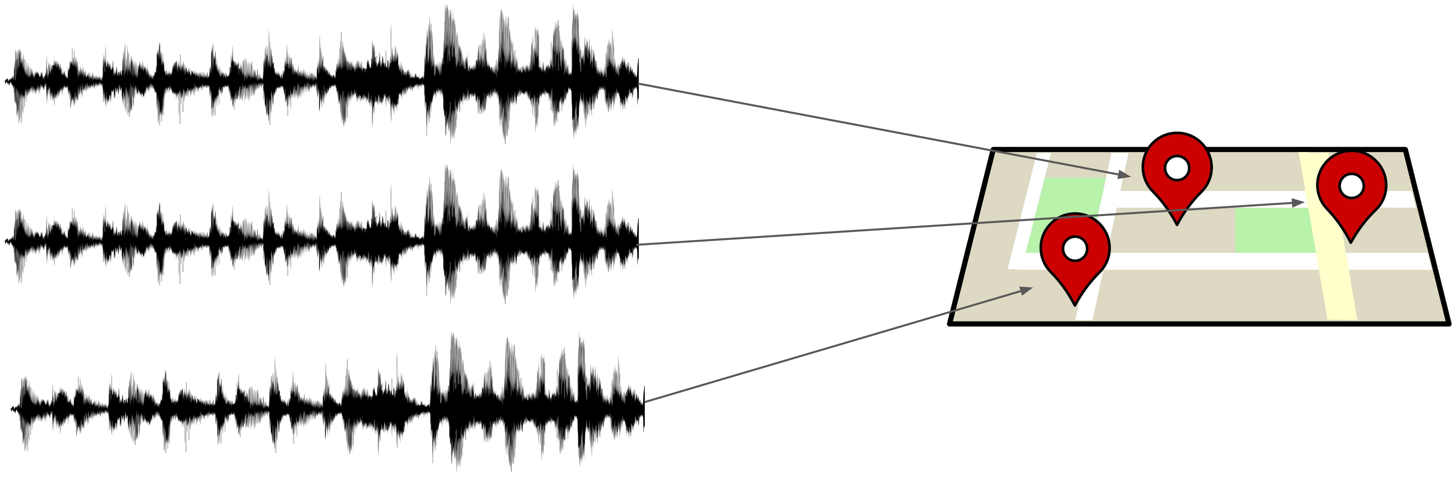
Where am I?
Acoustic Location Classification with Temporal Lags

Master’s Thesis

Stefan Kuhs
2020-05-28

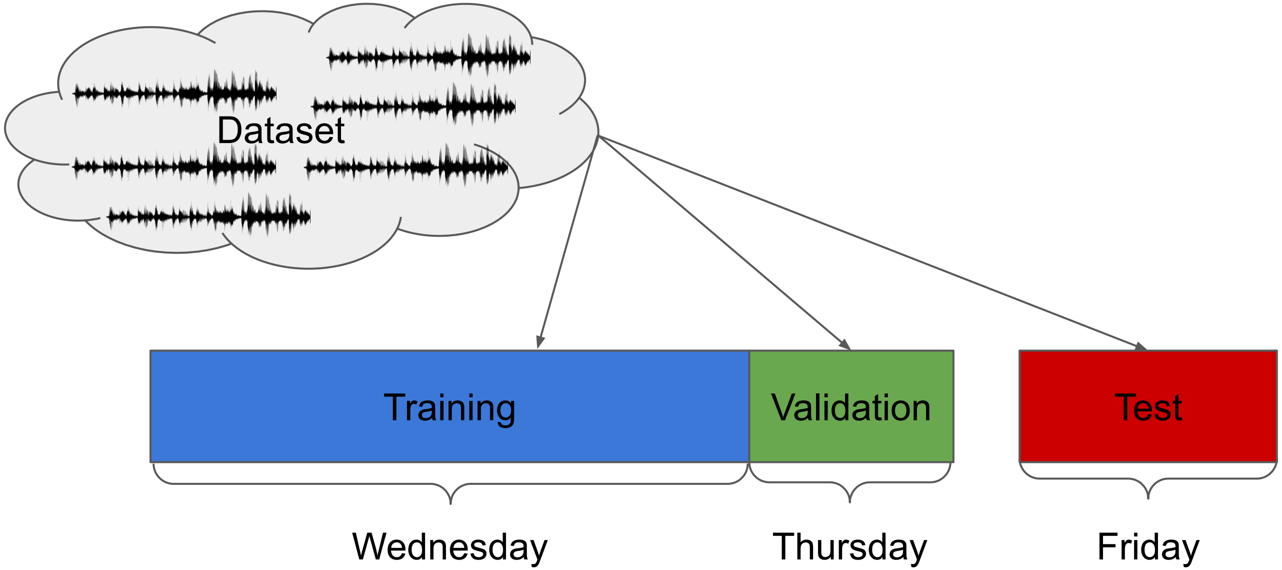
Acoustic Location Classification - **ALC**

- Artificial classification of audio samples to locations
- **Temporal** Constraints (Lags)



