Detecting Presence Of Speech In Acoustic Data Obtained From Beehives

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

data science

Pascal Janetzky
janetzky@informatik.uni-wuerzburg.de

Padraig Davidson
davidson@informatik.uni-wuerzburg.de

Michael Steininger
steininger@informatik.uni-wuerzburg.de

Anna Krause
anna.krause@informatik.uni-wuerzburg.de

Andreas Hotho
hotho@informatik.uni-wuerzburg.de

Motivation

As part of the We4Bee project, smart beehives are distributed to educational institutions all over the world. The hives are equipped with multiple sensors, among them microphones. The recording started last year and is still running.

In this process, audio data is collected, which supports the monitoring of bee colonies on the communication level. Such recording allow us to monitor, e.g., swarming behaviour and colony health.

Challenge

The beehives are predominantly placed in public space, where we require the written permission of all users. Due to the project's open and distributed nature, this is prohibitive. Further, the frequencies of human speech largely overlap with the bees' humming.

Approach

We construct an initial dataset and use Neural Networks to detect the presence of speech. This eliminates the need for obtaining written permission, since we can discard all positive samples and only upload approved recordings to the cloud.

Dataset

We took audio records from three days from August, September, October 2020 and manually labelled a portion of them as negative ("0", no speech) or positive ("1", with speech). Each sample is 60s and 44.1 kHz.

	0	1
Train	119	16
Validation	27	3
Test	25	10

Methods

We utilize three CNNs in a Siamese setup and input the raw audio. The *kapre* (Choi et al., 2017) package is used to generate Mel spectrograms as part of the forward pass.

- **Bulbul** (Grill & Schlüter, 2017): BatchNorm (BN) followed by four conv-relu-pool stacks
- **Saeed** (Saeed, 2016): Merge Mel spectrogram with delta features, followed by four conv-bn-relu-pool stacks
- **ESC** (Piczak, 2015): Mel spectrogram merged with its delta, followed by conv-dropout-pool-conv-pool-dense

During test time, we use a kNN classifier (trained on the learned training embeddings) to predict a test sample's class.

tl;dr

Use Siamese Neural Networks to handle imbalanced dataset, add trained kNN classifier during test time

