Learning to Tag Environmental Sounds in Nightlong Audio

Master thesis

Mohd Saif Khan

Overview

- Data and its challenges
- Related work and the model used
- Hypotheses
- Results and the Decision support system

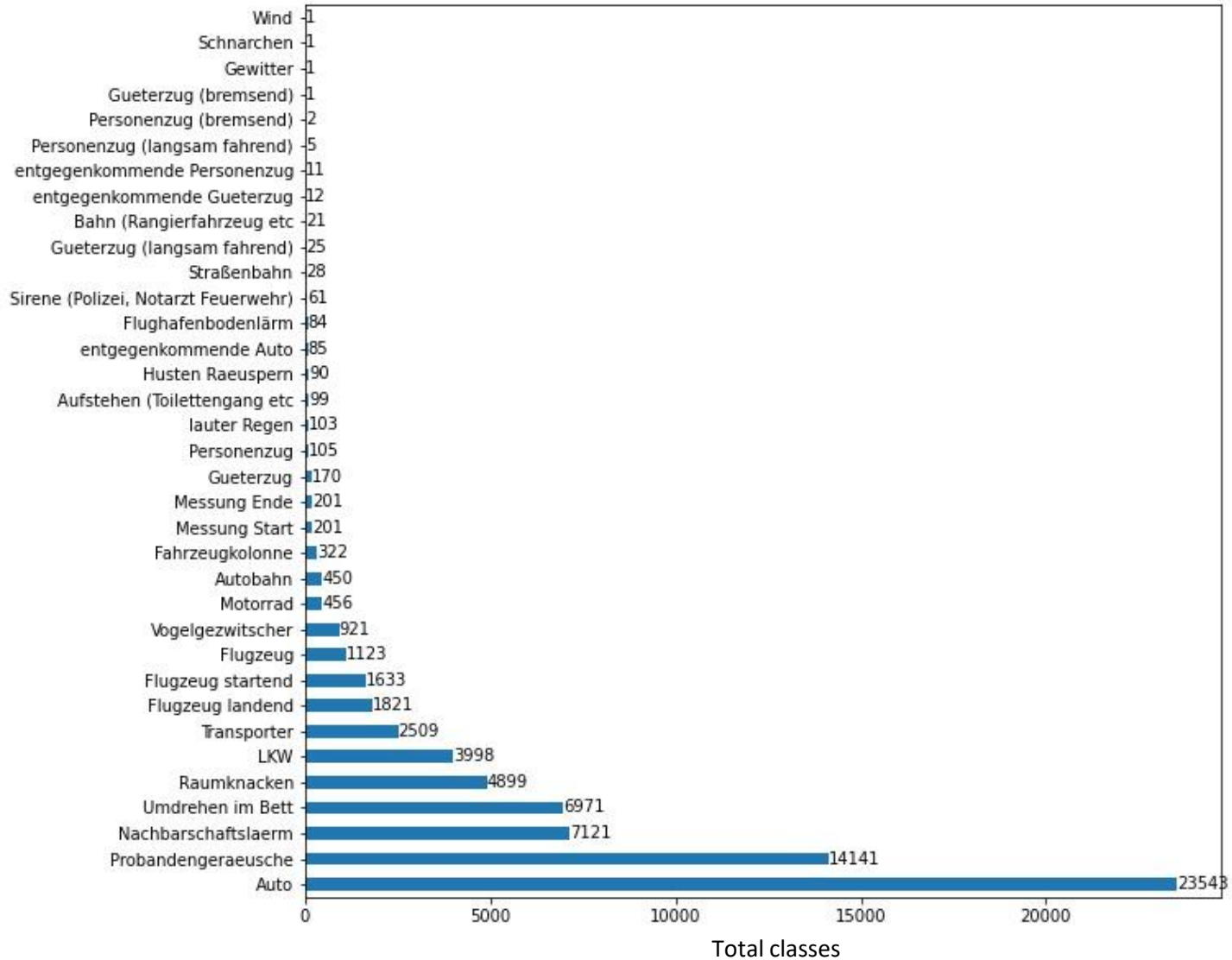
Data

- The German Aerospace Center is investigating effects of aircraft noise on sleep
- Participants consent to sleep in a controlled environment to obtain physiological data.
- 8-10 hours of audio is recorded and each sound class in the audio is tagged manually by annotators.

Challenges with data

- Class imbalance
- Average audio lengths of different classes
- Noise
- Weak labelling and misclassification

Class imbalance



Average audio lengths of different classes

- Due to different audio lengths, the audios must be padded or cut that could lead to loss of information.