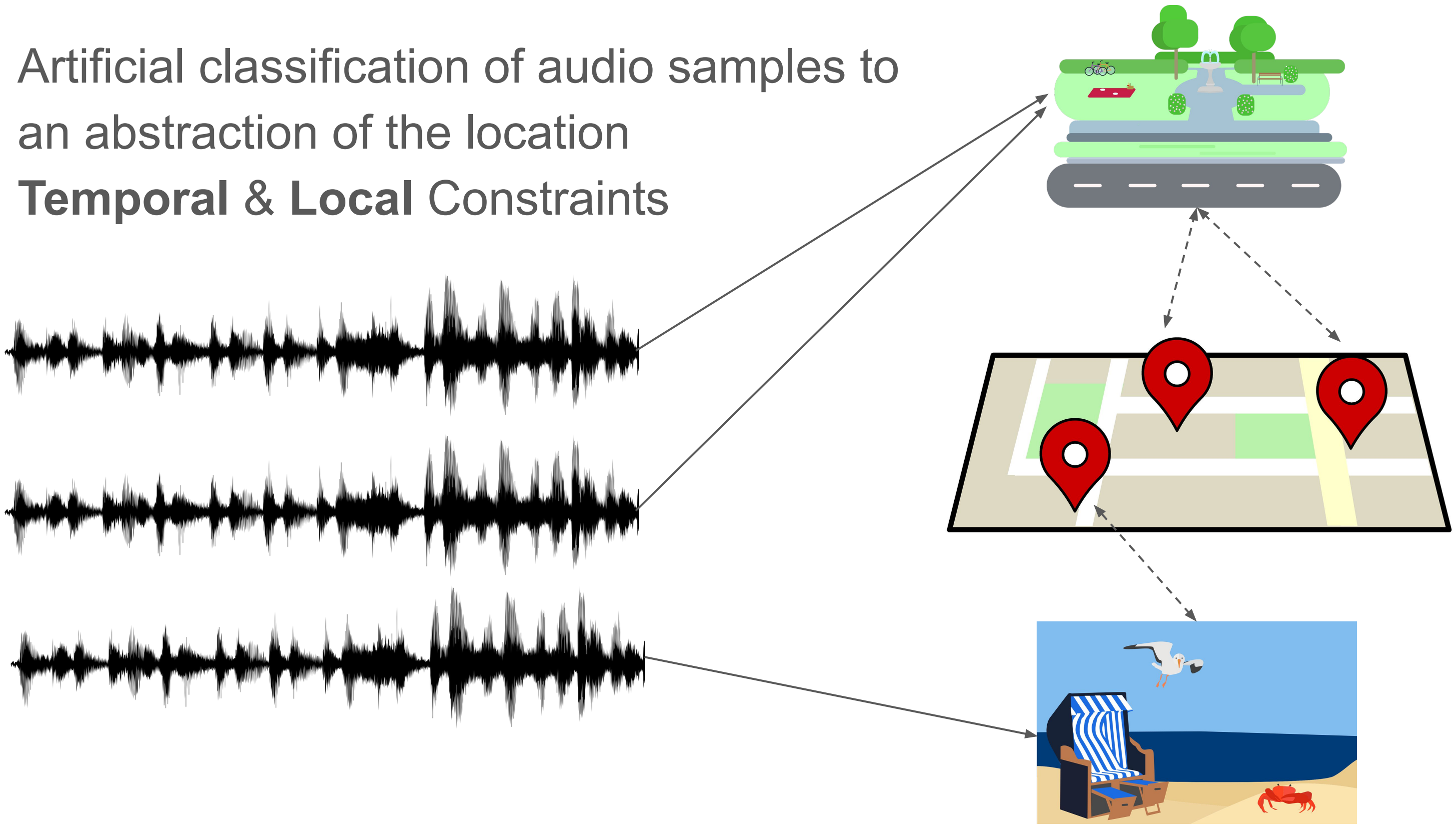
Temporal Constraints (Lags) ⏳

- Constrain audio samples on a temporal basis



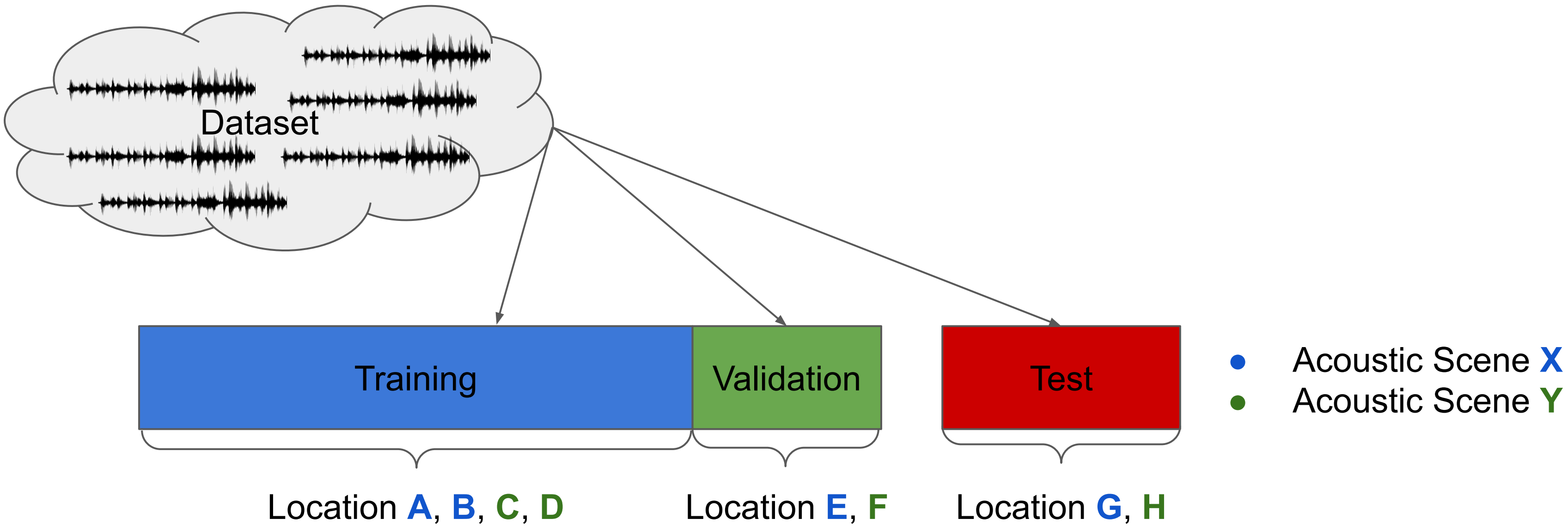
Acoustic Scene Classification - **ASC**

- Artificial classification of audio samples to an abstraction of the location
- **Temporal** & **Local** Constraints