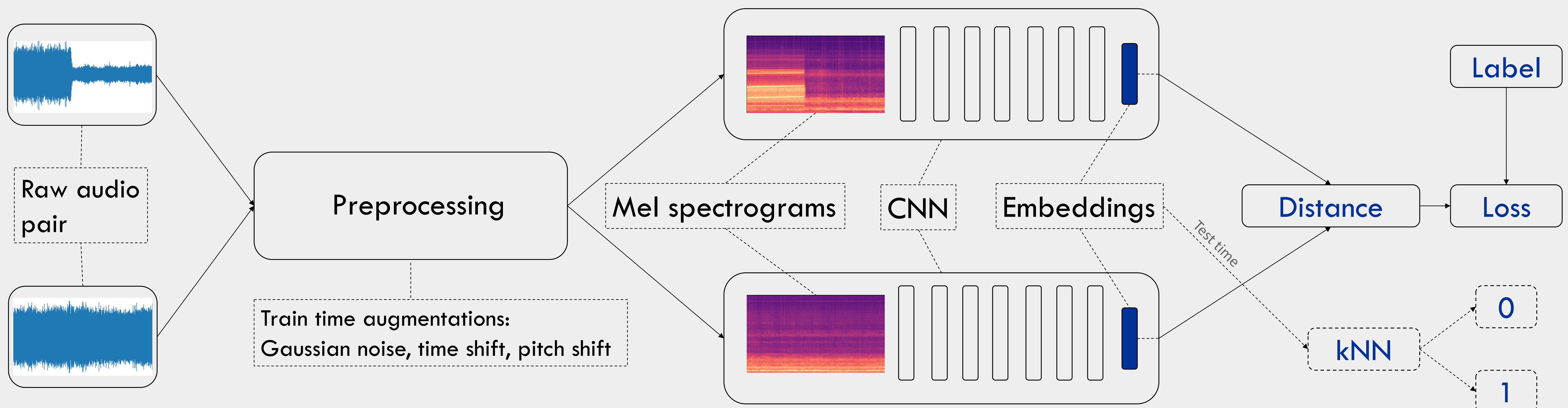
Results

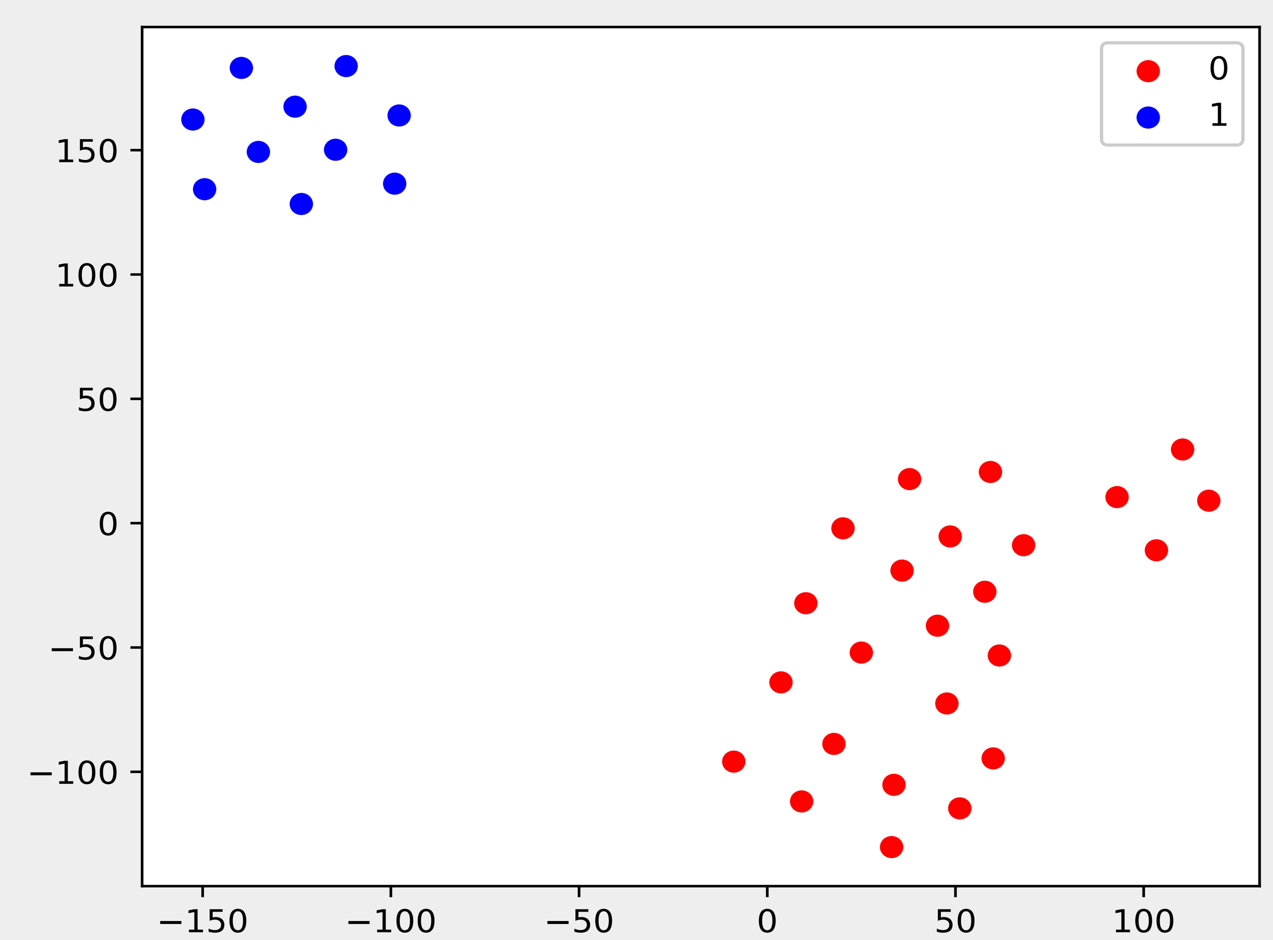
We find that pairing common Siamese convolutional neural networks with kNNs is a viable approach. The usage of augmentations during training leads to very good recall and F1 scores. The learned embeddings reflect this.

	Saeed		Bulbul		ESC	
k	AUROC	Recall speech	AUROC	Recall speech	AUROC	Recall speech
1	1.0 $\pm$ 0.0	1.0 $\pm$ 0.0	1.0 $\pm$ 0.0	1.0 $\pm$ 0.0	0.495 $\pm$ 0.020	0.04 $\pm$ 0.09
3	1.0 $\pm$ 0.0	1.0 $\pm$ 0.0	1.0 $\pm$ 0.0	1.0 $\pm$ 0.0	0.51 $\pm$ 0.05	0.04 $\pm$ 0.09
5	1.0 $\pm$ 0.0	1.0 $\pm$ 0.0	1.0 $\pm$ 0.0	1.0 $\pm$ 0.0	0.52 $\pm$ 0.07	0.06 $\pm$ 0.13

AUROC and recall scores for the Siamese Networks for different numbers of nearest neighbours (k). Both **Saeed** and **Bulbul** achieve perfect scores. The **ESC** network scores worse, which we attribute to the more shallow architecture.

Siamese Network





A t-SNE plot of the learned embeddings for the test data. The **Bulbul** network can perfectly separate the classes.

Test embeddings

Choi, K., Joo, D., & Kim, J. (2017). Kapre: On-gpu audio preprocessing layers for a quick implementation of deep neural network models with keras. *arXiv preprint arXiv:1706.05781*
Grill, T. & Schlüter, J."Two convolutional neural networks for bird detection in audio signals," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1764–1768
Piczak, K. J. , "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6
Saeed, A. (2016) Urban Sound Classification. [Online]. Available: http://aqibsaeed.github.io/2016-09-24-urban-sound-classification-part-2/