A Study on Robustness to Perturbations for Representations of Environmental Sound

Sangeeta Srivastava[1], Ho-Hsiang Wu[2], Joao Rulff[2], Magdalena Fuentes[2],
Mark Cartwright[3], Claudio Silva[2], Anish Arora[1], Juan Pablo Bello[2]

[1]The Ohio State University, Columbus, OH, USA
[2]New York University, New York, NY, USA
[3]New Jersey Institute of Technology, New Jersey, NY, USA

***Abstract*—Audio applications involving environmental sound analysis increasingly use general-purpose audio representations, also known as embeddings, for transfer learning. Recently, Holistic Evaluation of Audio Representations (HEAR) evaluated twenty-nine embedding models on nineteen diverse tasks. However, the evaluation's effectiveness depends on the variation already captured within a given dataset. Therefore, for a given data domain, it is unclear how the representations would be affected by the variations caused by myriad microphones' range and acoustic conditions – commonly known as *channel effects*. We aim to extend HEAR to evaluate invariance to channel effects in this work. To accomplish this, we imitate channel effects by injecting perturbations to the audio signal and measure the *shift* in the new (perturbed) embeddings with three distance measures, making the evaluation domain-dependent but not task-dependent. Combined with the downstream performance, it helps us make a more informed prediction of how robust the embeddings are to the channel effects. We evaluate two embeddings – YAMNet, and OpenL^3 on monophonic (UrbanSound8K) and polyphonic (SONYC-UST) urban datasets. We show that one distance measure does not suffice in such task-independent evaluation. Although Fréchet Audio Distance (FAD) correlates with the trend of the performance drop in the downstream task most accurately, we show that we need to study FAD in conjunction with the other distances to get a clear understanding of the overall effect of the perturbation. In terms of the embedding performance, we find OpenL^3 to be more robust than YAMNet, which aligns with the HEAR evaluation.**

***Index Terms*—Self-supervised learning, robust audio embeddings, transfer learning, acoustic perturbations, urban sound**

I. Introduction

The scarcity of a large amount of labeled data for supervised learning in applications related to environmental sounds has popularized the use of representation learning and transfer learning [1]–[5] in such applications. As part of this learning paradigm, a network is pre-trained on an *upstream* task, which has the availability of large datasets to learn generic representations or embeddings that are transferable across a variety of related target *downstream* application(s). Bengio et al. [6] defines good representations as one that are *expressive* enough to capture a considerable number of possible input configurations, are *invariant* to most local changes of the input, and *disentangles* the factors of variation in the input.

With the increase in the number of learning frameworks and architectures to learn such invariant representations, there is a need for evaluation benchmarks to test the generalization of the embedding models and empirically compare them. Holistic Audio Representation Evaluation Suite (HARES) [7] and Holistic Evaluation of Audio Representations (HEAR) [8] are two efforts at this front that test the invariance of various audio representations to downstream domains and tasks. The HEAR challenge, in particular, is the most extensive effort to date and includes an evaluation of twenty-nine audio embedding models on nineteen diverse downstream tasks. However, both HEAR and HARES evaluation methodologies have several limitations. Firstly, they are dependent on the included tasks, as well as the quantity and distribution of the training and test sets of those tasks. Thus, they are not informative as to how the embeddings will perform on unseen tasks. Secondly, they do not provide information on the stability of the audio embeddings in response to specific changes in the same data domain. Hence, not only might they give a limited understanding of what to expect when employing them in real-world applications under various conditions, but they also require inspection and analysis of their test sets to gain understanding of their stability. Lastly, they rely on the availability of annotated data for evaluation. This has the inherent drawback of requiring human annotations, especially for data collected from real-world deployments [9], [10].

In this work, we propose a path to address these limitations by using the following two steps in the evaluation framework: **(i)** we propose an alternative, yet complementary, testing scenario that includes *invariance to channel effects*. To accomplish this, we artificially degrade audio signals [11]–[13] by applying different mathematical transformations or *perturbations*, and **(ii)** we leverage distance metrics that quantify the *shift* in the embedding space directly, making the evaluation independent of the task but still dependent on the data domain. We correlate the metrics with the downstream results to corroborate the findings and establish the relationship between the perturbations and the downstream evaluation. We leverage two publicly available audio embedding models, *OpenL*^{3} [1] and *YAMNet*[a], to build on the findings from the

This work is partially supported by the NSF award 1544753.

