Diplomarbeitspräsentation

FAKULTÄT FÜR !NFORMATIK

Faculty of Informatics

Sound Event Detection with Deep Neural Networks

Masterstudium:

Computational Intelligence

Seyedehanahid Naghibzadehjalali

Technische Universität Wien

Information System engineering

Arbeitsbereich: Information and Software Engineering group

Betreuer: Ao.Univ.-Prof. Dr. Andreas rauber

Introduction

PROBLEM DEFINITION

- The studies on this topic are related to the **cocktail party problem** (refers to the remarkable ability of the brain in selective attention)

GOAL

- Goal is to use an intelligent system to automatically detect if any of the sound events within the given acoustic signals

APPLICATION AREA

- Military and security/surveillance applications
- Long term remote monitoring
- Sound indexing
- Smart home/ cities systems

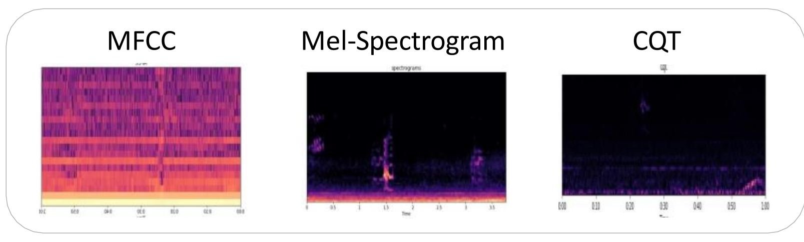
CONTRIBUTION OF THE PROJECT

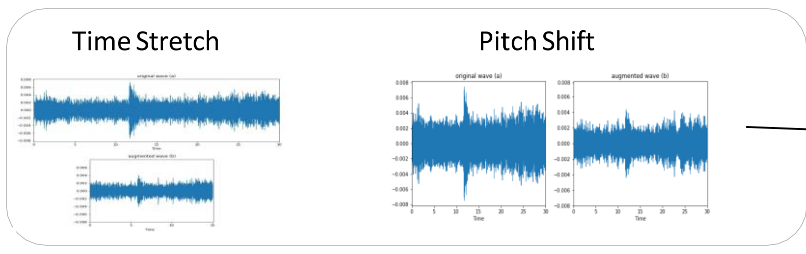
- Utilizing multiple deep learning architecture on Sound Event Detection task (SED)
  - Long Short Term Memories (LSTMs)
  - Bi-Directional LSTMs
  - Convolutional Neural Networks
- Comparing the performance of deep learning architectures on three input representation techniques
  - Mel Frequency Coefficient Cepstrals
  - Constant-Q Transforms
  - Log-Amplitude Mel-Spectrograms
- Generalizing the model using techniques such as Data augmentation and dropout
- Evaluating the models on 2 different datasets provided by DCASE community
  - Monophonic Rare Sound Event Detection
  - Polyphonic Real Life Street Sound Event Detection

