A Study on Robustness to Perturbations for Representations of Environmental Sound

Sangeeta Srivastava, Ho-Hsiang Wu, Joao Rulff, Magdalena Fuentes, Mark Cartwright, Claudio Silva, Anish Arora, Juan Pablo Bello

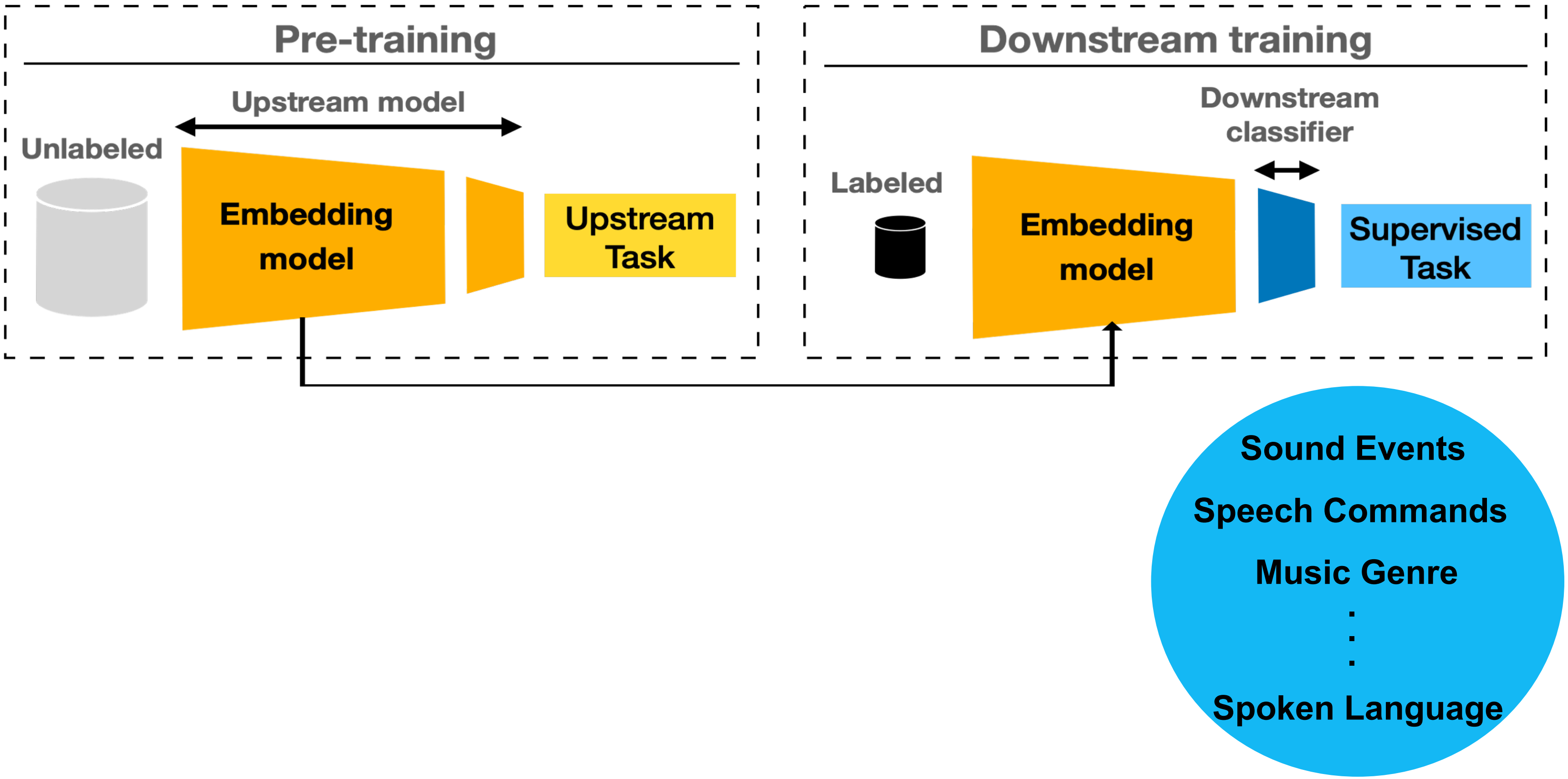
NYU

SONYC SOUNDS OF NEW YORK CITY

THE OHIO STATE UNIVERSITY

30th EUSIPCO BELGRADE/SERBIA 29 AUG - 2 SEP 2022

Self-supervised Learning



Evaluating Generalization

Generalization Question
Does the audio embeddings yield a good classifier for all audio related tasks/datasets?