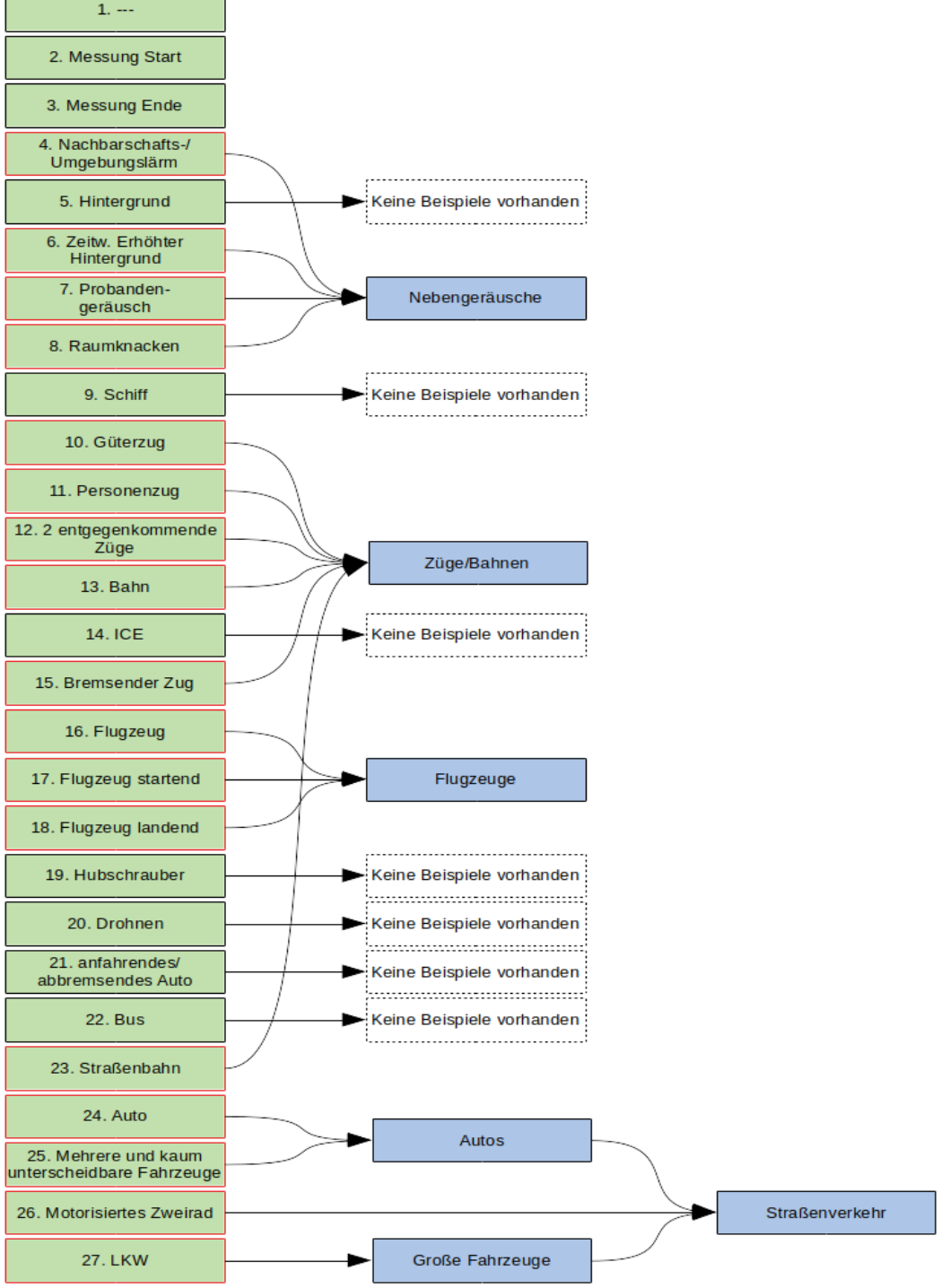
Class	Count	Total Duration (hh:mm:ss)	Average Length (s)
Auto	23448	107:37:56	16.52
Probandengeraeusche	14061	139:28:02	35.71
Nachbarschaftslaerm	7088	51:36:58	26.22
Umdrehen im Bett	6920	39:48:37	20.71
Raumknacken	4848	06:27:14	4.79
Flugzeug landend	1821	33:32:15	66.3
Flugzeug startend	1633	32:45:47	72.23
Flugzeug	1109	19:54:25	64.62
.	.	.	.
.	.	.	.
.	.	.	.
Wind	1	00:00:11	11

Noise

- Background noise or mixture of multiple sounds simultaneously.
- Some participants snored during the night or used fans during the night.

Weak labelling or misclassification

- All events are not annotated.
- Subjectivity of annotator's perception. The sound of moving fan can be confused with the sound of Auto etc.