[a]https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

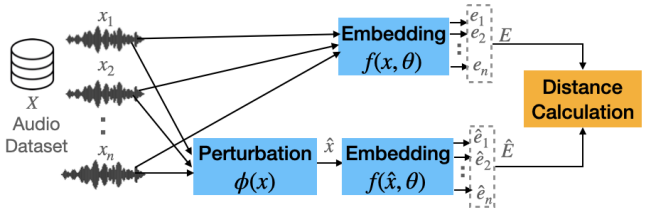


Fig. 1: Pipeline to evaluate the robustness of embeddings by calculating distance between the original and the new audio embeddings, E and $\hat{E}$, respectively.

HEAR challenge, where OpenL^3 and YAMNet are among the best and worst-performing models, respectively.

Our contributions can be summarized as follows:

1) We propose a methodology to evaluate the robustness of audio embeddings against channel effects, both qualitatively and quantitatively, in a task-free setting.
2) We investigate the effectiveness and limitations of correlating three distance measures quantifying change or shift in pairwise distances, topology, and distribution with the downstream performance.
3) Mimicking channel effects with four fundamental perturbations: high pass (HP) and low pass (LP) filtering, gain, and reverberation, we show that embeddings are more robust to changes in gain and reverberation than in HP and LP.
4) HEAR shows OpenL^3 to perform better than YAMNet. We get a similar conclusion, but with a closer inspection of the performance differences in each of the four perturbations.

II. Robustness to Perturbation

Let $X = \{x_i\}_{i=1}^n$ be a dataset of n audio snippets and θ be the parameters of the upstream embedding model $f(x,\theta) \rightarrow e_x$ that maps audio input x to a d-dimensional embedding $e_x \in \mathbb{R}^d$. $E = \{e_i\}_{i=1}^n$ is the set of all such n embeddings. Also, consider a transformation function $\phi(x) \rightarrow \hat{x}$ that perturbs the audio signal x to $\hat{x}$. The new audio set $\hat{X}$ then produces a new embedding space $\hat{E} = \{\hat{e}_i\}_{i=1}^n$. The robustness problem is then stated as follows: the embedding space $\hat{E}$ produced by the upstream model on the perturbed audio set should not change the semantics of the audio signal i.e. distance between E and $\hat{E}$ is small. We list different distance metrics to measure the variation between the two embedding spaces in section III-A.

We investigate four perturbations, namely *high pass* and *low pass* filtering, *gain* and *reverberation*. These perturbations are inspired by channel effects that arise when deploying environmental audio sensing devices, and simulate varying conditions both in the acoustic propagation from the source to the recording device and in the recording device itself. However, these perturbations are common in many microphone recording situations. Table I lists the range of values on which we explore each perturbation.

High and Low Pass filtering: Since low-cost microphones may not have a full frequency range response, we use high-pass and low-pass filters to approximate various frequency responses to test the representations' ability to be mic-invariant. While several applications focused on urban sound monitoring [14] use MEMS mics with a frequency range of 20-20k Hz, the sampling frequency of 44.1 or 48 kHz is an expensive option for low-power micro-controllers. Recently, Lopez et al. [15] leverage mics with a frequency range of 63-8k Hz instead. Besides the inherent differences in a mic's design, external factors like water, wind, and dust can change the frequency response. For example, water clogged inside the mic windscreen can attenuate the signal, especially at higher frequencies [16] and low pass filters can also simulate this.

Reverberation: Typically, the sources of environmental sounds are outdoors, for example, construction noise, honking, and aircraft, to name a few. However, people hearing these sounds can be in outdoor areas like streets with many buildings or indoor areas with walls and furniture. Sound waves reflect from such obstacles several times before reaching the ear. The sound reflections mix to create what is known as reverberation. We evaluate the representations for different listener environments by modeling the reverberation time of space, defined as the time required for the sound level to decay by 60 dB after the signal has stopped.

Gain: Due to infrastructure requirements, it is common to place the microphones far from the sound sources when collecting environmental sounds. For a spherical wave, the sound pressure level (SPL) decreases by 6 dB (one-half) per doubling of distance from the source. For line sources such as traffic noise, the decay rate varies between 3 and 4 dB [17]. In order to test the near-field performance of the learned representations, we vary the gain of the signal.

TABLE I: Range of values for each perturbation (pert.) type for OpenL^3 and YAMNet

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

III. Experimental Design

Fig. 1 shows the pipeline to calculate a distance measure to quantify the effect of a perturbation ϕ on an embedding space E.