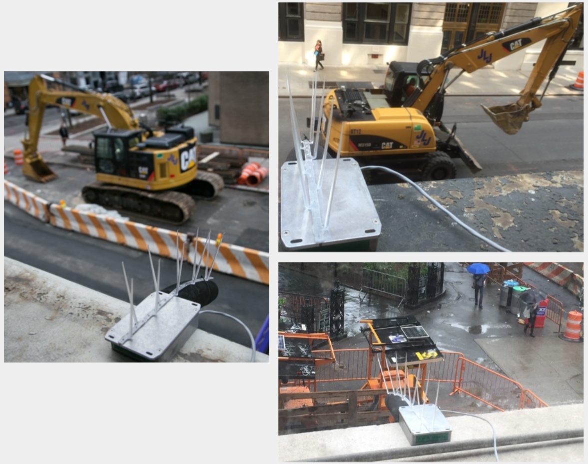
Evaluation suite	#audio models	#tasks
HARES[1] (Holistic Audio Representation Evaluation Suite)	13	12
HEAR[2] (Holistic Evaluation of Audio Representations)	29	19

[1] Turian, Joseph, et al. "HEAR: Holistic Evaluation of Audio Representations." arXiv preprint arXiv:2203.03022 (2022).
[2] Wang, Luyu, et al. "Towards learning universal audio representations." ICASSP 2022.

Limitations of the Evaluation Suites

- Lack of variations within the dataset
  - Evaluation dependent on the variability already captured in the datasets
  - Evolving test scenarios in the same data domain
- Need for annotations
  - Dependency on the presence of labels in the downstream tasks

Case Study: Environmental Sound Detection



New deployment

New acoustic conditions

- Channel effects - Variations in:
  - Acoustic conditions
  - Microphone ranges

Evaluating Robustness

Robustness Question

If there is a change in the input that does not change the semantics of the sound, does the new embedding space also preserve them?

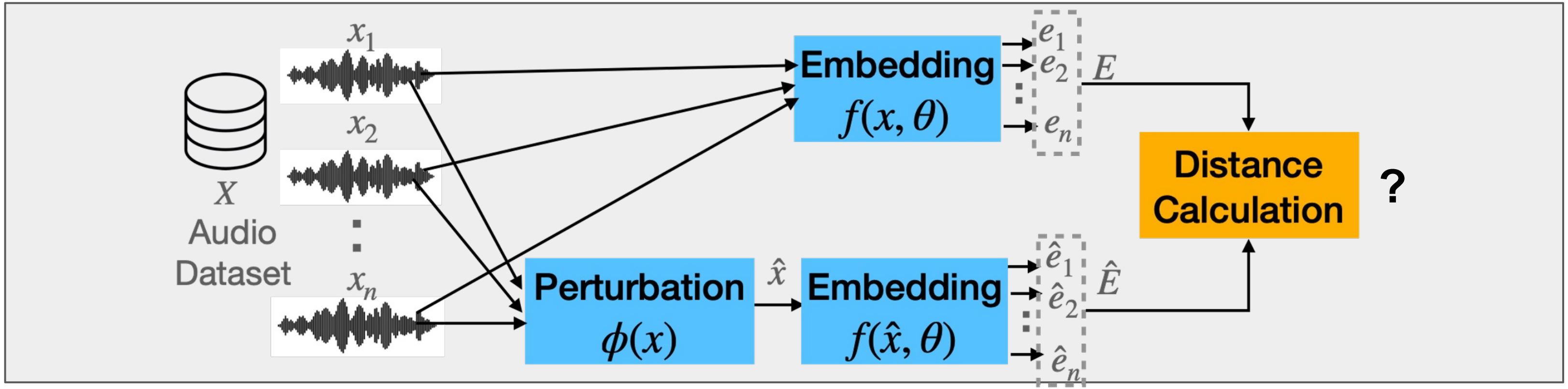
Goal

Evaluate the **robustness of the audio embeddings** against variations caused by myriad microphones' range and acoustic conditions (i.e. **channel effects**) for environmental sound detection

