Siren Sounds as Acoustic Landmarks for Content Verification

Jakob Abeßer, Juan Manuel Rodríguez Mejía, Luca Cuccovillo, Patrick Aichroth

(jakob.abesser@idmt.fraunhofer.de)

Fraunhofer IDMT

Motivation

- **Siren sounds** are **perceptually** very **prominent**, even in complex (urban) sound scenes
- Hypotheses:
  - Characteristic fundamental frequency patterns (**pitch contours**) allow for **recognizing different siren types**
  - Siren sounds can serve as **acoustic landmarks** to **verify claims about the location of ambient audio recordings**

Goals

- **Supervised learning** experiment (using deep residual network) to classify the **siren type (3 classes)** and **country of origin (9 classes)**
- Study on **similarity between pitch contours** across siren classes
  - Study high-dimensional **embedding space derived from computer vision model**
  - Shared pitch contours can lead to **possibly ill-defined classification task**

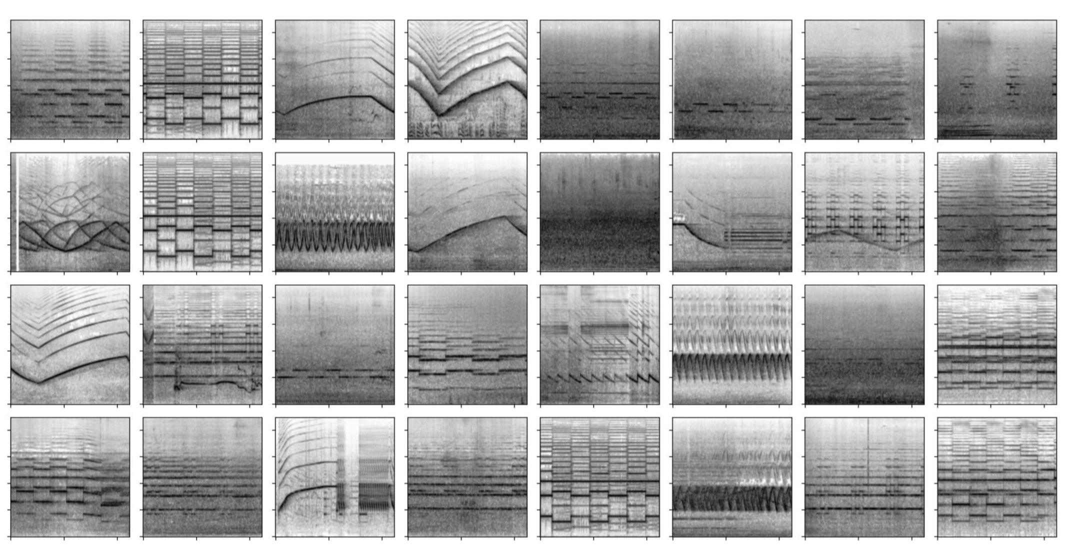
Challenges

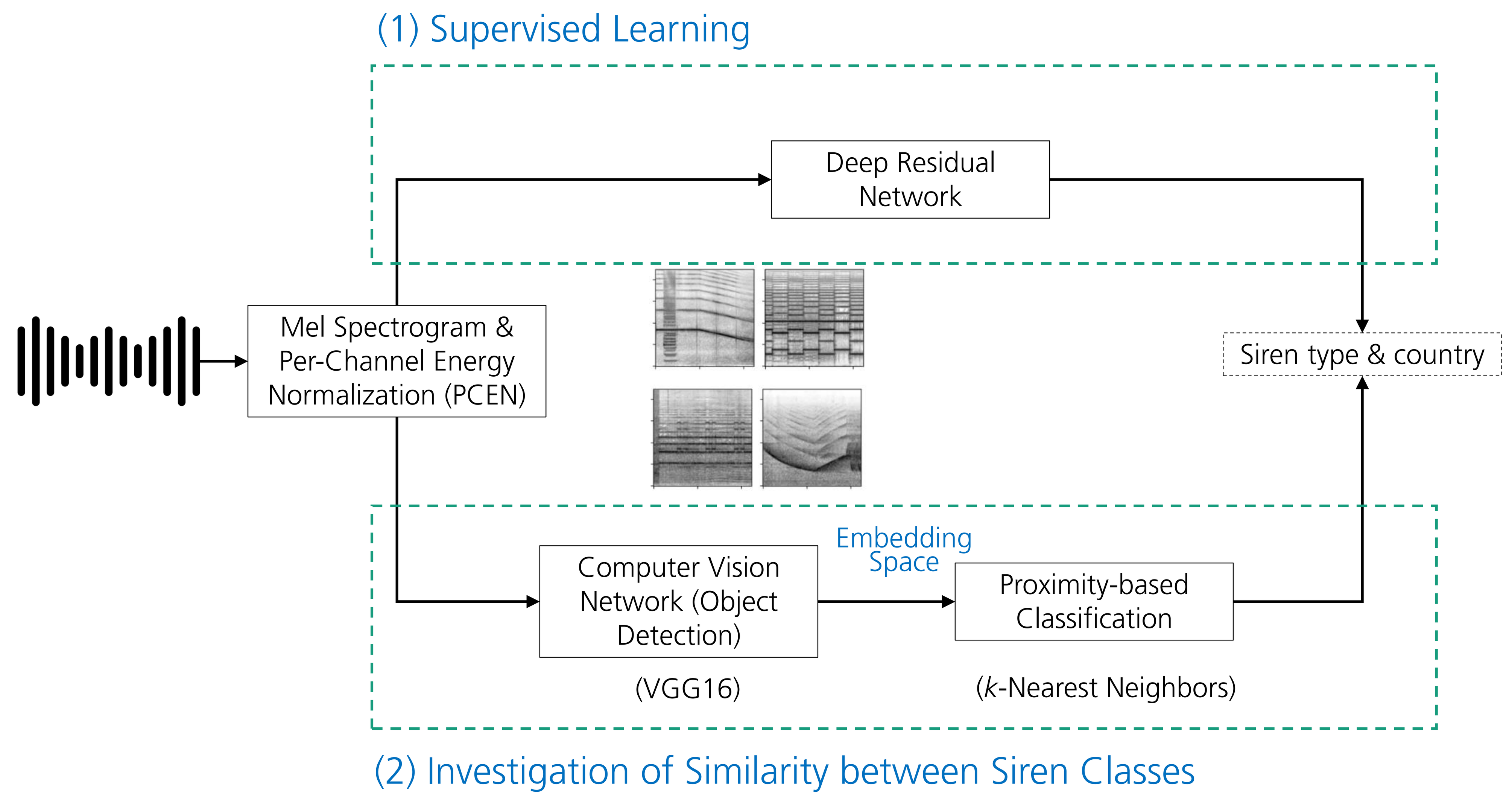
- Some siren sounds are used internationally, which makes it impossible to assign them to specific regions / countries
- Pitch changes due to Doppler shift (produced by passing vehicles) need to be addressed

Dataset & Audio Processing

Category	Classes
Type (3)	Police, Firefighters, Ambulance
Country (9+1)	Canada, China, France, Germany, India, Italy, Japan, Spain, USA, (Non-Siren)

- Mel spectrogram
  - Log-magnitude (or)
  - Per-Channel Energy Normalization (PCEN)
- 44,100 Hz sample rate
- FFT size 2048, Hop length: 1024
- Spectrogram patch duration
  - 2.5s
  - 1.48s
- Patch examples illustrate **large variety of pitch contour types** (stable, sweeps, alternating, etc.)





(1) Supervised Learning

Method

- Deep Residual Neural Network
  - Convolutional Block (64 filters, 5x5 kernel size, ReLU)
  - 4 Residual Blocks (64/64/128/128 filters)
  - Global Average Pooling
  - Around 800k parameters
- Data Augmentation
  - Random Rotate, Grid Distortion, Spec Augment, Random Erasing, Random Brightness, Mixup (all with probability of p=0.5)

Experiment

- Joint Siren type & country classification (27+1 classes)
  - Example: Germany-Police, France-Ambulance, etc.