A. Distance Metrics to Evaluate Robustness

We utilize a toy dataset of five random coordinates in Fig. 2 to motivate the usage of different distance measures. The dataset is intended to be simple and illustrative rather than directly related to the perturbations in this paper. Even minor perturbations, such as those in $1a$ and $1b$, change the pairwise distances between the old and new points. However, the distance between them remains preserved in the new space, which is evident from the hierarchical clustering in $2a$ and $2b$. Similarly, scaling by a constant factor of 2 clusters the new points the same as that in the original dataset but changes

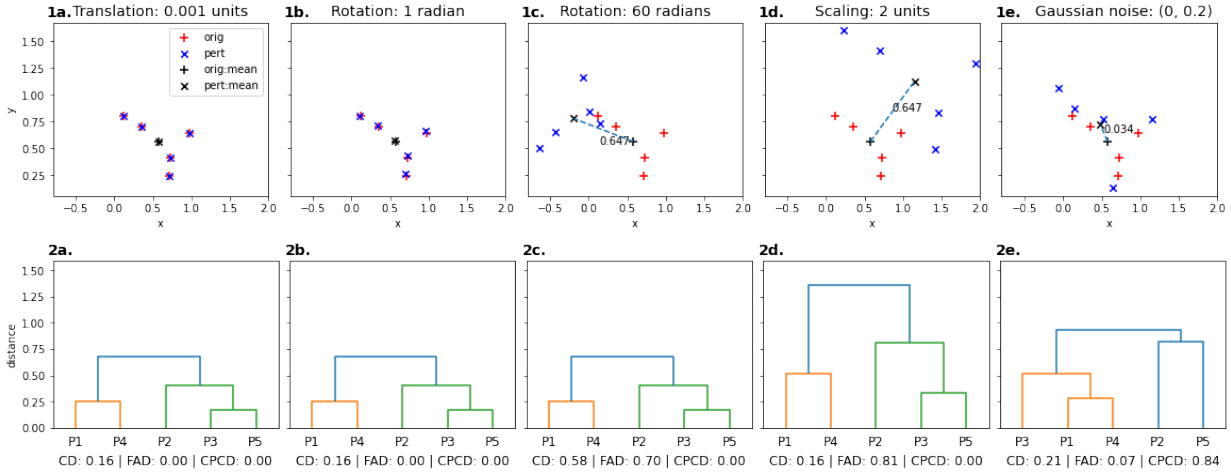


Fig. 2: Cosine Distance (CD), Fréchet Audio Distance (FAD) and Cophenetic Correlation Distance (CPCD) to measure change in pairwise distance, distribution and topology respectively for a toy dataset of five coordinates. The text on the dotted blue line denotes the distance between the original (orig) and the perturbed (pert) mean.

the mean of the new distribution. Motivated by the unique qualitative and quantitative information provided by various metrics, we investigate distance measures to evaluate the shift between the original and the perturbed embeddings in three aspects: (1) pairwise distances, (2) relative pairwise distances (as in hierarchical clustering topology), and (3) distribution.

Pairwise: When comparing embeddings, a common method has been to use some pairwise distance. We choose *cosine similarity* because it is common to normalize embeddings before training the downstream classifier. To change similarity into distance, we use *cosine distance* (CD). To generate a single distance value for the full embedding set, we find the mean of all the CDs.

Topology: As shown in Fig. 2 $2a$-$2d$, there may be scenarios in which pairwise distances might significantly change, even when the relative distances between the data points do not vary as observed in clustering. In such situations, classes may still be well-separated in the embedding spaces, but new data may be required to represent those class distributions. To make the pairwise study less stringent and distance-free, we evaluate the total change in the pairwise proximity of the embeddings in E and $\hat{E}$. We use agglomerative clustering with Euclidean distance and average linkage criterion to create dendrograms for the original and perturbed embeddings. The branching patterns (also known as topology) in the two dendrograms might differ in terms of the embedding positions in the leaf set. To quantify the difference, we calculate the *Pearson correlation coefficient (PCC)* (Equation 1) between the cophenetic distance matrices, C_o and C_p, for the dendrograms corresponding to the original and perturbed embeddings. We utilize Equation 2 to convert the correlation into a distance metric, which we refer to as *cophenetic correlation distance (CPCD)*.

$$PCC = \frac{cov(C_o, C_p)}{\sqrt{var(C_o)var(C_p)}} \quad (1)$$

$$CPCD = 1 - PCC \quad (2)$$

where cov and var correspond to covariance and variance, respectively.