Proposed Solution

Proposed Solution
Variability: Artificial degradation of signals by applying different mathematical transformations or **perturbations**
Task-free: Distance metrics to quantify **shift in the embedding space** directly

Experimental Pipeline



X
UrbanSound8K
SONYC-UST

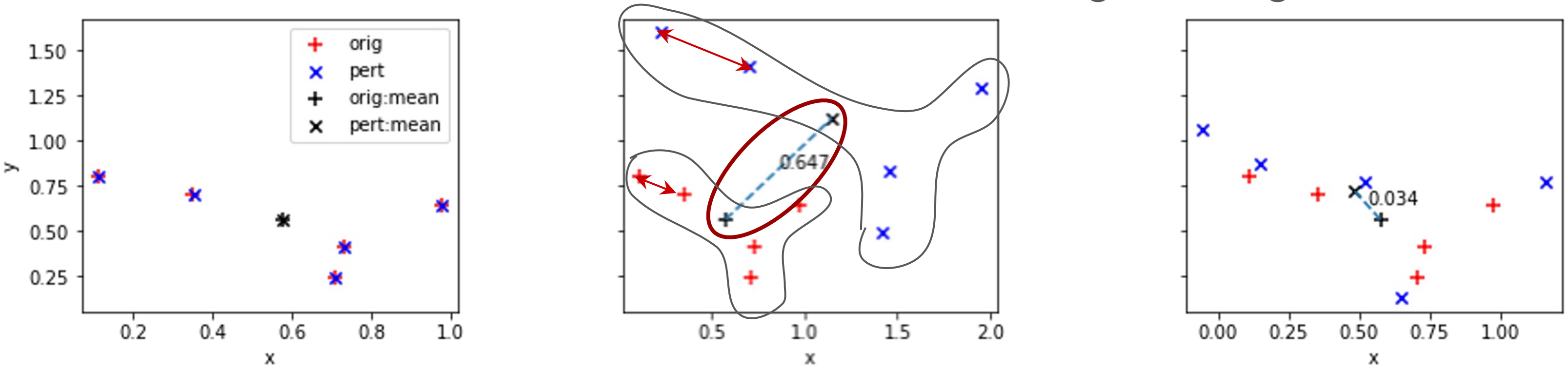
ϕ

Pert. Type	Pert. Values
High Pass	{100, 200, 400, 800, 1600, 4k} Hz
Low Pass	{8k, 4k, 1600, 800, 400} Hz
Reverberation	{25, 50, 75, 100} %
Gain	{3, 6, 10, 20, 30} dB

θ
OpenL^3
YAMNet

Shift in the Embedding Space

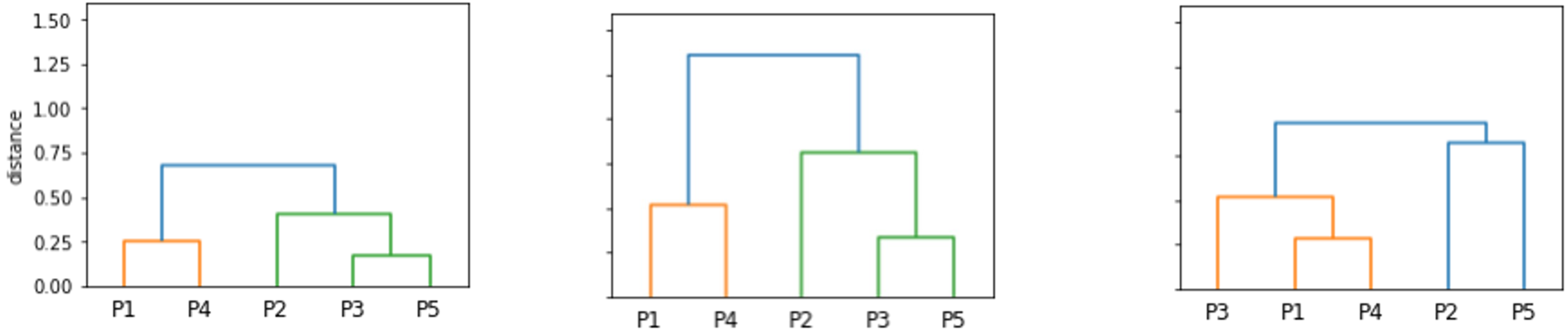
- Metrics should inform on how the classifications might change