Patch Length (s)	Spectrogram Type	Accuracy*
1.4	Log Mel	0.46 (0.58)
1.4	PCEN	0.55 (0.75)
2.5	Log Mel	0.50 (0.64)
2.5	PCEN	0.57 (0.72)

*Accuracy values per Spectral Patch or aggregated per Test-File (in brackets)

Results

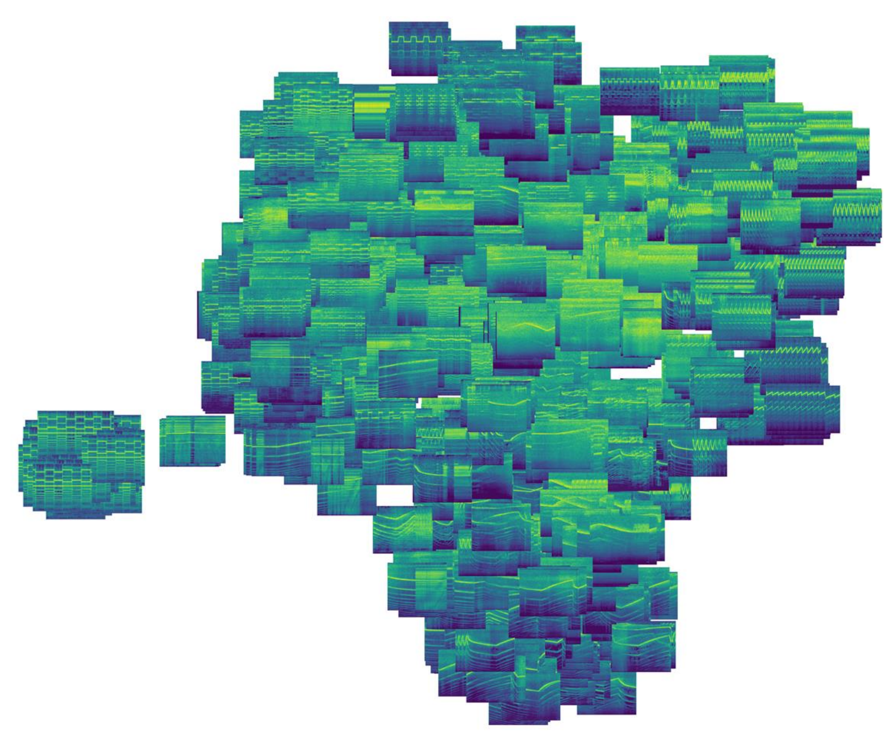
- **PCEN is better** spectrogram representations than log-magnitude Mel spectrogram (Log Mel)
  - PCEN suppresses background noise & enhances salient foreground sounds
- **Longer spectrogram patches (2.5 s) improved results**
  - Hypothesis: especially relevant for slowly changing pitch contours

(2) Investigation of Similarity between Siren Classes

Method

- Convert spectrogram patches into images (224x224 pixels)
- Compute embeddings from pre-trained VGG16 model

t-SNE embedding space visualization



Confusion matrix for siren country classification (A=0.44)

	Canada	China	France	Germany	India	Italy	Japan	Spain	USA
Canada	0.64	0.12	0.12	0.00	0.00	0.00	0.04	0.08	0.00
China	0.27	0.20	0.20	0.07	0.07	0.07	0.13	0.00	0.00
France	0.00	0.06	0.65	0.06	0.00	0.06	0.12	0.06	0.00
Germany	0.00	0.00	0.41	0.45	0.00	0.09	0.05	0.00	0.00
India	0.25	0.08	0.12	0.04	0.21	0.04	0.17	0.04	0.04
Italy	0.00	0.05	0.63	0.11	0.00	0.21	0.00	0.00	0.00
Japan	0.00	0.12	0.12	0.06	0.00	0.00	0.65	0.06	0.00
Spain	0.15	0.00	0.35	0.15	0.00	0.10	0.05	0.20	0.00
USA	0.19	0.10	0.33	0.05	0.10	0.00	0.05	0.00	0.19

Results

- CV model embeddings **clusters similar pitch contours** (**interpretable embedding space** structure)
- Confusion matrix reveals **shared siren sounds across siren types and countries** (demonstrate that classification task is somewhat **ill-defined**)

DAGA 2024 HANNOVER

vera.ai