Acoustic Event Detection Using Speaker Recognition Techniques: Model Optimization and Explainable Features

Mattson Ogg and Ben Skerritt-Davis

Johns Hopkins University Applied Physics Laboratory

Objectives

Aim 1: Examine whether speaker-ID approaches can be successfully applied to acoustic event recognition.

- Sub-aim: Study adaptations to speaker-ID methods to better fit non-speech sound recognition.

Aim 2: Provide insight into how networks perform acoustic event recognition.

What acoustic features are most important for distinguishing among different classes of sounds?

- Deploy representational similarity analysis (*RSA*, Kriegeskorte & Kievit, 2013) to explore the information encoded in network embeddings.

Study Design

Used the FSDKaggle2018 (Fonseca et al., 2018) dataset: Experiments carried out on a validation partition that we created by holding out 1/3 of the manually labelled training examples.

Benchmarked performance against the published baseline (mAP@3 = 0.70).

Harder baseline derived from Google YAMNet algorithm (Hershey et al., 2016): single fully connected layer between YAMNet embeddings and 41 output units for the corpus' target classes: Validation performance: accuracy = 0.79, mAP@3 = 0.86; Test performance: accuracy = 0.78, mAP@3 = 0.85.

We implemented a Time-Delay Neural Network (TDNN) modelled after Snyder et al., 2017 in PyTorch:

- 5-Layer TDNN operating at the frame level (512 units per layer).
- 1500 units to calculate stats pooling (3000 after mean and std calculation).
- 2 fully-connected layers operating at the segment (file) level.
- Trained for 100 epochs measuring the accuracy on the validation partition.

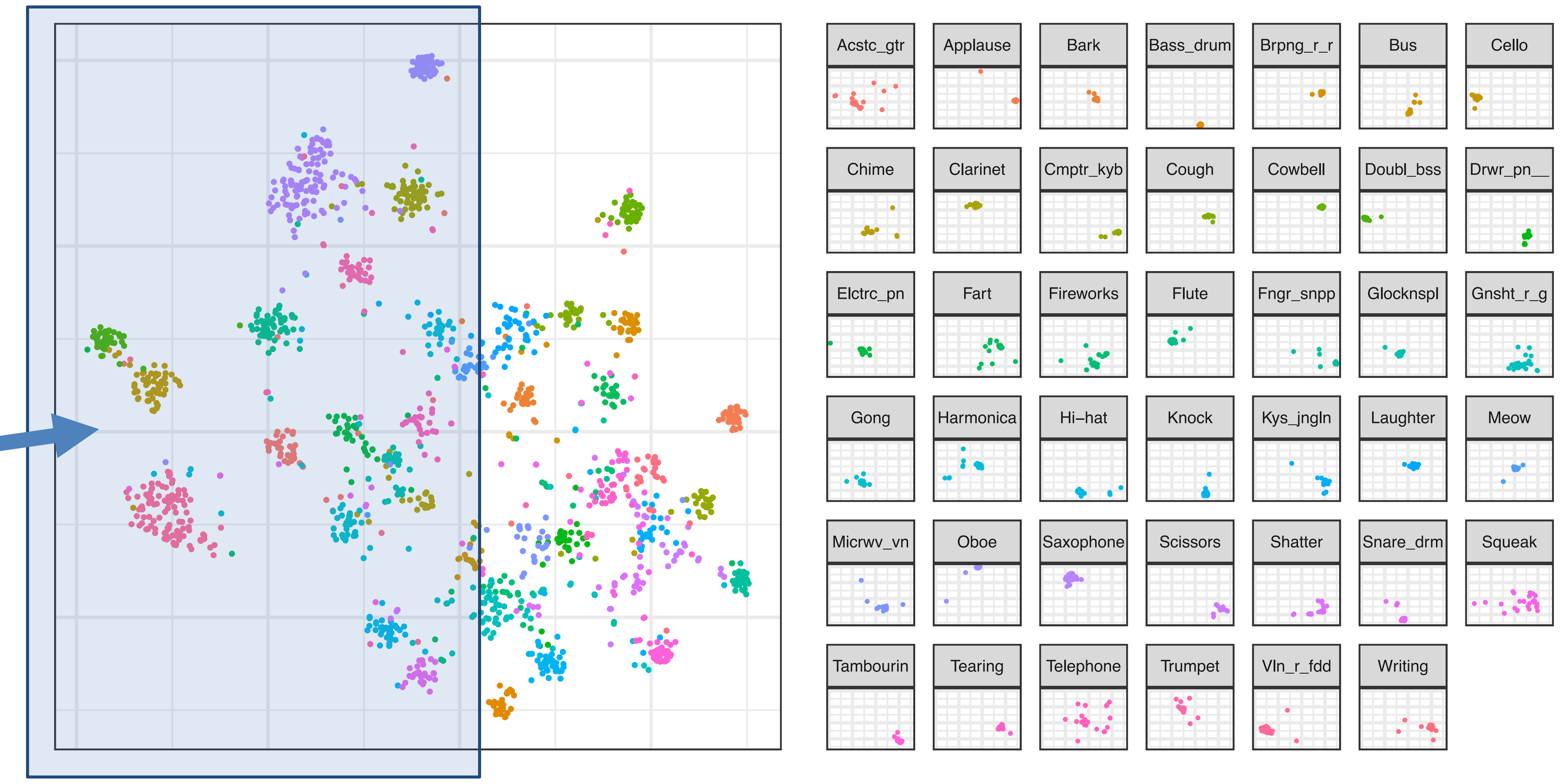
Results: Speaker-ID TDNN Model Optimization Experiments