Change in mean

Increase in pairwise distances

Similar overall structure



Metrics

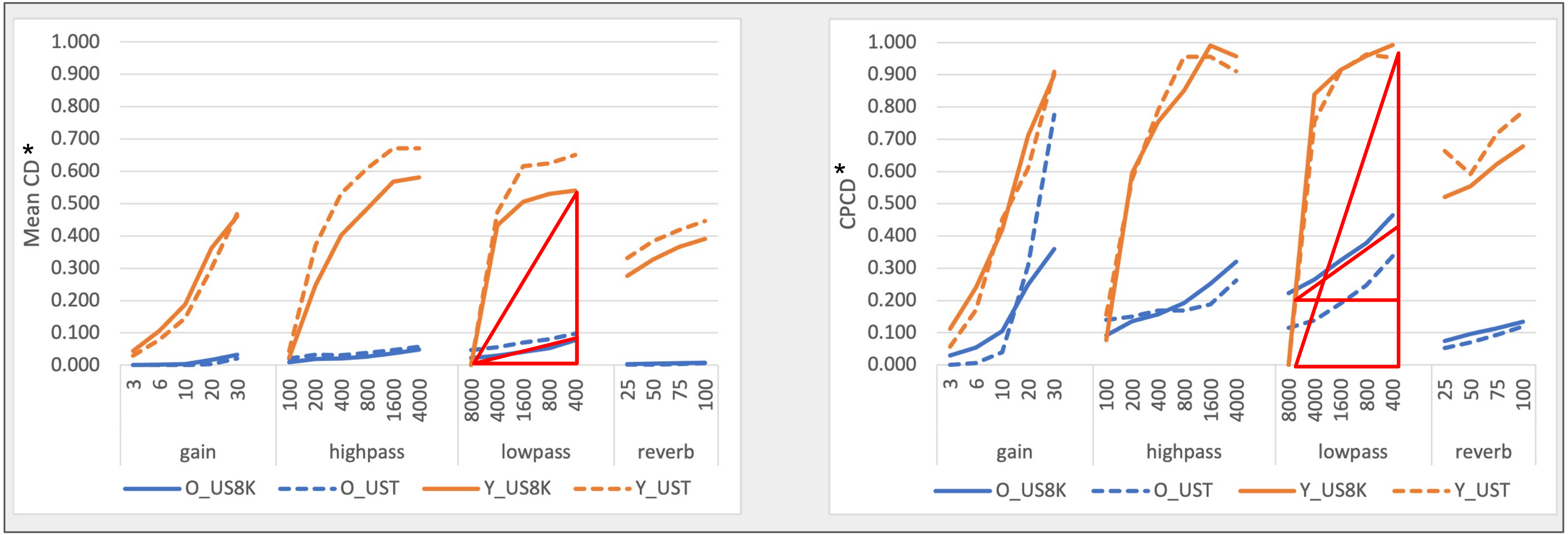
- Absolute pairwise distances
  - Mean Cosine Distance (CD)
- Relative distances
  - How much the dendrograms of the hierarchical clustering change?
  - Cophenetic Correlation Distance (CPCD)
- Distribution shift
  - Assumption: Distributions are Gaussian
  - Fréchet Audio Distance (FAD[3])

[3] K. Kilgour, et al. "Fréchet Audio Distance: A reference-free metric for evaluating music enhancement algorithms." INTERSPEECH, 2019.