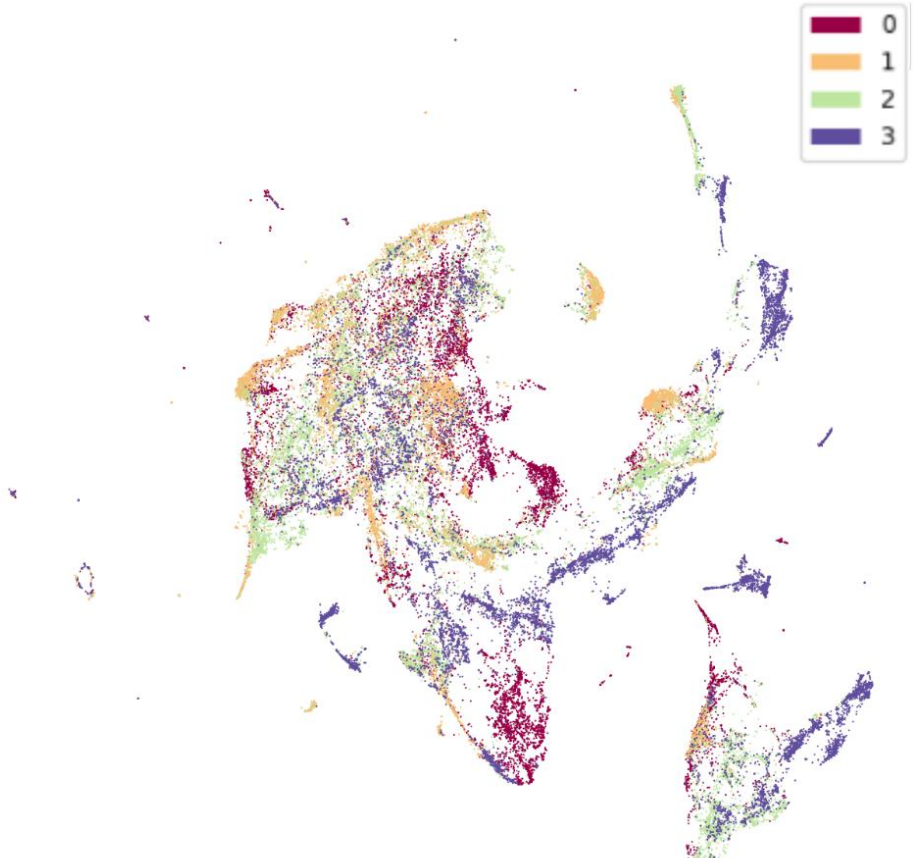
Sound classes and their hierarchy

- There was a total of 27 different sound event.
- These 27 classes were grouped together to form 5 broad classes.
- 3 classes (**Airplane (*Flugzeug*), Cars (*Autos*), and Neighborhood Noise (*Nebengeraeusche*)** along with ***Silence*** class are used.

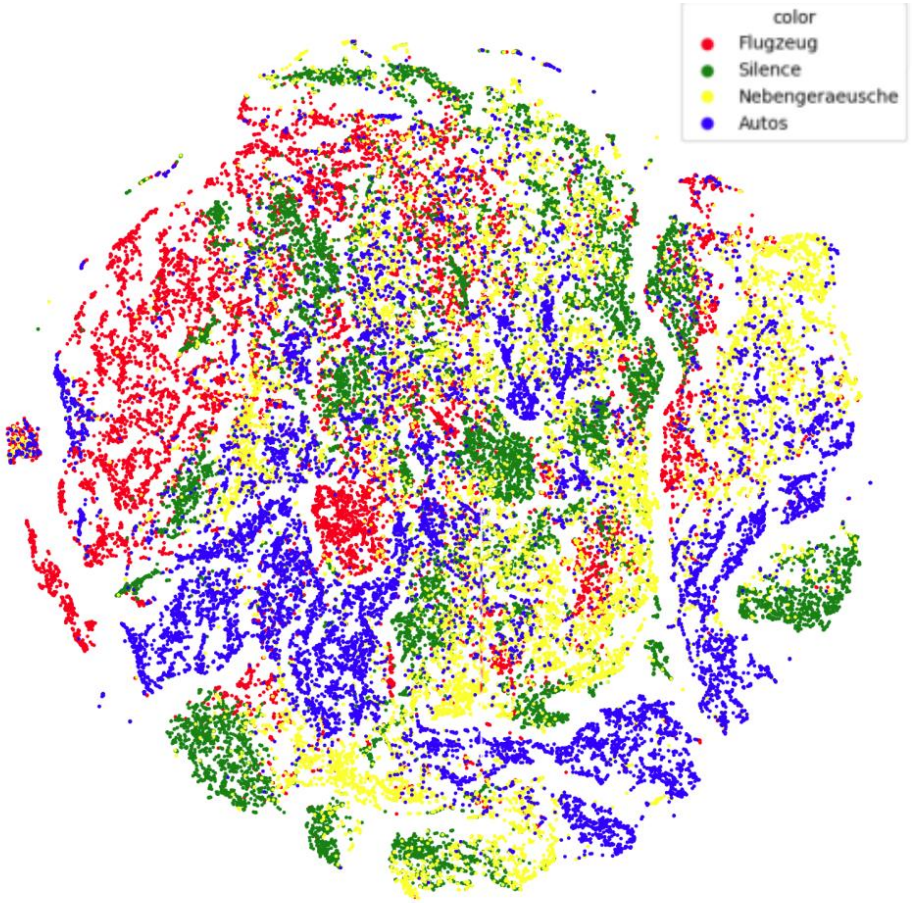



1. https://www.colourbox.com/vector/alzheimer-old-man-cartoon-vector-37220328
2. https://www.onallcylinders.com/2015/08/21/our-top-10-cartoon-cars-of-all-time/
3. https://www.pngitem.com/middle/iRxbxRJ_vector-illustration-of-commercial-airplane-passenger-airplane-gif/
4. https://freesvg.org/quiet-symbol

Insights on data by dimensionality reduction



UMAP



TSNE

Related Work

- Sound event classification
  - Gaussian Mixture models with KL divergence *(Aucouturier et al. (2007))*
  - KNN and k-medoids clustering to create bag of features and then applied Support vector machine *(Briggs et al. (2012) applied)*
  - RNN's have been used as they are efficient in modelling sequences *(Phan et al(2017))*
  - CNN has been used to classify audio signals *(Eghbal-zadeh et al(2017), Dorfer et al(2018))*

Research questions

- In reference to the challenges, what is the suitable method for classification of sound events?
- What are the optimal window size and overlap sizes in order to meet the target start and end point detection accuracy?
- How can we create semi-automatic decision support system that helps annotators to tag the audio recordings?

