Konstantinos Drossos, Emre Cakir, Dmitriy Serdyuk, and Tuomas Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), November 2018, pp. 138–142.

A-CRNN

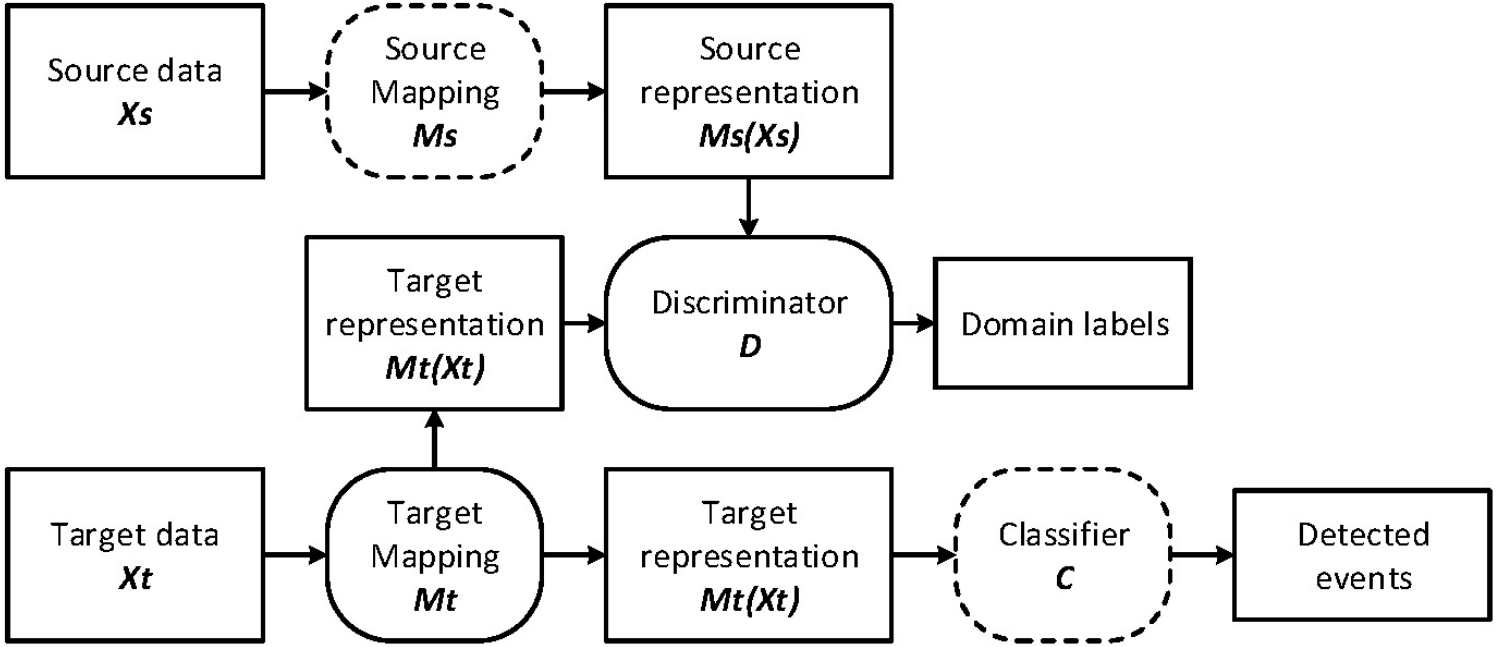
- Pre-training step
  - Train a model for the source domain



$$\min_{M_s,C} L_s = -\frac{1}{N_s}\sum_{n=1}^{N_s}\sum_{k=1}^{K} \mathbb{1}_{[k=Y_s^n]} \log C(M_s(X_s^n))$$