Evaluation

- How do the representation types OpenL^3 and YAMNet compare?
- How does the shift compare with the downstream performance?
- What is the effect in each perturbation type?

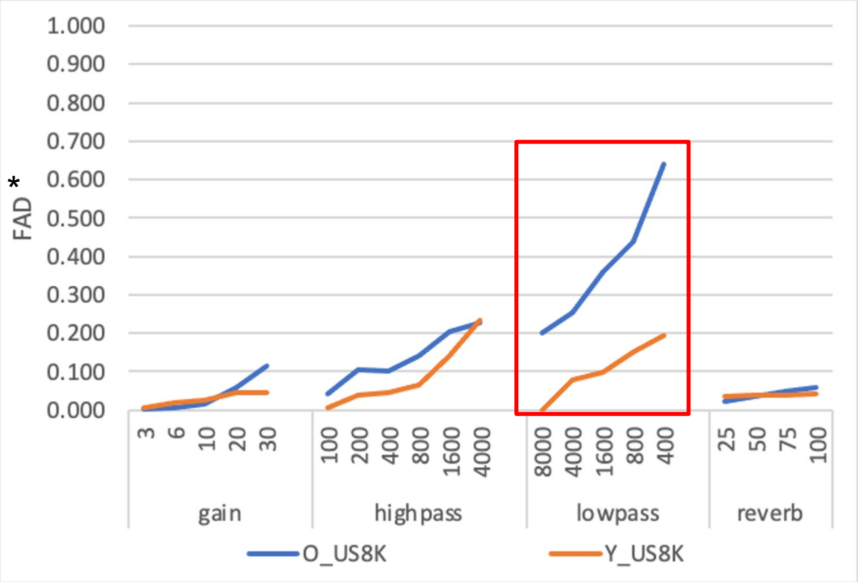
Comparison of Representation Types

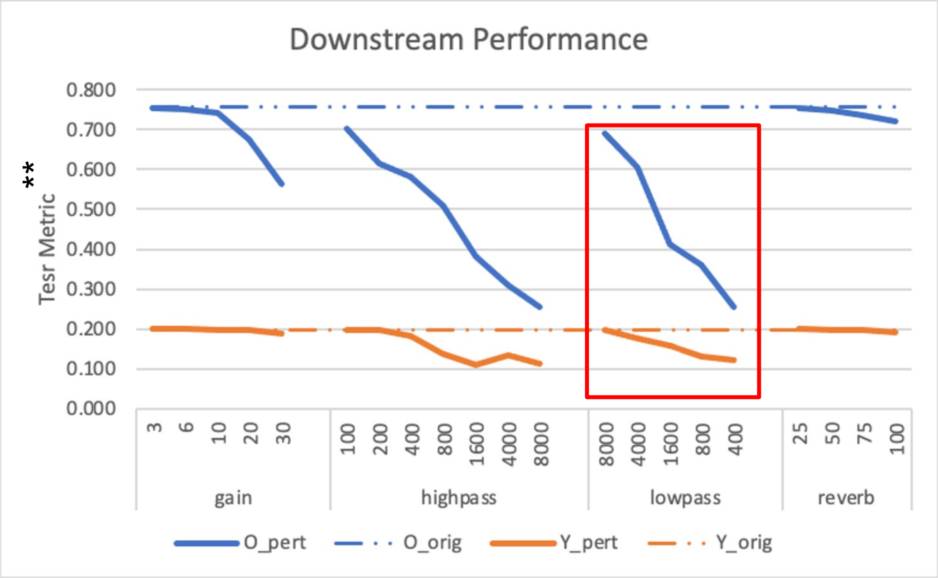


** Smaller value is preferred*

- YAMNet exhibits higher sensitivity as compared to OpenL^3
  - Larger slope -> more sensitive to change