Feature- Mel Filter Banks

- It contains information about both time and frequency components of an audio signal.
- The output is a spectrogram that can be represented in a 2D image.

Mel Filter Banks

- Mel filter bank representation with 128 filters
- The repetitive nature of the siren can be observed in Figure 2.

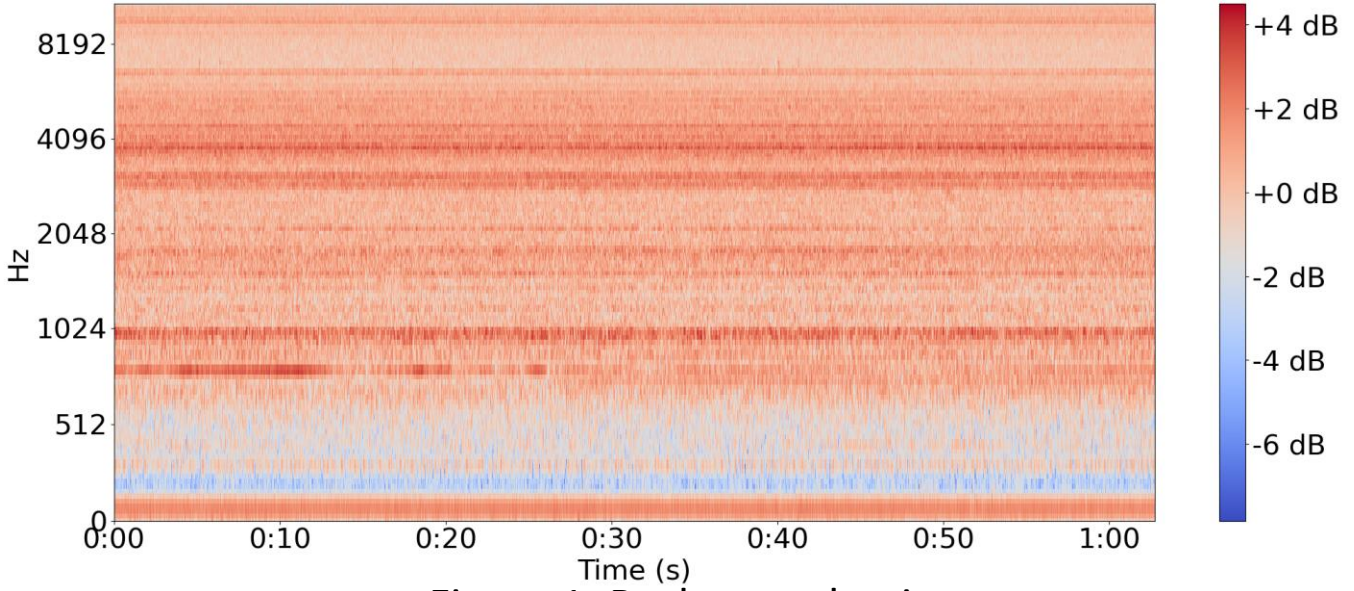


Figure 1: Background noise

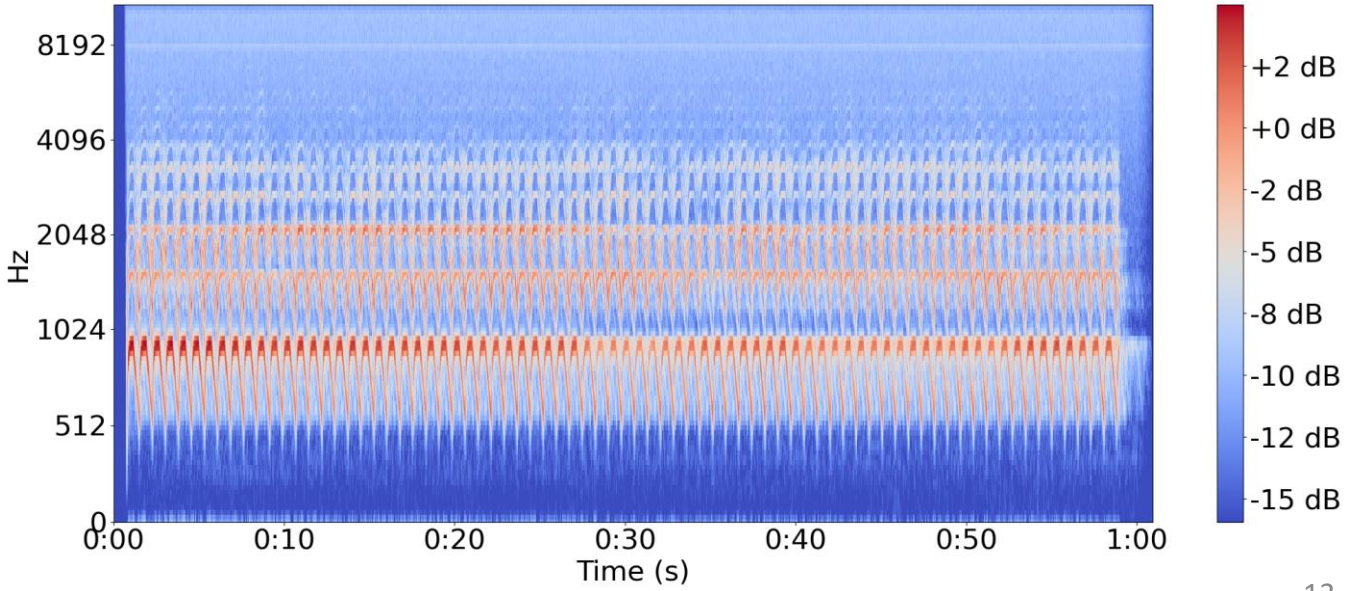
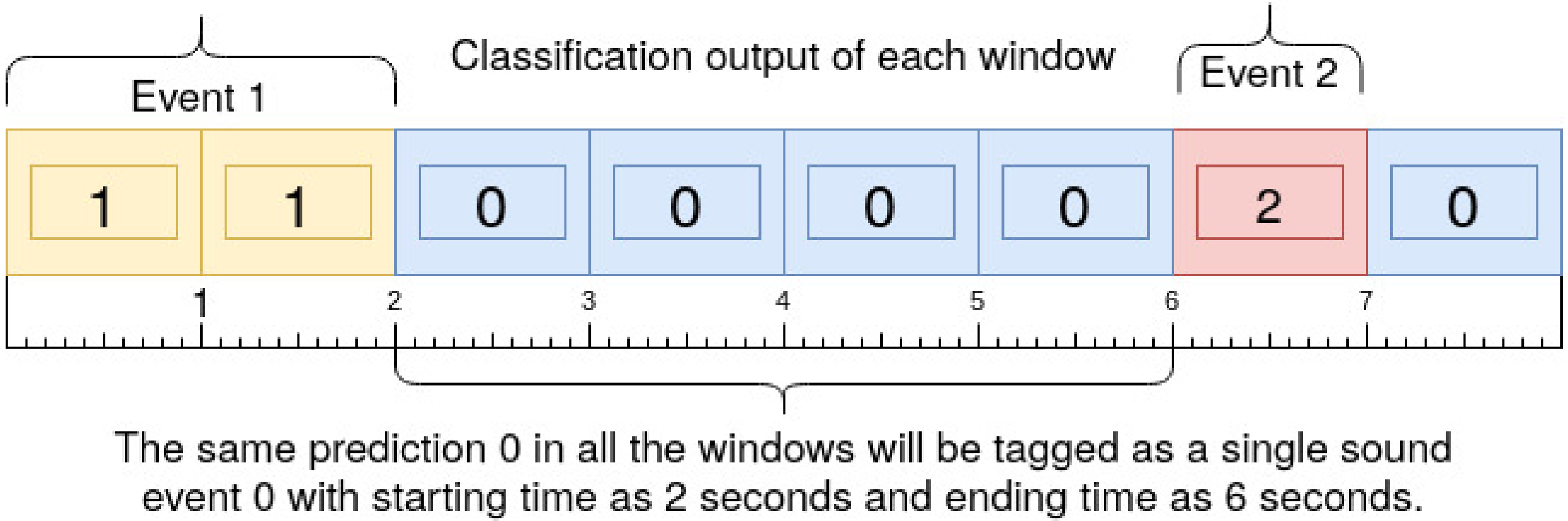


Figure 2: Siren