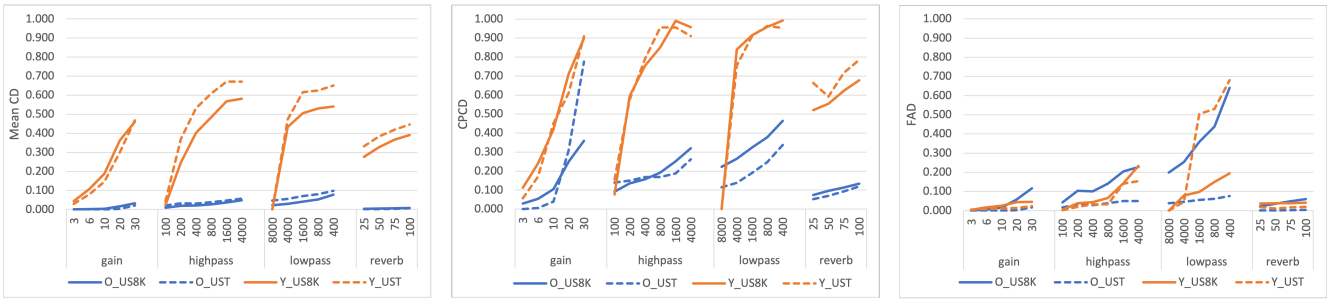
Distribution: In order to get the variation in the distribution within the embedding space, we leverage the *Fréchet Audio Distance (FAD)*. Initially proposed for music enhancement application, Kilgour et al. [18] use FAD to compare the embedding statistics generated on a large reference set of clean music with the embedding statistics generated on an evaluation set of enhanced noisy signals. In this work, we use FAD to compare the statistics between the original and the perturbed embedding set. The Fréchet distance (also known as Wasserstein-2 distance) between the Gaussian of the original embeddings $\mathcal{N}_o(\mu_o, \Sigma_o)$ and the perturbed embeddings $\mathcal{N}_p(\mu_p, \Sigma_p)$ is then computed as:

$$FAD(\mathcal{N}_o, \mathcal{N}_p) = ||\mu_o - \mu_p||_2^2 + tr(\Sigma_o + \Sigma_p - 2\sqrt{\Sigma_o \Sigma_p}) \quad (3)$$

where μ represent the mean, Σ the covariance matrix, and tr the trace. Unlike the cosine and the correlation distance, FAD is oblivious to the way the embeddings are related to one other, as illustrated in $1c$ and $1d$ in Fig. 2. It is primarily used to investigate the change in the overall embedding distribution.

B. Datasets

We study the robustness of both *OpenL*^{3} and *YAMNet* for two popular datasets, namely *UrbanSound8K (US8K)* [19], and *SONYC Urban Sound Tagging (UST)* [20]. These datasets complement those used in the HEAR challenge for environmental sound. For the UST dataset, we use all the 1380 recordings with verified annotations in v2.3. As for the US8K samples, we use all the ∼8k samples. To simplify the analysis, we sample one embedding per clip, which we select by computing the sound pressure levels (SPL) and retrieving the embedding where the SPL is highest. The assumption behind this is that the highest SPL level correlates with the presence of a labeled sound source.



(a) (b) (c)

Fig. 3: CD, CPCD, and FAD for US8K and UST datasets for four perturbation types. The x-axis values for each perturbation range from low to high changes. The sampling frequency of YAMNet (Y) and OpenL^3 (O) is 16kHz and 44.1kHz, respectively.

C. Metrics for Downstream Evaluation

We assess the effect of perturbation not only on the embeddings but also on the classification metrics on the downstream datasets, i.e. US8K and UST. Specifically, we train a logistic regression model with the original embeddings E and evaluate its performance on embeddings $\hat{E}$ perturbed with different types and values.

US8K: We employ cross-validation accuracy as well as *mean silhouette score* to compare the quality of classification and clustering of the embeddings before and after the perturbation.

UST: As for the UST dataset, we use macro-averaged areas under the precision-recall curve (macro-AUPRC) as the primary evaluation metric. We do not use silhouette analysis for UST because it is a multi-label dataset and one embedding can be part of multiple classes at the same time.