A-CRNN

- Adversarial training step
  - Adapt the mapping to fit the target domain

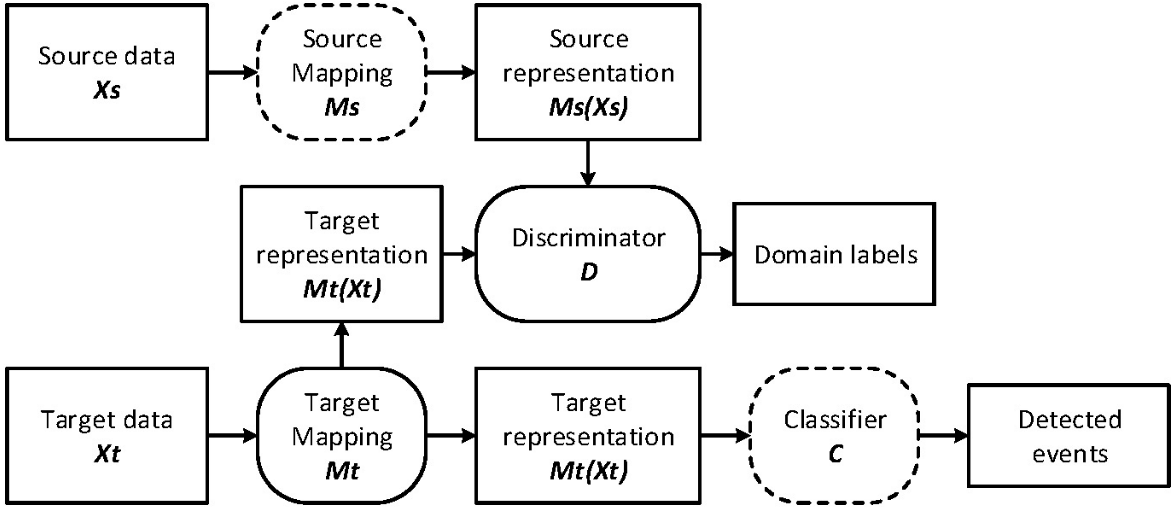


A-CRNN

- Adversarial training step

$$\min_{D} L_d = -\frac{1}{N_s}\sum_{n=1}^{N_s}\log D(M_s(X_s^n)) - \frac{1}{N_t}\sum_{n=1}^{N_t}\log(1 - D(M_t(X_t^n)))$$

$$\min_{M_t} L_t = -\frac{1}{N_t}\sum_{n=1}^{N_t}\log D(M_t(X_t^n)) - \frac{1}{N_s}\sum_{n=1}^{N_s}\sum_{k=1}^{K}\mathbb{1}_{[k=Y_s^n]}\log C(M_t(X_s^n))$$



A-CRNN

- Testing step
  - Test the model on the target domain

A-CRNN

- Detailed architecture
  - CNN mappings (both source and target mapping)

Input	log Mel-band energies
Convolutional layers	128 filters of shape 3 x 3, ReLU, 1 x 5 max pooling
	128 filters of shape 3 x 3, ReLU, 1 x 2 max pooling
	128 filters of shape 3 x 3, ReLU, 1 x 2 max pooling

A-CRNN

- Detailed architecture
  - RNN classifier

Recurrent layers	32 units, GRU, tanh
	32 units, GRU, tanh
Fully connected layers	16 units, time distributed, ReLU
	5 units, time distributed, Sigmoid