Local Constraints

- Constrain audio samples based on locations



Objective

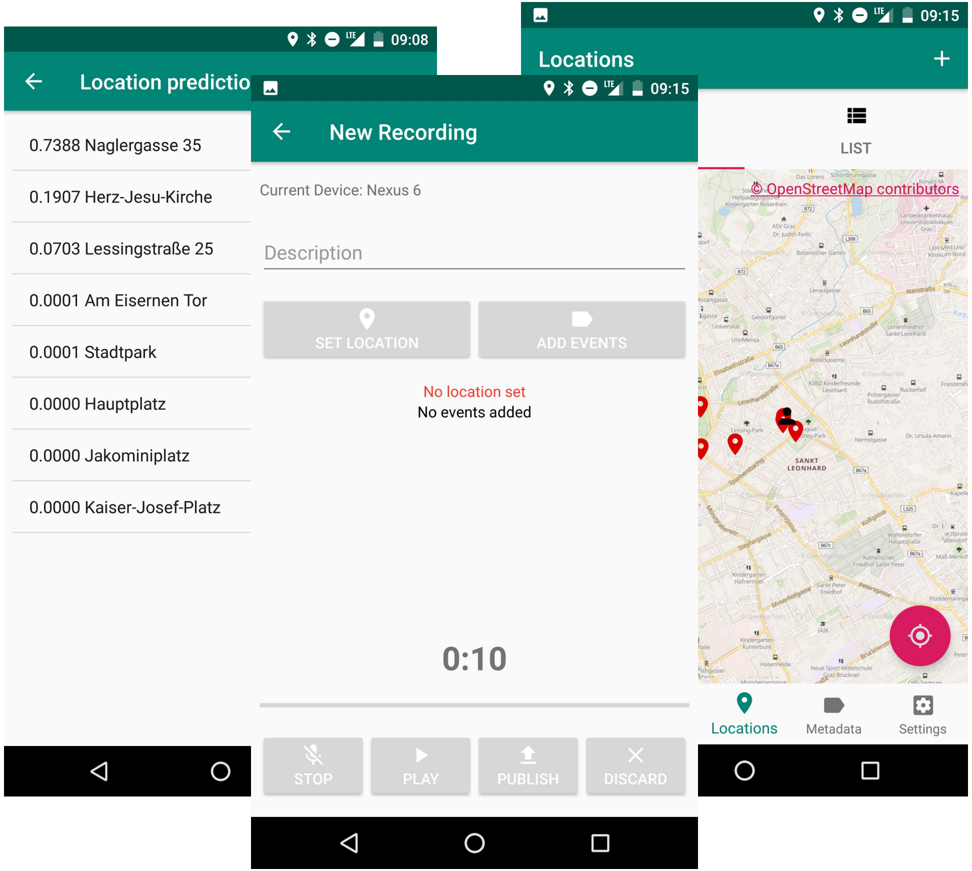
- Acoustic Location Classification
    - Capabilities
    - Limitations

- Relationship between **ASC** and **ALC**
    - Properties
    - Difficulties

- Impact of Constraints
    - **Local** & **Temporal**
    - **ASC** & **ALC**

Collecting audio data

- Client-server infrastructure
- Dedicated Android application
- Audio recordings
  - Consistent properties
  - Unprocessed
  - Monophonic
  - 48 kHz sample rate
  - Bit depth of 16 bit
- Metadata
  - Locations
  - Acoustic Scenes
  - Events
- Cloud-based classification
  - Convolutional Neural Network
  - **ASC** & **ALC**



Dataset for Acoustic Location Classification (**Graz** DS)

- 8 locations within target region Graz (AUT)
- 2 acoustic scenes
  - Public square
  - Urban green space
- 3 consecutive working days
  - Wednesday, Thursday, Friday
- 5 minutes per day and location
- Intraday time frame
  - 9:00 am to 12:00 am
  - Temporal intraday identity
- Minimal microphone movement
- *Unchanged position at locations*

Dataset for Acoustic Location Classification (Graz DS)

- Public squares
  - Am Eisernen Tor
  - Hauptplatz
  - Jakominiplatz
  - Kaiser-Josef-Platz

- Urban green spaces
  - Herz-Jesu-Kirche
  - Lessingstraße 25
  - Naglergasse 35
  - Stadtpark

Dataset for ALC (Graz DS) vs. TAU dataset

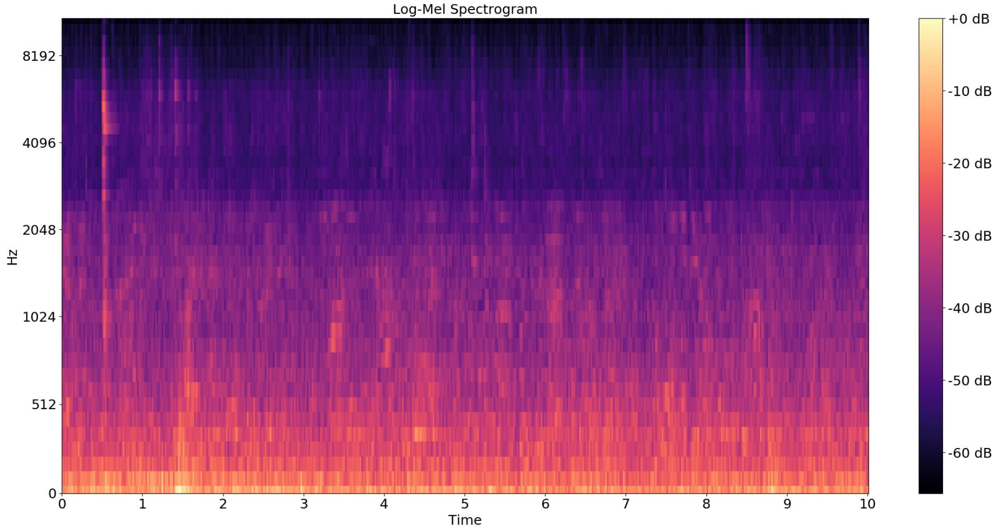
	Dataset for ALC (Graz DS)	TAU dataset
No. of 10 s audio samples	**720**	**14400** (20x larger)
No. of recordings	24 (8 locations x 3 days)	~ 1000
Duration of one recording	~ 300	~ 144 (2 - 3 minutes)
Acoustic scenes	2	10
Locations	8	514
Sample rate	48 kHz	48 kHz
Bit depth	16 bit	24 bit
Channels	1	2 (binaural)
Recording solution	Nexus 6 with dedicated software	Professional audio recorder and in-ear microphone
Applicable for ALC & ASC with Temporal Constraints	**YES**	No