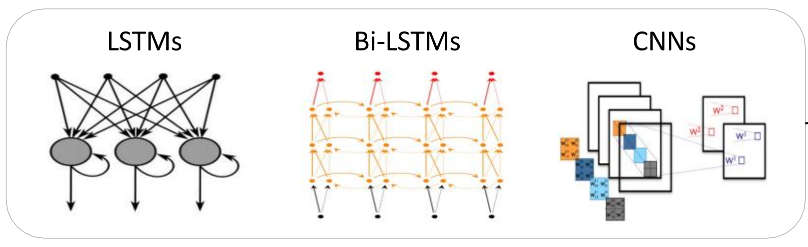
Related Work

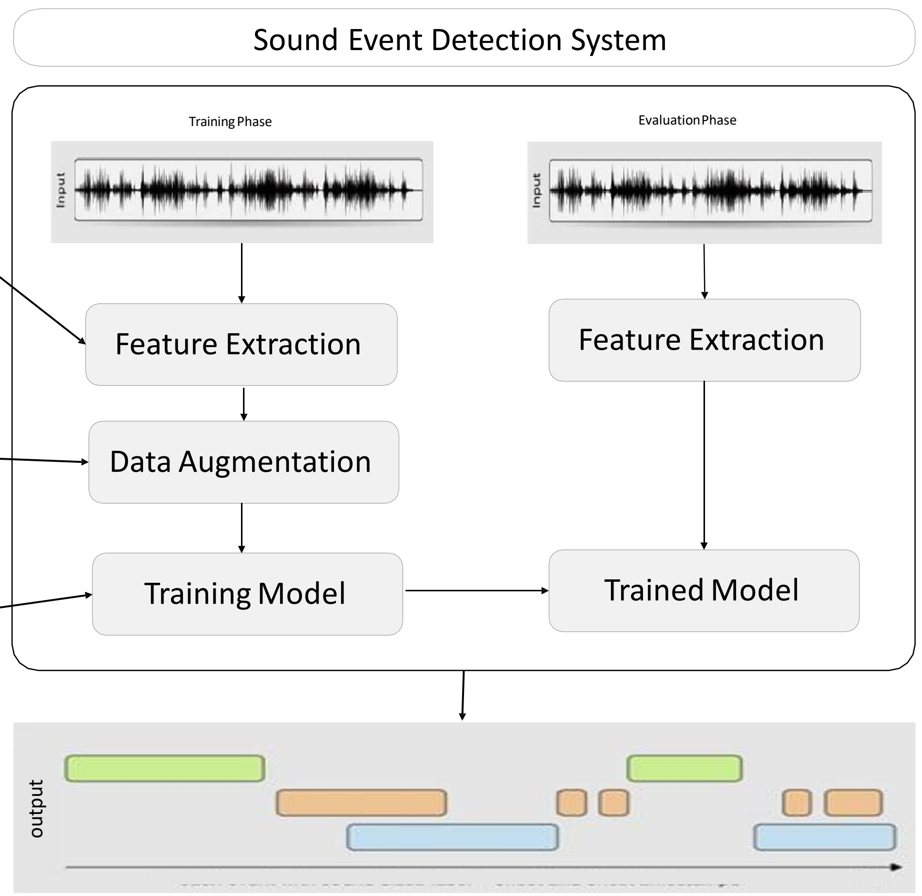
- **Use of Bi-directional Long Short Term Memory** extracts the full content in an input sequence [1]
- **Use of Mel-band energy as features** for Deep Neural Networks enhanced the performance of model [2]
- **Audio manipulation for data augmentation** improves the reliability of the prediction [3]

Methodology







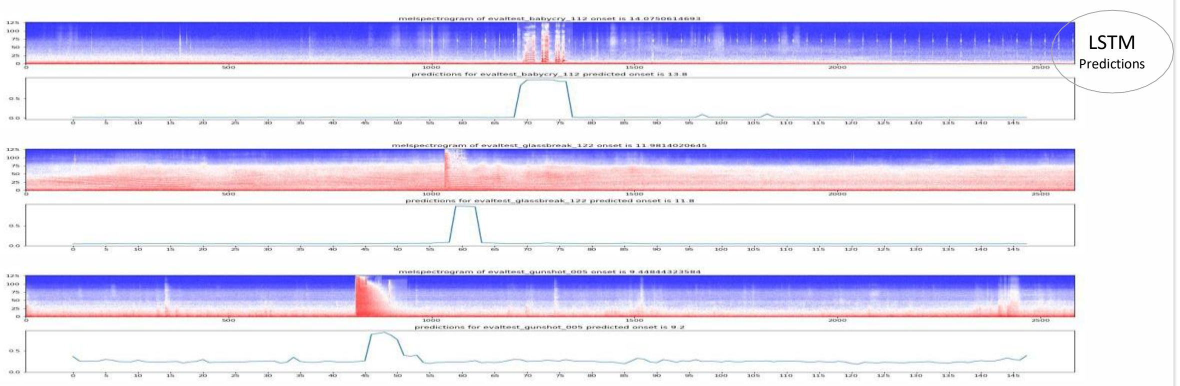


Visualization of the Experimental Results

MODEL EVALUATION (Rare Sound Event Detection)