Model- Audio Spectrogram Transformer(AST)

- Transformer based audio classifier that is SOTA on open-source datasets(Audioset, ESC-50, etc.).(Gong et al (2021))
- Uses spectrogram of audios as input.
- Can be trained with audios with different input window lengths without needing to change the entire architecture.
- Trained models with **30, 15, 10**, and **5** seconds input window sizes.

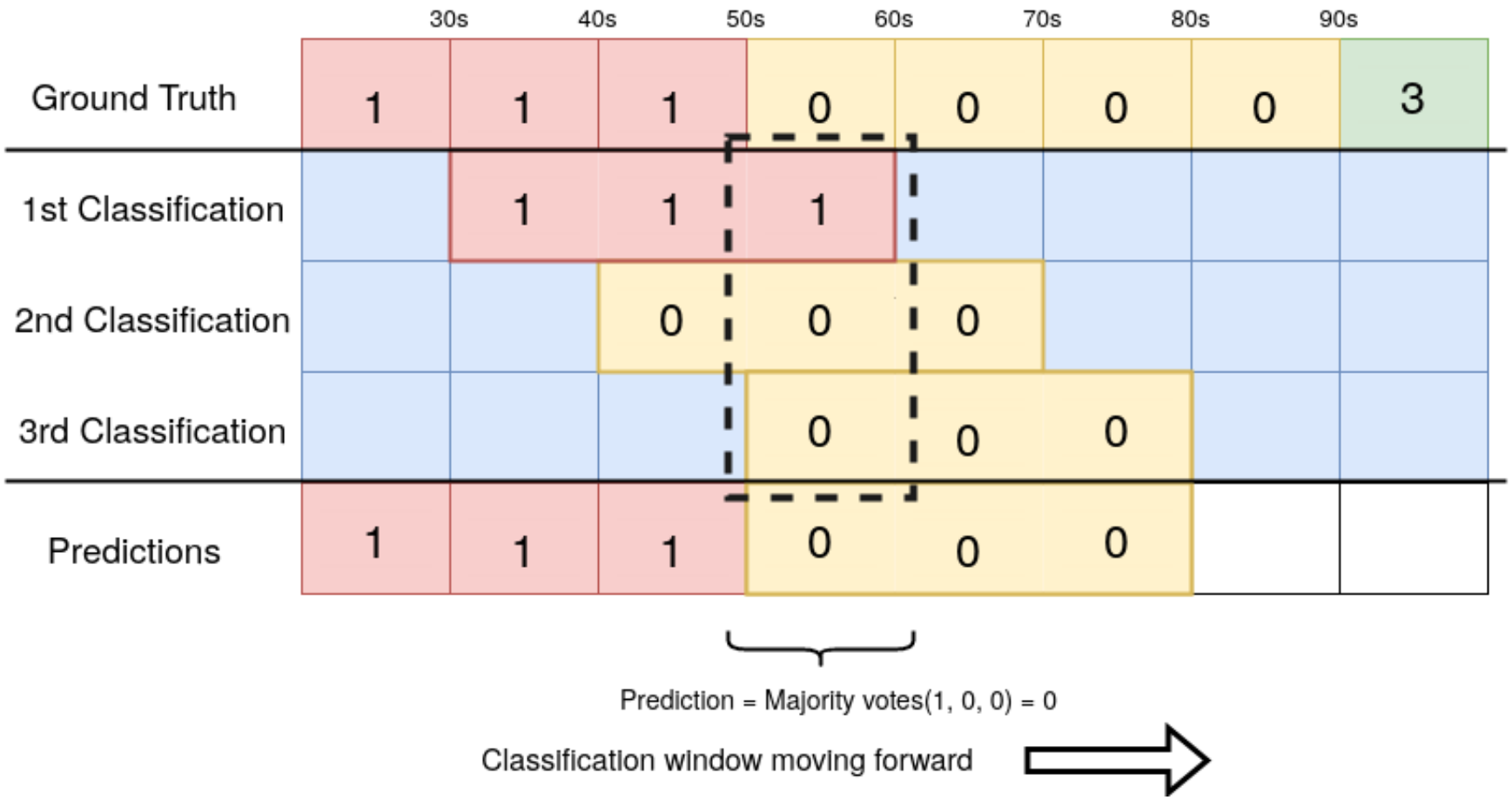
Approach for Audio Tagging(non-overlapping windows)