Classification A-Z

- Preprocessing
  - 10 s audio samples (**Graz** DS 5 min recordings)
  - Monophonic (**TAU** DS binaural)
  - 48 kHz sample rate
  - 16 bit (**TAU** DS 24 bit) bit depth
- Feature extraction
  - 40-band log-mel spectrograms
- Split into training, validation, and test set
  - Remove classes with an insufficient number of samples
  - Preserve class-wise distribution (Normal distribution)
  - Local & Temporal Constraints
- Normalize to zero mean & unit variance
- Train Convolutional Neural Network (CNN) and evaluate predictions

Features & Convolutional Neural Network (In-depth)

- DCASE 2019 Baseline for **ASC**
- 40-band log-mel spectrograms of 10 s audio samples
  - 40 ms window size
  - 20 ms overlap
  - Hamming windows
- 2 convolutional layers
  - Max Pooling
- Softmax output layer
- Adam
- Early Stopping, Dropout
- Batch normalization
- Cross-validation, Independent training runs



Experiments

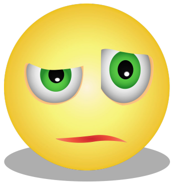
	Description	Dataset	Unconstrained	Local	Temporal
A	Acoustic scene classification	**TAU**	76,45 $_{+/-1,78}$	57,19 $_{+/-\ 2,30}$	*N/A*
B	**ASC** on Urban Parks & Public Squares	**TAU**	*N/A*	89,45 $_{+/-\ 0,51}$	*N/A*
C	**ALC** on Urban Parks	**TAU**	84,27 $_{+/-1,75}$	*N/A*	*N/A*
D	Acoustic scene classification	**Graz**	98,74 $_{+/-0,70}$	**88,15** $_{\mathbf{+/-\ 10,48}}$	95,14 $_{+/-2,37}$
E	**ASC** transferred from TAU dataset ❗	**TAU | Graz**	*N/A*	**93,00** $_{+/-1,51}$	**93,00** $_{+/-1,51}$
F	**ASC** by humans ❗	**Graz**	*N/A*	*N/A*	**95,83** $_{\mathbf{+/-5,89}}$
G	Acoustic location classification	**Graz**	**85,17** $_{\mathbf{+/-1,70}}$	*N/A*	**63,82** $_{\mathbf{+/-4,49}}$
H	**ALC** with varying number of training samples ❗	**Graz**	*N/A*	*N/A*	**[26,32, 66,04]** $_{+/-9,46}$

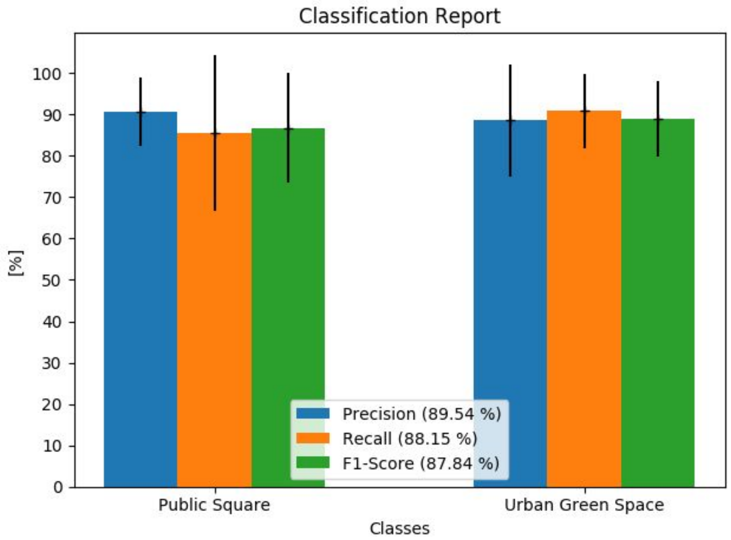
1-30 per location,
8-240 in total,
approx. 20 x 10 s
samples per location

Locally Constrained **ASC**

- **Graz** dataset
- 4-run Cross-validation
- 4-2-2 split
- Locations equally distributed w.r.t. the Acoustic Scenes

- High Std Dev between CV runs
  - Precision 9.15 %
  - Recall 10.48 %
  - F_1 score 10.99 %





Locally Constrained **ASC** (In-depth)

CV	Target	Location	Misclassification
1	Public square	Kaiser-Josef-Platz	7,33
	Urban green space	Stadtpark	0
2	Public square	Am Eisernen Tor	**43,89**
	Urban green space	Herz-Jesu-Kirche	10,11
3	Public square	Hauptplatz	5,56
	Urban green space	Lessingstraße 25	**20,00**
4	Public square	Jakominiplatz	1,22
	Urban green space	Naglergasse 35	6,67

Unique soundscape unsupported by training data

Construction work based sound emissions on the first day

Transfer model

- Acoustic Scene Classification
  - Train on subset of **TAU** dataset
  - Evaluate on **Graz** dataset

- **TAU** dataset ↔ **Graz** dataset
  - Urban Park ↔ Urban Green Space
  - Public Square ↔ Public Square

- Implies **Local** and **Temporal** Constraints

Transfer model (In comparison)

Training	**Evaluation**	**Constraints**	**Accuracy**
TAU dataset	**Graz** dataset	*Temporal & Local*	93,00
Graz dataset	**Graz** dataset		98,74
Graz dataset	**Graz** dataset	Local	88,15
Graz dataset	**Graz** dataset	Temporal	95,14

Humans versus Machines

- Acoustic Scene Classification
  - **Temporally** Constrained ⏳
  - Subset of **Graz** dataset

- 3 Test persons
  - 2 test persons resident in Graz
  - 1 test person involved in the collection process

- Training
  - 1 10 s audio sample for each of the 8 locations
- Evaluation
  - 2 10 s audio samples for each of the 8 locations

Humans versus Machines (Results)

	Precision	Recall	F_1	Support
Public square	93,33	100,00	96,30	8
Urban green space	100,00	91,67	95,24	8
Average	96,67	95,83	95,77	
+/-	4,71	5,89	5,99	