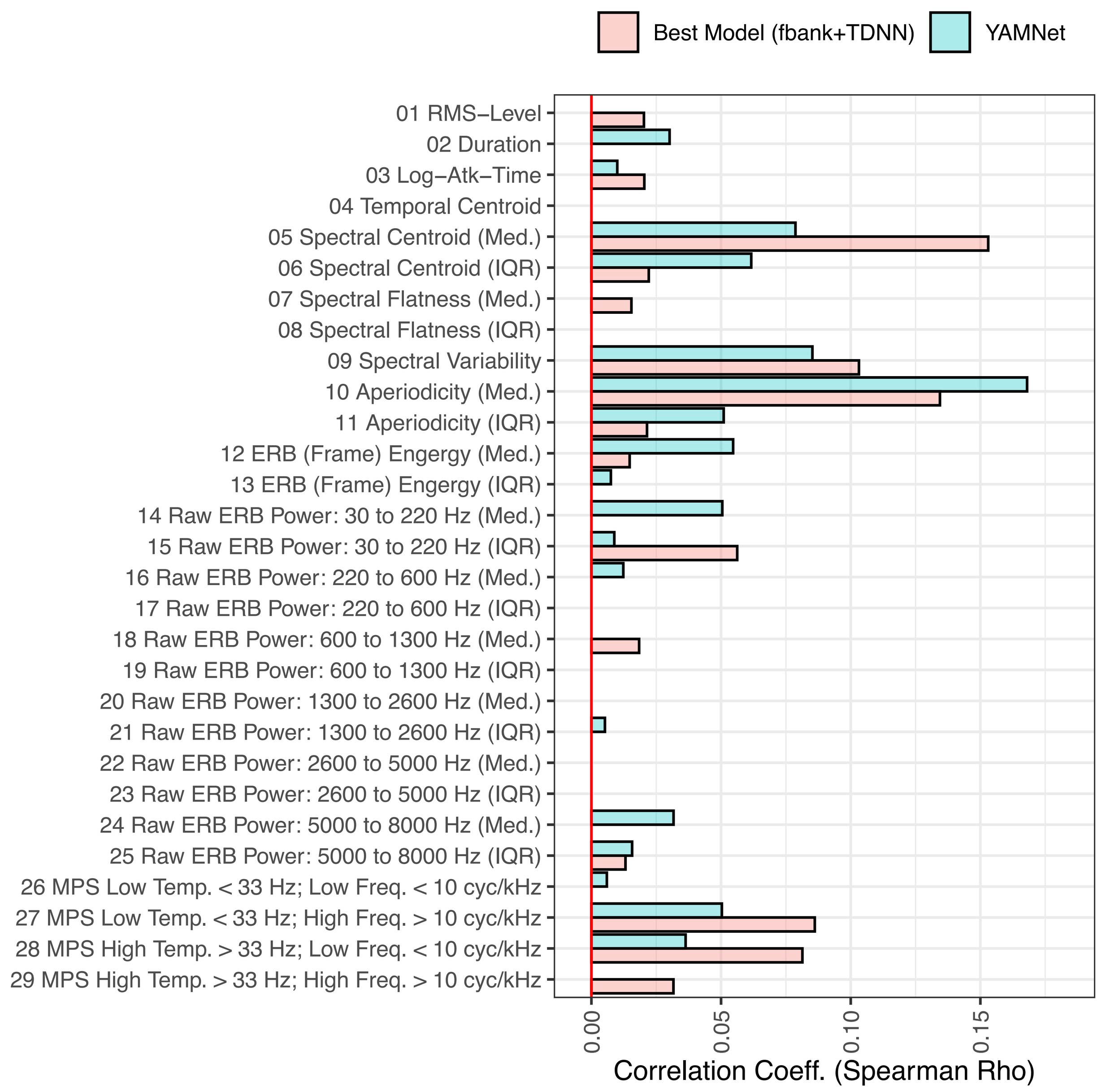
	Accuracy				mAP@3			
	Cochleagram	Mel-Filterbank	MFCC	Spectrogram	Cochleagram	Mel-Filterbank	MFCC	Spectrogram
Initial Baseline TDNN Model	0.73(Δ0)	0.74(Δ0)	0.24(Δ0)	0.75(Δ0)	0.79(Δ0)	0.8(Δ0)	0.32(Δ0)	0.81(Δ0)
Diff. Maps	0.74(Δ0.016)	0.75(Δ0.012)	0.16(Δ-0.078)	0.75(Δ0.002)	0.8(Δ0.006)	0.81(Δ0.007)	0.24(Δ-0.087)	0.81(Δ-0.005)
Reverb Aug.	0.77(Δ0.043)	0.79(Δ0.043)	0.7(Δ0.464)	0.77(Δ0.022)	0.82(Δ0.03)	0.84(Δ0.035)	0.78(Δ0.453)	0.82(Δ0.011)
Speed Aug.	0.83(Δ0.099)	0.88(Δ0.134)	0.76(Δ0.526)	0.87(Δ0.117)	0.87(Δ0.079)	0.91(Δ0.103)	0.82(Δ0.493)	0.9(Δ0.089)
Smaller Net: 256 Units	0.72(Δ-0.004)	0.76(Δ0.019)	0.45(Δ0.218)	0.74(Δ-0.01)	0.78(Δ-0.012)	0.82(Δ0.013)	0.56(Δ0.234)	0.8(Δ-0.011)
Larger Net: 1024 Units	0.74(Δ0.015)	0.75(Δ0.007)	0.4(Δ0.166)	0.76(Δ0.007)	0.8(Δ0.009)	0.81(Δ0.007)	0.5(Δ0.178)	0.81(Δ0.001)
Reduced Context-Layer	0.72(Δ-0.011)	0.74(Δ-0.002)	0.43(Δ0.194)	0.73(Δ-0.02)	0.78(Δ-0.018)	0.81(Δ0.002)	0.54(Δ0.216)	0.8(Δ-0.015)
Added Context-Layer	0.74(Δ0.007)	0.75(Δ0.011)	0.56(Δ0.326)	0.74(Δ-0.012)	0.79(Δ-0.001)	0.81(Δ0.001)	0.65(Δ0.332)	0.8(Δ-0.015)
Batch-Norm, Drop-Out	0.76(Δ0.031)	0.8(Δ0.055)	0.63(Δ0.39)	0.78(Δ0.033)	0.81(Δ0.016)	0.85(Δ0.044)	0.71(Δ0.392)	0.84(Δ0.026)
Speed+Reverb Aug.	NA	0.87(Δ0.126)	NA	NA	NA	0.9(Δ0.093)	NA	NA
Speed+Reverb, Diff. Maps, Batch-Norm+Drop-Out	NA	0.86(Δ0.113)	NA	NA	NA	0.89(Δ0.084)	NA	NA
Speed+Reverb, Batch-Norm+Drop-Out	NA	0.85(Δ0.109)	NA	NA	NA	0.89(Δ0.084)	NA	NA
Speed+Reverb, Diff. Maps	NA	0.87(Δ0.129)	NA	NA	NA	0.91(Δ0.1)	NA	NA