Distance Metrics and Downstream Evaluation

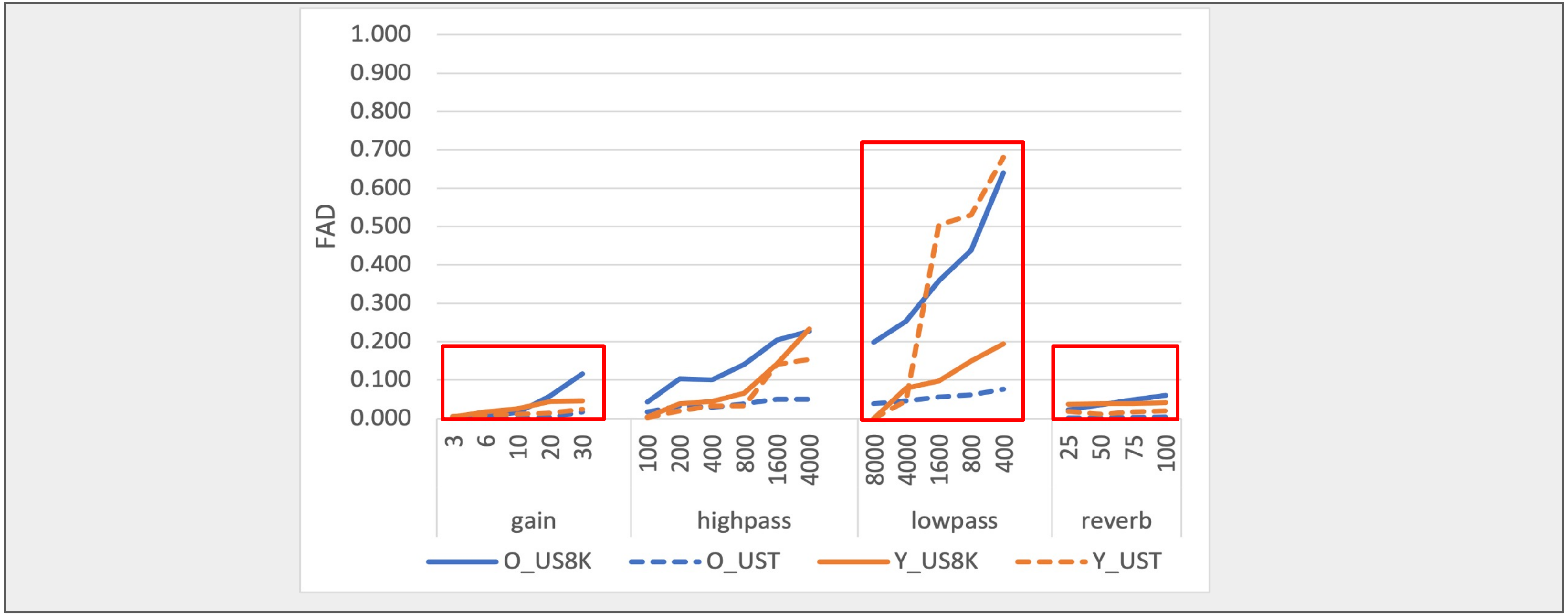




** Smaller value is preferred*
*** Larger value is preferred*

- FAD inversely correlates with downstream performance as perturbation severity increases

Comparison of Perturbation Types



- Embeddings more robust to gain and reverb than to high- and low-pass filtering
- OpenL[3] changes significantly with low-pass filtering
  - Codec-related *shortcuts*[4] in self-supervised learning

[4] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in ICCV. 2015.

Conclusion

- **Contributions**
  - Evaluate robustness of audio embeddings against channel effects in a task-free setting
  - OpenL^3 performs better than YAMNet (in line with HEAR results)
  - FAD has high inverse correlation with downstream performance
    - May be used for data augmentation
  - Embeddings more robust to changes in gain and reverberation than in high/low pass filtering
- **Limitations**
  - Still preliminary analysis
  - Distance show correlation but further work is needed for them to be actual predictors
- **Future Work**
  - Extending the analysis to more datasets/embeddings
  - Formalizing the theory

Thank You