- 2 test persons with perfect classification
- Test person not resident in Graz
  - 2 x Herz-Jesu-Kirche misclassified as Public Square
    - Noise of passing vehicles within test samples

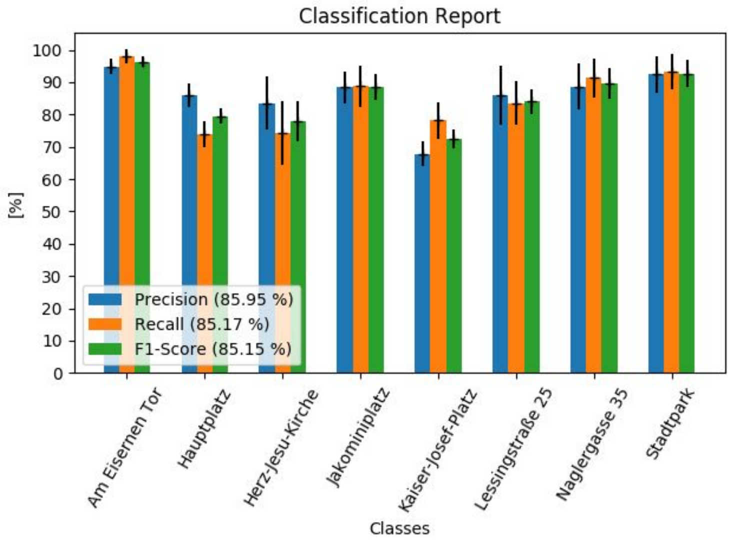
Humans versus Machines (In comparison)

Type	Dataset	Constraints	Accuracy
ASC by humans	Subset of **Graz** dataset	Temporal	95,83
Artificial **ASC**	**Graz** dataset		98,74
Artificial **ASC**	**Graz** dataset	Local	88,15
Artificial **ASC**	**Graz** dataset	Temporal	95,14

- *Humans outperform artificial counterpart!*

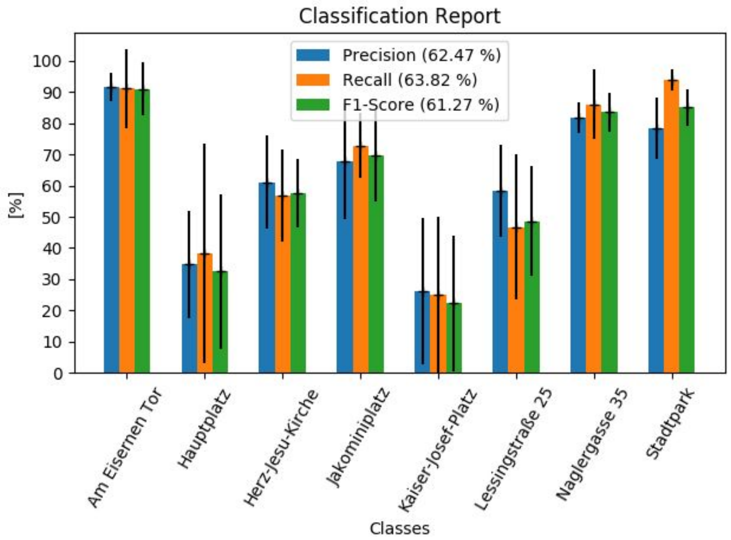
Unconstrained **ALC**

- **Graz** dataset
- <u>Reminder:</u>
    - 8 locations
    - 3 days
- Training, validation, and test set
    - 33 % - 33 % - 33 %
    - Benchmark for **Temporally** Constrained **ALC**
- 10 independent training runs
- Decent classification scores
    - ~ 85 % Precision, Recall, F_1 score



Temporally Constrained **ALC**

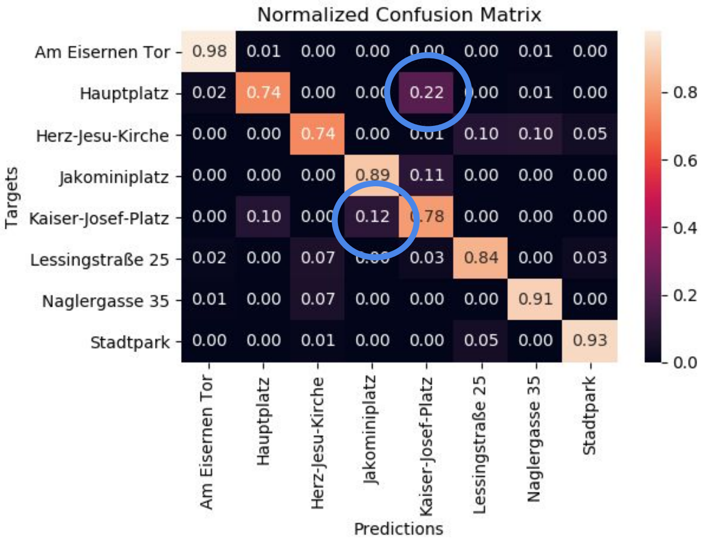
- **Graz** dataset
- Training, validation, and test set
  - 33 % - 33 % - 33 %
  - 1 day for training / validation / testing
- CV with 6 runs
  - Permutations of the days

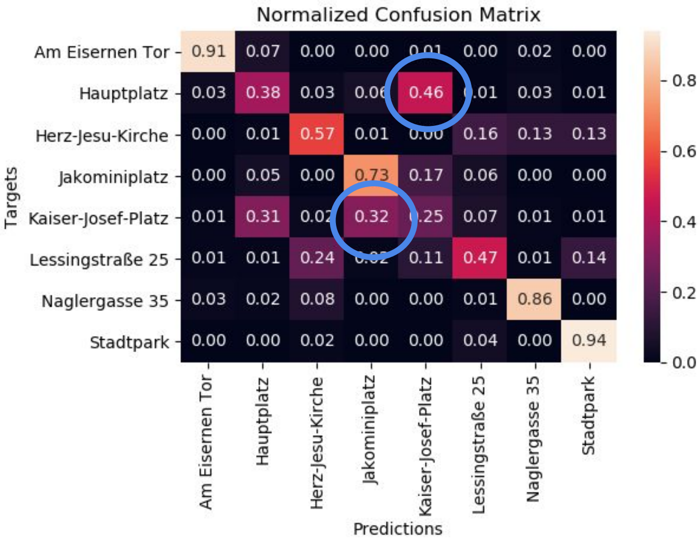


- Comparatively bad performance
- High Std Dev between CV runs for particular classes