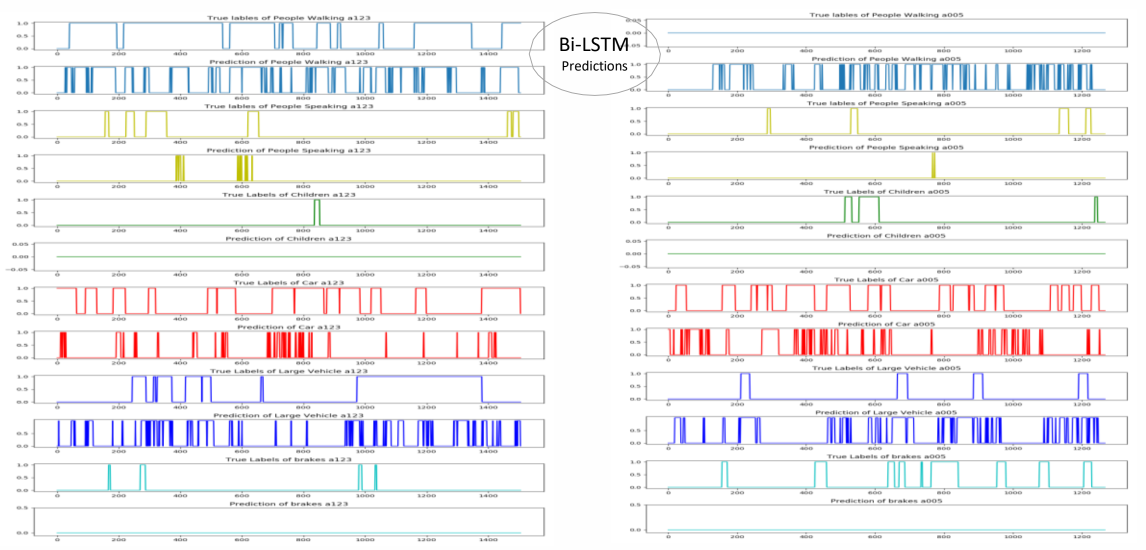
Comparison of Different Input Representation (Model : LSTMs)						
Model	Mel-Spectrogram		CQT		MFCCs	
Classes	Err	F1	Err	F1	Err	F1
Babycry	0.28	77.35%	0.27	73.26%	0.39	76.22%
Glassbreak	0.45	79.43%	0.71	29.48%	0.74	67.37%
Gunshot	0.54	68.52%	0.65	44.44%	0.46	61.40%
Average	**0.42**	**75.10%**	0.54	49.06%	0.53	68.33%

Model	Baseline (MLP)		LSTM		BLSTM		CNN	
Classes	Err	F1	Err	F1	Err	F1	Err	F1
Babycry	0.67	72.00%	0.27	77.84%	0.40	69.43%	0.24	83.17%
Glassbreak	0.22	88.50%	0.34	81.05%	0.33	76.27%	0.24	84.17%
Gunshot	0.69	57.40%	0.53	69.53%	0.69	41.47%	0.44	58.04%
Average	0.53	72.70%	**0.38**	**76.16%**	0.47	62.34%	**0.30**	**75.12%**



MODEL EVALUATION (Real Life Street Sound Event Detection)

Model	Baseline (MLP)		LSTM		BLSTM		CNN	
Classes	Err	F1	Err	F1	Err	F1	Err	F1
People Walking	1.44	33.5%	0.89	13.51%	0.91	15.94%	0.92	9.29%
People Speaking	1.29	8.6%	0.92	12.21%	0.95	6%	0.97	2.71%
Children	2.66	0.0%	1	0.0%	0.99	1.11%	0.98	2.88%
Car	0.76	65.1%	0.63	42.35%	0.7	34.35%	0.74	28.82%
Large Vehicle	1.44	42.7%	0.87	14.66%	0.91	9.71%	0.84	16.38%
Brake	0.98	4.1%	1.	0.0%	0.97	3.28%	1	0.0%
Average	0.93	42.08%	0.89	41.02%	0.93	31.72%	0.77	28.31%



Conclusions

- Deep Learning Appoaches are well suited for SED tasks
- Data Augmentation reduced the False Positive Rates
- Mel spectrograms are more appropriate for Deep Neural Networks
- Polyphonic SED requires more advanced signal processing

Future Work

- Apply Hybrid models such as C-RNN which have shown robustness on feature learning.
- Apply attention layer to improve model's performance in SED
- Investigation of Multi-channel Audio Analysis

References

[1] Hayashi T; Watanabe S; Toda T; 2016. Bidirectional lstm-hmm hybrid system for polyphonic sound event detection.

[2] Adavvane S; Parascandolo G; Heittola T; Virtanen T; 2016. Sound event detection in multichannel audio using spatial and harmonic features, DCASE2016.

[3] Cui X; Goel V; Kingsbury B; 2015. Data augmentation for deep neural network acoustic modelling, TASLP15.

Kontakt: anahid.jalali@gmail.com