Hypotheses

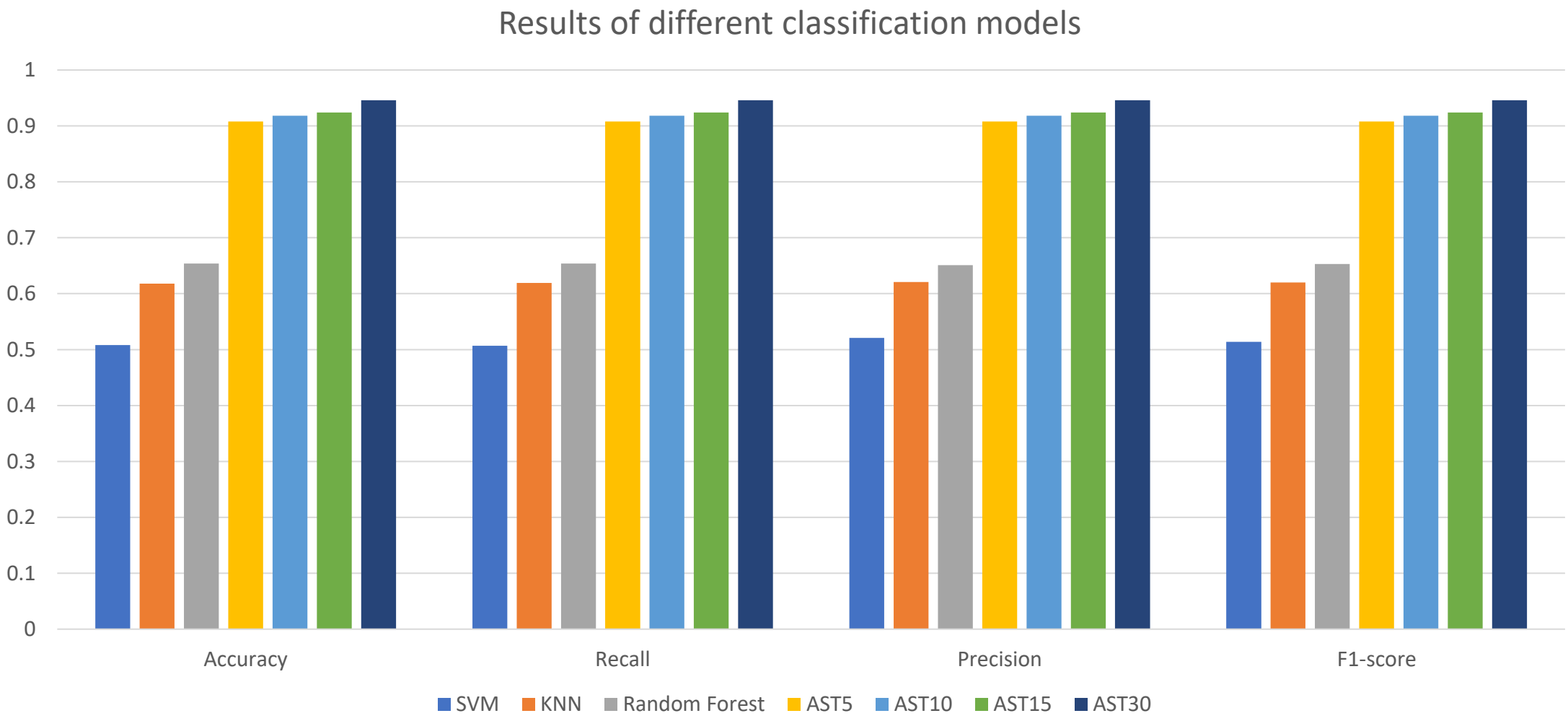
- ***Hypothesis 1***: Different input window sizes would directly affect the performance of classes with different average audio lengths.
- ***Hypothesis 2***: The classification of sound signals with an **overlapping window** will improve the starting point and ending point accuracy of the target sound.