Comparatively bad performance

Unconstrained vs. **Temporally** Constrained ALC

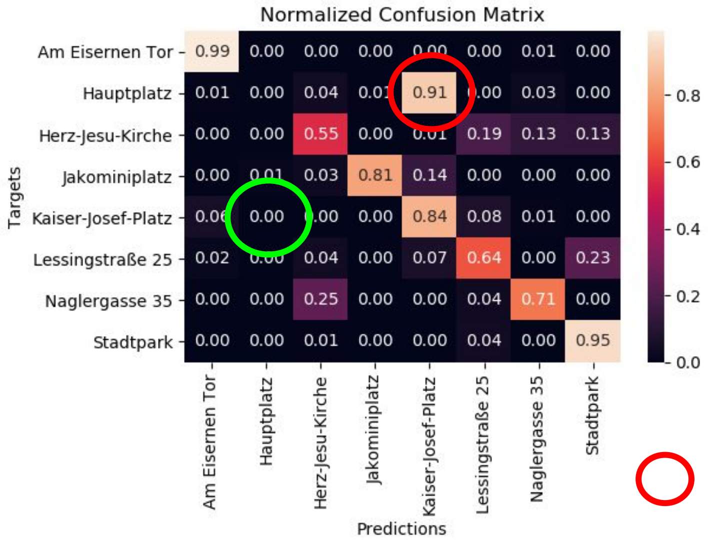




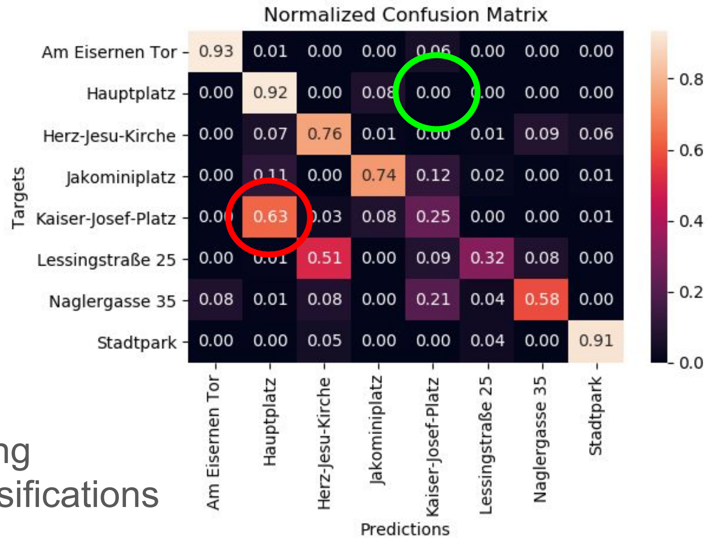
- Without Constraints
  - Higher classification sore
  - Falsified generalization estimate
- **Temporal** Constraints
  - Dampened classification score
  - **Susceptible classes preserved**

High Std Dev between CV runs for particular classes

1st vs. 4th CV run - **Temporally** Constrained **ALC**



Emerging Misclassifications



- 1st CV run
  - 1st day training
  - 2nd day validation
  - 3rd day testing
- 4th CV run
  - 1st day testing
  - 2nd day validation
  - 3rd day training

Training set size

- Acoustic Location Classification
  - **Temporally** Constrained ⌛
  - **Graz** dataset
  - 3-fold Cross-validation
  - 6 training-validation-test permutations

- Increasing number of training samples
  - Start with 1 sample per location
  - Till 30 samples per location

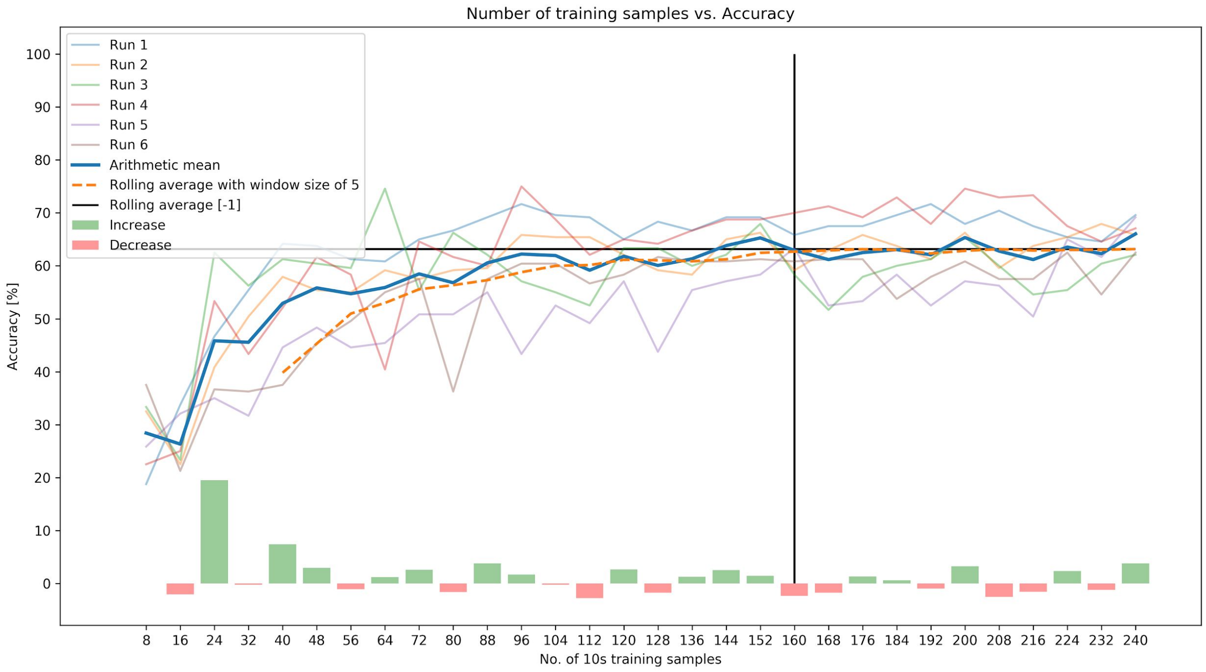
- Fixed validation and test set

Training set size (Results)

- 160 10 s audio samples
- 200 s of audio data per location

- Visual estimation
  - Humanly biased

- Biased towards the underlying dataset



Future work

- Increase dataset
  - Daily recordings
  - Time frame over several weeks or months
  - More locations
    - Similar and different ones!
  - *Monitor increasing intra-class/ decreasing inter-class variability* [Schmidhofer2018]

- Adapt model complexity, optimize hyperparameters, etc.