IV. Evaluation

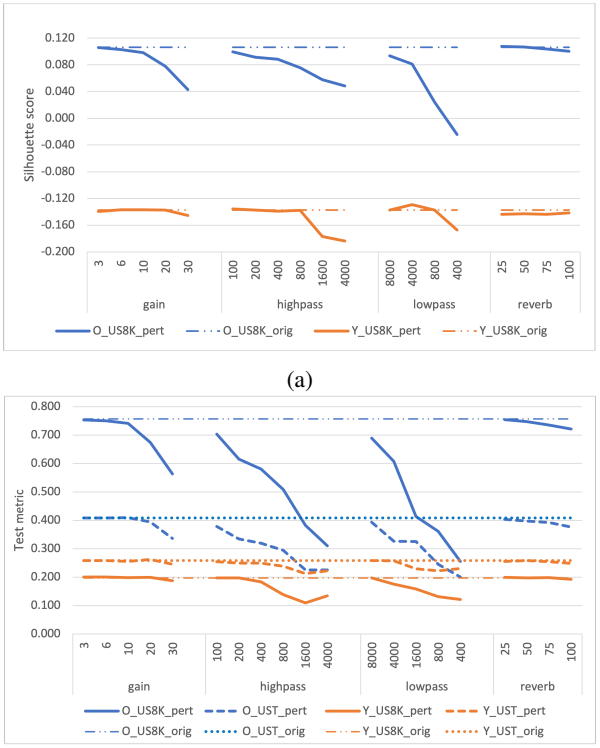
A. Comparison of representation types

Looking at Fig. 3, for both CD and CPCD, YAMNet shows a larger distance (higher sensitivity) as compared to OpenL^3. To get a deeper understanding of YAMNet's sensitivity to pairwise relations, we calculate the *silhouette* scores of the embeddings of each class in US8K (c.f. Section IV-B).

Although OpenL^3's distribution show more variation than YAMNet for US8K (c.f. Fig. 3c) when perturbed, large values of CPCD for YAMNet in Fig. 3b indicate that the YAMNet's pairwise relationships change significantly in the perturbed space, and can possibly affect the downstream performance. Fig. 4b confirms this hypothesis. Note that in order to re-scale FAD to a 0-1 scale, we use Min-Max scaling within a dataset to normalize FAD scores, which somewhat skews the comparison but has no effect on the overall trend.

B. Distance metrics and downstream evaluation

We compare the trends of the distance measures with the downstream evaluation metrics. In Fig. 4a, we observe that YAMNet produces a negative silhouette score of -0.14 even for the original representations, meaning that embeddings of the same class lack the two qualities of separability from embeddings of other classes and cluster compactness. Even



(a)

(b)

Fig. 4: (a) Silhouette scores of OpenL^3 and YAMNet for US8K. (b) compares the classification accuracy of the original (orig) and perturbed (pert) embeddings for US8K and UST.

a tiny modification can change the pairwise groups in this scenario. This is also reflected in Fig. 3a and Fig. 3b.

The trends in FAD closely approximate the performance drops in Fig. 4b as the severity of the perturbation increases. We notice that OpenL^3 has a steeper change in accuracy than YAMNet. Nevertheless, even when OpenL^3 produces the lowest accuracy (high pass at 8k Hz), it is better than YAMNet. One can infer the same by considering both FAD and CPCD simultaneously, as stated in Section IV-A. The mean CD might have a neutralizing effect. For an example, let us consider two embeddings, e_1 and e_2 in E. If the CD decreases for the

$(e_1, \hat{e}_1)$ pair by 0.3 but increases for $(e_2, \hat{e}_2)$ by 0.3, the mean change would remain unaffected. Both CD and CPCD utilize pairwise information and are comparatively more sensitive to noise and outliers. We recommend always supplementing such pairwise metrics with information from other robust metrics like FAD.

Furthermore, because FAD best reflects the performance loss, it may be used for data augmentation to make the downstream classifier more robust. When it comes to determining what values to utilize for augmentation, we can see from FAD (c.f. Fig. 3c) that each embedding and dataset combination appears to have different inflection points, i.e., where changes in distance and performance drop more dramatically. We believe that this value is a useful indicator of how much perturbation to use for the augmentation, as larger values may be associated with dramatic changes in the signal, while smaller values may not make a significant difference. In future research, we'll investigate the use of inflection points for augmentation.