Hypothesis 2



Results

Results of Different models



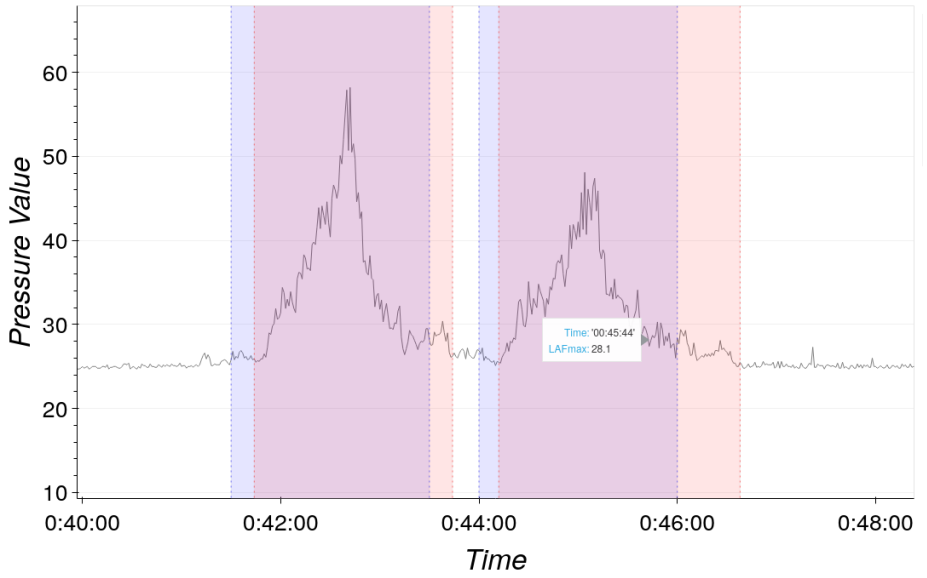
Results

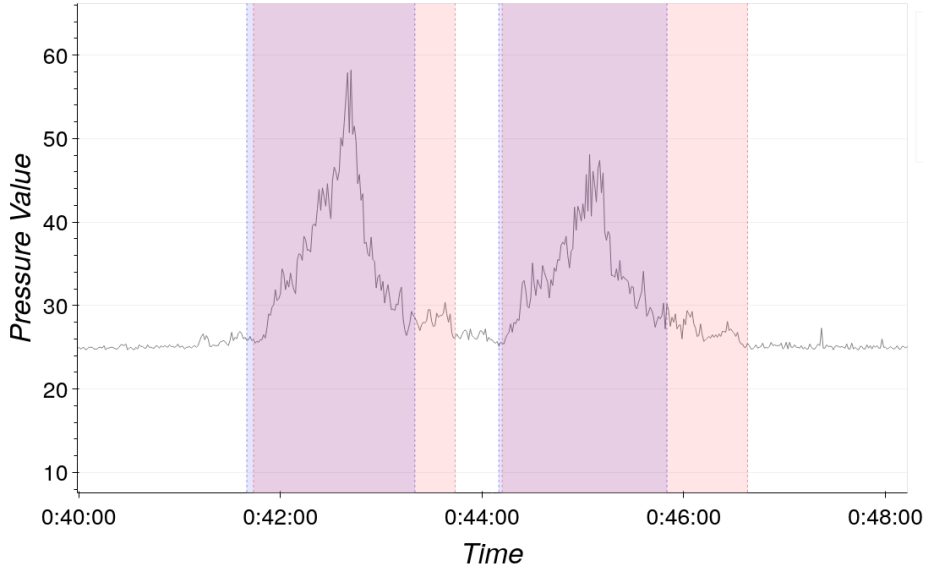
- **Hypothesis 1: Not successful**
- AST_{30} with a window of 30 seconds is better for each class irrespective of the length of the audio.

	Model	Flugzeug	Silence	Nebengeraeusche	Autos
Recall	AST_5	0.921	0.944	0.869	0.898
	AST_{10}	0.908	0.979	0.876	0.909
	AST_{15}	0.925	0.976	0.857	0.936
	AST_{30}	**0.928**	**0.981**	**0.918**	**0.957**
Precision	AST_5	0.897	0.942	0.882	0.91
	AST_{10}	0.922	0.936	0.892	0.921
	AST_{15}	0.914	0.959	**0.915**	0.907
	AST_{30}	**0.949**	**0.982**	0.911	**0.942**
F1-Score	AST_5	0.909	0.943	0.875	0.904
	AST_{10}	0.915	0.957	0.884	0.915
	AST_{15}	0.919	0.968	0.885	0.921
	AST_{30}	**0.938**	**0.982**	**0.914**	**0.95**
Average audio length		68.01	…	27.35	16.52