- Related topics
  - Mismatched recording devices [Mesaros2018]
  - Indoor locations [Tarzia2011]
  - Sound Event Localization and Detection (SELD) [Adavanne2018, Adavanne2019]
  - Artificial dataset augmentation [Chen2019]
  - Transfer Learning (TL) [Pan2009]

Findings & Conclusion

- Constraints are necessary for **ASC** & **ALC**
  - Prevent biased evaluation
  - Reliable generalization estimates
  - Dampens the classification score
  - More samples required (e.g., **ASC** on **Graz** DS)
    - Obtain a well generalized representation for acoustic scenes
- Transferring models - **ASC**
  - **TAU** dataset ⟶ **Graz** dataset
- *Humans outperform machines* - **ASC**
- Approx. 200 s of audio data per location w.r.t. **Temporally** Constrained **ALC**

References

Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen. "A multi-device dataset for urban acoustic scene classification." *arXiv preprint arXiv:1807.09840* (2018).

Tarzia, Stephen P., et al. "Indoor localization without infrastructure using the acoustic background spectrum." *Proceedings of the 9th international conference on Mobile systems, applications, and services*. 2011.

Chen, Hangting, et al. "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling." *arXiv preprint arXiv:1907.06639* (2019).

Adavanne, Sharath, et al. "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks." *IEEE Journal of Selected Topics in Signal Processing* 13.1 (2018): 34-48.

Adavanne, Sharath, Archontis Politis, and Tuomas Virtanen. "A multi-room reverberant dataset for sound event localization and detection." *arXiv preprint arXiv:1905.08546* (2019).

References

Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

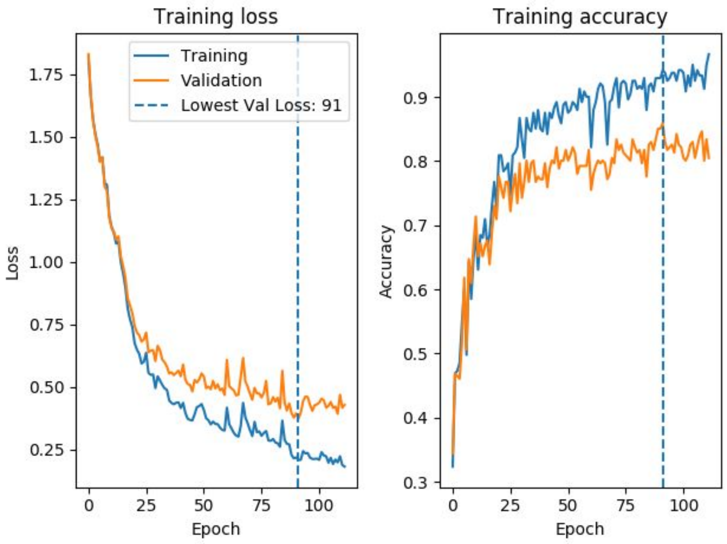
Imoto, Keisuke, and Nobutaka Ono. "Online acoustic scene analysis based on nonparametric Bayesian model." *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016.

Sahoo, Doyen, et al. "Online deep learning: Learning deep neural networks on the fly." *arXiv preprint arXiv:1711.03705* (2017).

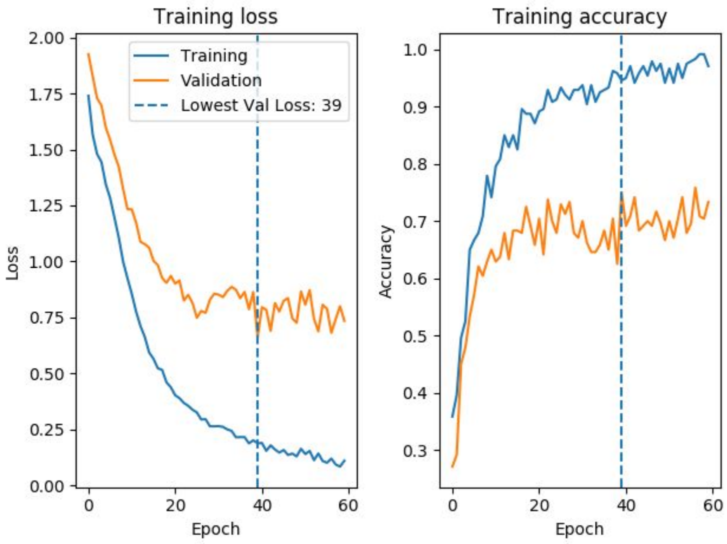
A. Schmidhofer, ‘Dataset generation guideline for acoustic transport mode detection’, Master’s thesis, Graz University of Technology, 2018.