Mel-Filterbank TDNN trained with speed-augmented data performed best (Test: accuracy = 0.82, mAP@3 = 0.86)

Results: Explainable Network Representations

t-SNE visualization of our model's embedding space revealed a structure that corresponded (at least) to the presence or absence of prominent pitch content. (e.g., Instruments, Chimes, Telephone ring etc.)





Representational Similarity Analysis:

Probed embeddings via Semi-partial Spearman correlations of dissimilarity matrices (DSM):

- *Network-DSMs*: cosine distances among a network's embeddings for each pair of test items.
- *Acoustic-DSMs*: difference between each pair of test items along well-studied acoustic dimensions.

Our model and the YAMNet model's embedding spaces were only modestly correlated (r_s = 0.31).

Both models' performance associated with aperiodicity, spectral centroid, and spectral variability.

- Similar features influence human listeners' perception (Ogg & Slevc, 2019)

References

N. Kriegeskorte and R. A. Kievit, "Representational geometry: Integrating cognition, computation, and the brain," *Trends Cognit. Sci.*, vol. 17, no. 8, pp. 401–412, Aug. 2013.

E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *Proc. IEEE DCASE*, 2018, pp. 69–73.

S. Hershey, S. et al., "CNN architectures for large-scale audio classification," in *IEEE ICASSP*, 2017, pp. 131–135.

D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, pp. 999–1003, 2017.

M. Ogg, and R. L. Slevc, "Acoustic correlates of auditory object and event perception: Speakers, musical timbres, and environmental sounds," *Front. Psychol.*, vol. 10, 1594, 2019.