Results

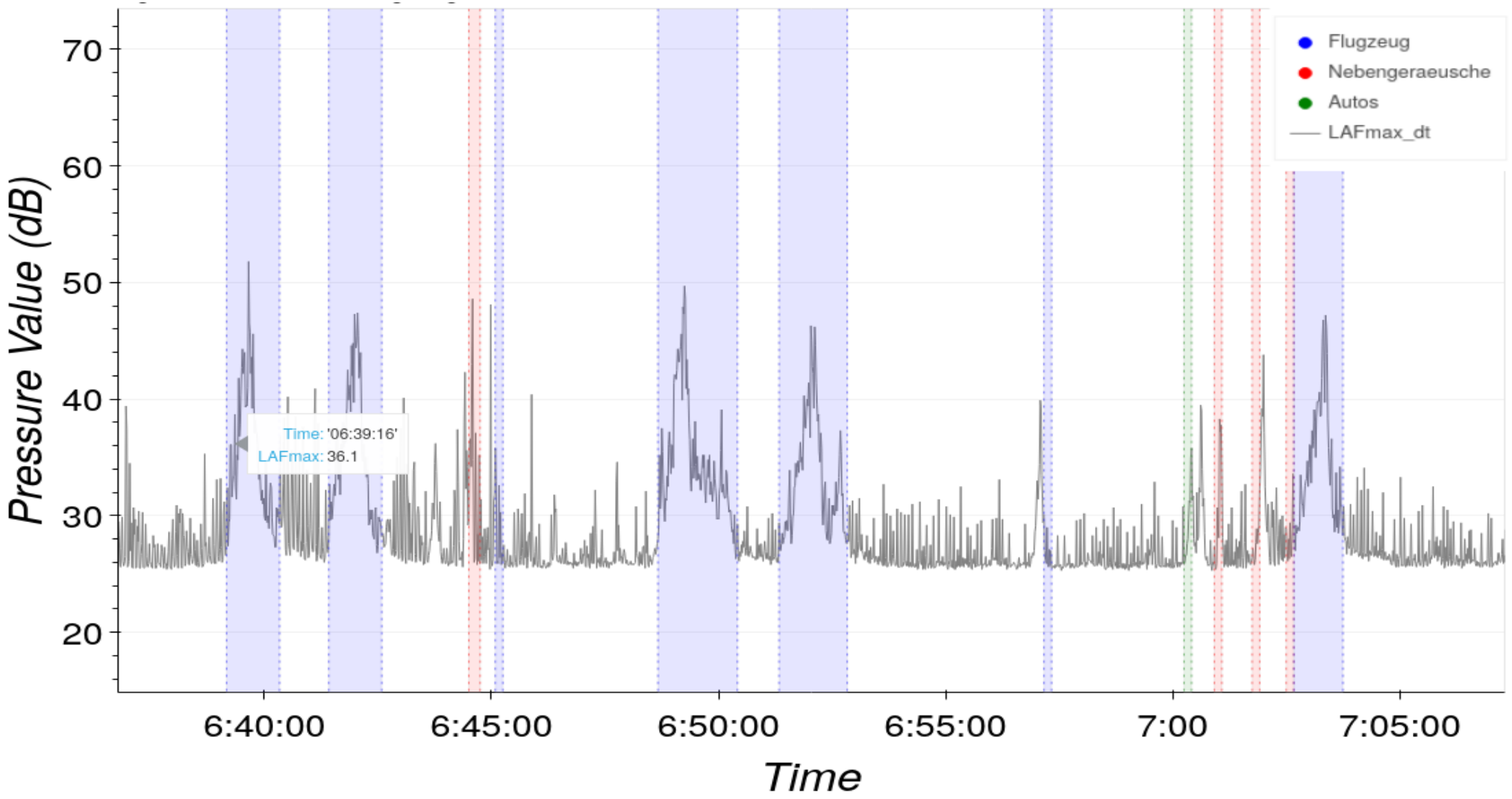
- **Hypothesis 2: Partially successful** as the starting timings of the sound events improved for the events that are **longer** in duration.
- Perform a type of smoothing that negatively impact the sounds with smaller lengths. But it is good for removing noise.





Orange is the ground truth and blue is the prediction.

Decision Support System



Results on a test recording

Overlapping window approach

Recording	Window Size (Overlap)	PSDS	Airplane		Car		Nebengerausche		Silence	
			F1	TPR	F1	TPR	F1	TPR	F1	TPR
135-0061-210521-225750-indoors	15	0.516	0.251	0.885	..	..	0.041	0.25	0.755	0.66
	30	**0.682**	**0.744**	**0.914**	..	..	**0.2**	**0.35**	**0.902**	**0.821**

Non-overlapping window approach

Recording	Window Size	PSDS	Airplane		Car		Nebengerausche		Silence	
			F1	TPR	F1	TPR	F1	TPR	F1	TPR
135-0061-210521-225750-indoors	5	0.151	0.101	0.942	..	..	0.013	0.35	0.177	0.142
	10	0.431	0.176	0.885	..	..	0.051	0.55	0.53	0.392
	15	0.565	0.312	**0.942**	..	..	0.054	0.3	0.767	0.678
	30	**0.708**	**0.688**	0.914	..	..	**0.222**	**0.4**	**0.92**	**0.857**

Conclusion and Future Work

- AST model achieves high performance
- The first hypothesis is not successful, and the second hypothesis is only partially successful.
- Created a Decision support system to tag the sound events

- More classes can be included.
- Classifier can be organized to predict multiple labels.
- Improved overlapping window approach

Thank you

Questions