Training-Validation Loss & Accuracy (**Graz** DS)

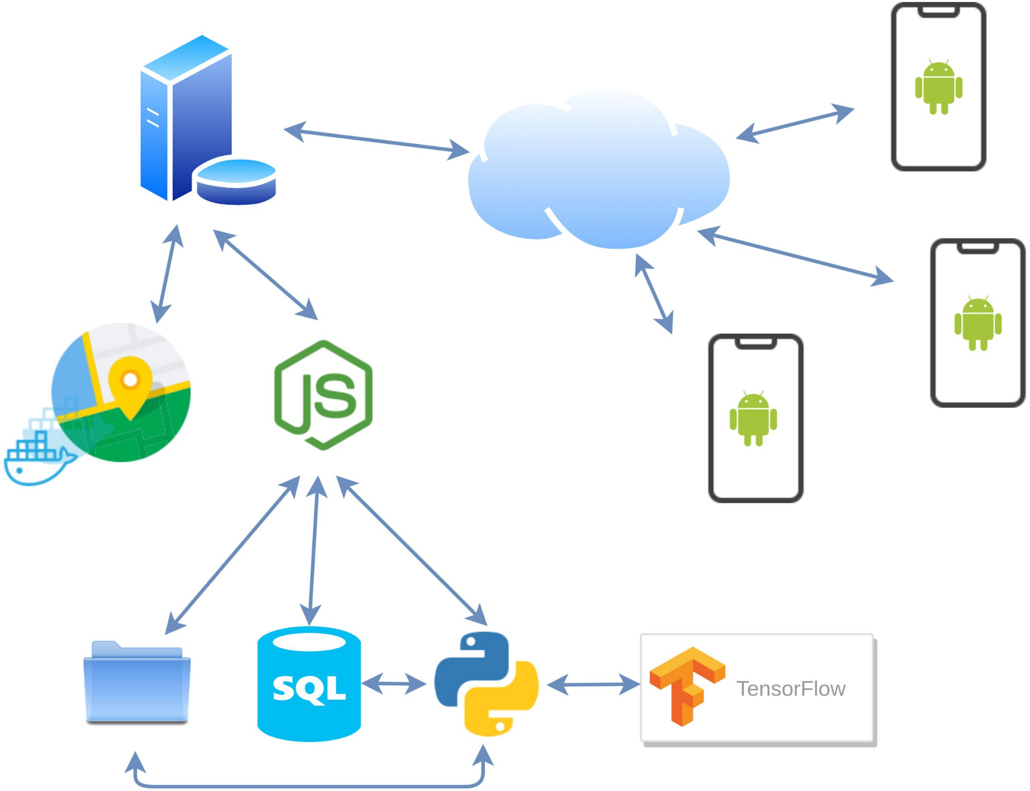
- **ALC without Constraints**



- **ALC** with **Temporal Constraints**



Infrastructure



Classification A-Z

1 Preprocessing
- 10 s sequences
- Monophonic
- 48 kHz sample rate
- 16 bit depth

2 Feature Extraction
- 40-band log-mel spectrograms
- Hamming windows
- 40 ms window size
- 50% overlap

3 Train-Test Split
- Training, validation, and test set
- Eliminate underrepresented classes
- Preserve class distributions
- Implement constraints

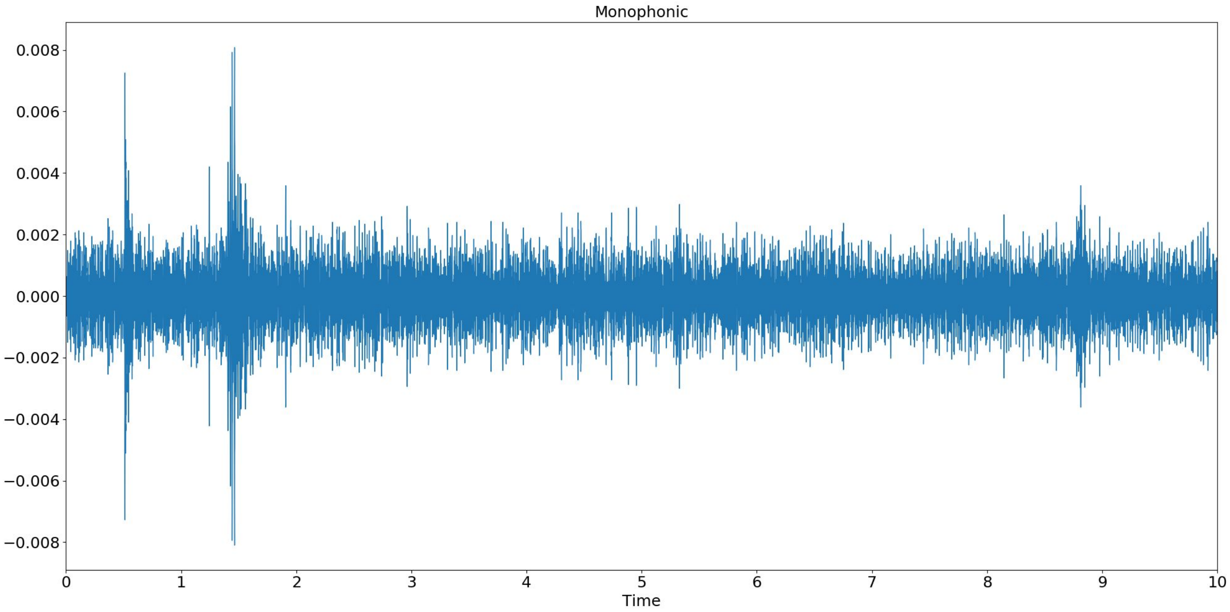
4 Normalization
- Zero mean & unit variance w.r.t. the training set

5 Training
- CNN
- Early stopping

6 Evaluation
- Precision
- Recall
- F1-score
- etc.

Recorded audio (In-depth)

- Unprocessed
- Monophonic
- 48 kHz sample rate
  - Nexus 6
- bit depth: 16 bit
  - Integer PCM



Overview

1. Collecting audio data
   - Client-server infrastructure
   - Dedicated Android application
   - Audio recordings and metadata
2. Datasets
   - Dataset for Acoustic Location Classification (**Graz** dataset)
   - TAU Urban Acoustic Scenes 2019, Development dataset [Mesaros2018] (**TAU** dataset)
3. Classification
   - Preprocessing, Feature extraction, etc.
   - Convolutional Neural Network
4. Evaluation
   - Experiments
   - Findings
5. Future Work

Summary

1. Set up infrastructure to collect audio samples and metadata with a dedicated Android application
2. Established dataset for Acoustic Location Classification
3. Implemented Classification A-Z
   - Preprocessing
   - Feature extraction
   - Constrained training & test splits
   - Convolutional Neural Network
4. Conducted several experiments
5. Evaluated and discussed the outcomes
6. Provided future outlook