C. Comparison of perturbation types

The embeddings are more robust to *gain* and *reverb* than to high and low pass filtering. This is not surprising because these perturbations do not significantly change the information contained in the signal (much less than low and high pass filtering), so the fact that the embeddings are robust to them is a good indication that the models are doing what we expect and they are mainly learning semantic information. The inflection point for FAD and CPCD at a gain of 10 dB indicates the presence of harmonic distortions associated with clipping. It is a bit surprising how much OpenL^3 embeddings change in response to low pass perturbations. We hypothesize this is due to a codec-related "shortcut" [21] in the self-supervised audio-visual correspondence task in which the model finds a relationship between high-frequency absence and image artifacts in low-bit-rate encodings.

V. Conclusion

We employ three distance metrics to estimate the effect of channel effects on two representations, OpenL^3 and YAMNet. We demonstrate that the downstream performance and the distance measures are complementary. Limiting the evaluation to downstream performance precludes a more in-depth study of the reason and extrapolation of the findings to other real-world test scenarios. Similarly, the analysis of distance measurements can be misleading when using only one metric. In our study, the combination of FAD and CPCD gave the most valuable insight and was representative of downstream trends. We recommend using FAD to choose among different perturbations for augmentation to make sound event detection models more robust.

In future work, we intend to repeat this study on a wide variety of embeddings and datasets and extend the analysis to include correlations between distance metrics and different sound event classes.

References

[1] J. Cramer, H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[2] A. Jansen, D. P. Ellis *et al.*, "Coincidence, categorization, and consolidation: Learning to recognize sounds with minimal supervision," in *ICASSP*. IEEE, 2020, pp. 121–125.

[3] L. Wang and A. v. d. Oord, "Multi-format contrastive learning of audio representations," *arXiv preprint arXiv:2103.06508*, 2021.

[4] S. Srivastava, Y. Wang, A. Tjandra, A. Kumar, C. Liu, K. Singh, and Y. Saraf, "Conformer-based self-supervised learning for non-speech audio tasks," *arXiv preprint arXiv:2110.07313*, 2021.

[5] J.-B. Alayrac, A. Recasens *et al.*, "Self-supervised multimodal versatile networks," *NeurIPS*, vol. 2, no. 6, p. 7, 2020.

[6] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[7] L. Wang, P. Luc, Y. Wu, A. Recasens, L. Smaira, A. Brock, A. Jaegle, J.-B. Alayrac, S. Dieleman, J. Carreira *et al.*, "Towards learning universal audio representations," *arXiv preprint arXiv:2111.12124*, 2021.

[8] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally *et al.*, "Hear 2021: Holistic evaluation of audio representations," *arXiv preprint arXiv:2203.03022*, 2022.

[9] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[10] C. E. Catlett, P. H. Beckman, R. Sankaran, and K. K. Galvin, "Array of things: a scientific research instrument in the public way: platform design and early lessons learned," in *Proceedings of the 2nd international workshop on science of smart city operations and platforms engineering*, 2017, pp. 26–33.

[11] D. de Benito-Gorrón, D. Ramos, and D. T. Toledano, "An analysis of sound event detection under acoustic degradation using multi-resolution systems," *Proc. IberSPEECH*, vol. 2021, pp. 36–40, 2021.

[12] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 86–90.

[13] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[14] J. Ardouin, L. Charpentier, M. Lagrange, F. Gontier, N. Fortin, D. Ecotière, J. Picaut, and C. Mietlicky, "An innovative low cost sensor for urban sound monitoring," in *INTER-noise and noise-con congress and conference proceedings*, vol. 258, no. 5. Institute of Noise Control Engineering, 2018, pp. 2226–2237.

[15] J. M. López, J. Alonso, C. Asensio, I. Pavón, L. Gascó, and G. de Arcas, "A digital signal processor based acoustic sensor for outdoor noise monitoring in smart cities," *Sensors*, vol. 20, no. 3, p. 605, 2020.

[16] C. Ribeiro, D. Ecotière, P. Cellard, and C. Rosin, "Uncertainties of the frequency response of wet microphone windscreens," *Applied acoustics*, vol. 78, pp. 11–18, 2014.

[17] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of acoustics*. John wiley & sons, 2000.

[18] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms." in *INTERSPEECH*, 2019, pp. 2350–2354.

[19] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[20] M. Cartwright, J. Cramer, A. E. M. Mendez, Y. Wang, H.-H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon *et al.*, "SONYC-UST-V2: An urban sound tagging dataset with spatiotemporal context," *arXiv preprint arXiv:2009.05188*, 2020.

[21] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.