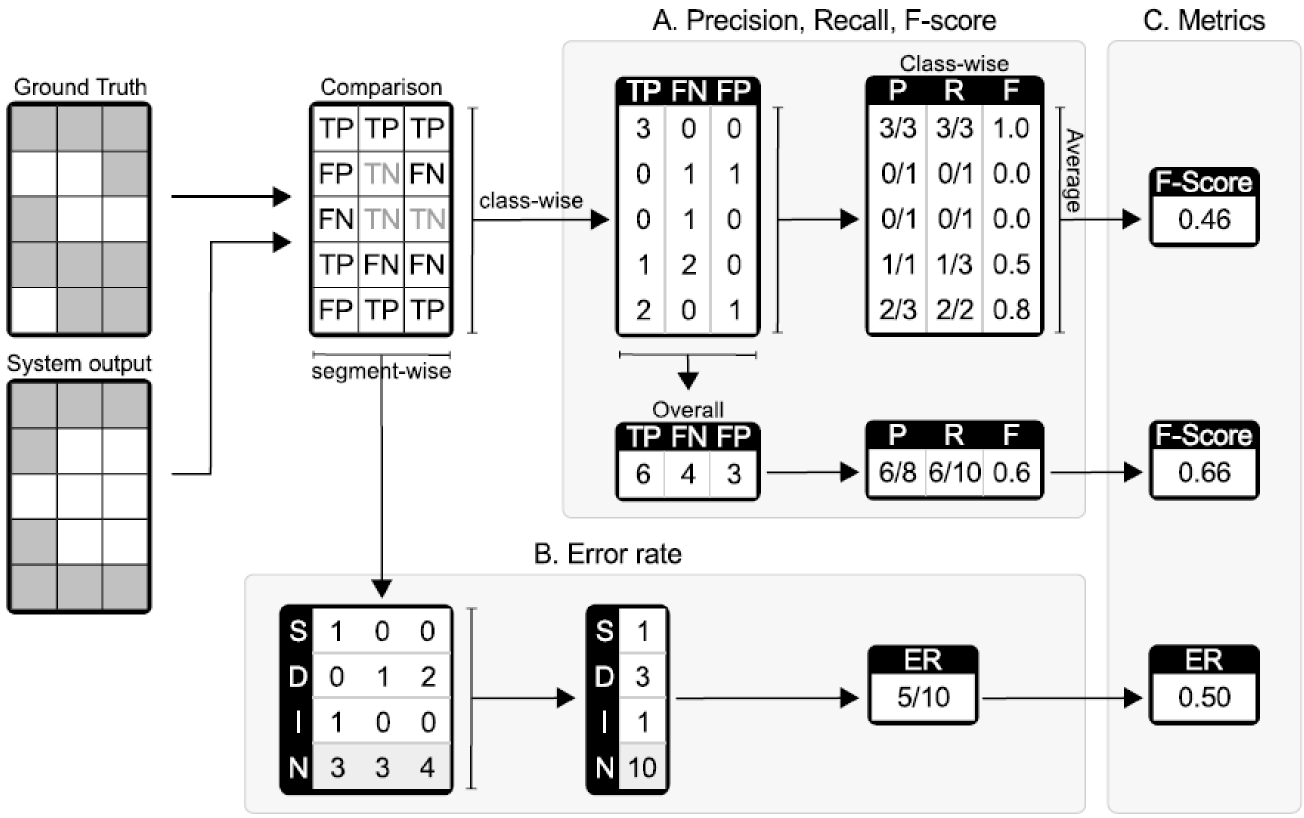
A-CRNN

- Detailed architecture
  - RNN discriminator

Recurrent layers	32 units, GRU, tanh
	32 units, GRU, tanh
	32 units, GRU, tanh
Fully connected layers	64 units, time distributed, ReLU
	64 units, time distributed, ReLU
	16 units, time distributed, ReLU
	2 units, time distributed, Softmax

A-CRNN

- Evaluation metrics [2]
  - F-score
  - Error rate



[2] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," Applied Sciences, vol. 6, no. 6, pp. 162, 2016.

A-CRNN

- Experiment results
  - Improvement on both source and target domains

Table 1: Results (Source: SG-high; Target: SG-low)

Model	Source domain		Target domain	
	F-score	Error rate	F-score	Error rate
CRNN	0.583	0.620	0.442	0.743
A-CRNN	0.590	0.609	0.480	0.688

A-CRNN

- Experiment results
  - Slight drop on source domain
  - Clear improvement on the target domain

Table 2: Results (Source: SG-high; Target: DCASE)

Model	Source domain		Target domain	
	F-score	Error rate	F-score	Error rate
CRNN	0.470	0.793	0.256	0.947
A-CRNN	0.427	0.869	0.458	0.826

Table 3: Results (Source: DCASE; Target: SG-high)

Model	Source domain		Target domain	
	F-score	Error rate	F-score	Error rate
CRNN	0.528	0.705	0.163	1.072
A-CRNN	0.514	0.716	0.301	0.960

A-CRNN

- Experiment results
  - Smaller improvement

Table 4: Results (Source: DCASE; Target: SG-low)

Model	Source domain		Target domain	
	F-score	Error rate	F-score	Error rate
CRNN	0.528	0.705	0.223	1.097
A-CRNN	0.511	0.757	0.295	0.936

A-CRNN

- Experiment results
  - Class-wise F-score

Table 6: Class-wise F-score on the target domain

Class name	DCASE to SG-high		DCASE to SG-low	
	CRNN	**A-CRNN**	**CRNN**	**A-CRNN**
car	0.256	0.473	0.357	0.479
children	0.072	0.005	0.235	0.005
large vehicle	0.119	0.004	0.109	0.024
people speaking	0.297	0.056	0.334	0.149
people walking	0.081	0.096	0.082	0.104

Future Work

- Other non-adapted model architectures

Future Work

- Other non-adapted model architectures

- Improve performance for certain classes

Future Work

- Other non-adapted model architectures
- Improve performance for certain classes
- Other domain shift aspects
- Semi-supervised domain adaptation model

Conclusions

- Problem addressed
  - Domain adaptation for sound event detection
- Solution
  - SG dataset
  - Domain adaptation model: A-CRNN

Thank you