AURA: PRIVACY-PRESERVING AUGMENTATION TO IMPROVE TEST SET DIVERSITY IN NOISE SUPPRESSION APPLICATIONS

Xavier Gitiaux ∗1, Aditya Khant2, Chandan Reddy †3, Jayant Gupchup2, Ross Cutler2

1. George Mason University, Fairfax VA, USA. Email: xgitiaux@gmu.edu
2. Microsoft Corporation, Redmond WA, USA. Email: firstname.lastname@microsoft.com
3. Email: chandan.ka@outlook.com

ABSTRACT
Noise suppression models running in production environments are commonly trained on publicly available datasets. However, this approach leads to regressions in production environments due to the lack of training/testing on representative customer data. Moreover, due to privacy reasons, developers cannot listen to customer content. This ‘ears-off’ situation motivates augmenting existing datasets in a privacy-preserving manner. In this paper, we present Aura, a solution to make existing noise suppression test sets more challenging and diverse while limiting the sampling budget. Aura is ‘ears-off’ because it relies on a feature extractor and a metric of speech quality, DNSMOS P.835, both pre-trained on data obtained from public sources. As an application of Aura, we augment a current benchmark test set in noise suppression by sampling audio files from a new batch of data of 20K clean speech clips from Librivox mixed with noise clips obtained from AudioSet. Aura makes the existing benchmark test set harder by 100% in DNSMOS P.835, a 26% improvement in Spearman’s rank correlation coefficient (SRCC) compared to random sampling and, identifies 73% out-of-distribution samples to augment the test set.

Index Terms— Machine Learning, Noise Suppression, Audio Datasets, Data Privacy

1. INTRODUCTION
Speech enhancement includes removing background noise and reverberation [1, 2]. Deep learning based approaches provide state-of-the-art improvement but typically involve supervised training on a large and diverse dataset of noisy speech and clean speech pairs [3]. Since real-world applications contain speech mixed with noise, most deep learning methods rely on synthetic data that mix clean and noisy speech [4]. However, synthetic data cannot fully capture real-world conditions leading to poor performance when deployed to large, diverse customer bases [5]. One solution is to identify new conditions to augment current test sets with additional representative data. However, labeling a large batch of customer data is expensive, and often prohibitive due to privacy and compliance reasons.

This paper presents Aura, an ‘ears-off’ methodology to identify scenarios in a sample-efficient manner that is: (i) representative of the customer workload, and (ii) challenging to noise suppression models. Real-world scenarios are instrumental to generate a low-variance estimate of how noise suppressors improve speech quality when used by customers. It is noteworthy that keeping the test set small (i.e., sample efficiency) ensures agility of development and deployment in production pipelines. Production workload datasets can be in the order of millions of clips so sample efficiency is an important practical challenge.

The main contribution of Aura is to operate in an ‘ears-off’ mode by relying on a pre-trained audio feature extractor [6] and a pre-trained accurate objective speech quality metric called DNSMOS P.835 [7]. Both the feature extractor and speech quality metric are convolution-based models trained on large-scale open-sourced processed data and do not risk memorizing customer private information. Aura maps customer data into an embedding space, partitions the data via clustering, and identifies the most challenging scenarios for noise suppressors in each cluster.

Since we cannot share results on customer data, we evaluate Aura’s sampling on simulated data. We apply Aura to rank 28 noise suppression models from the DNS Challenge [4] in terms of overall mean opinion scores (MOS) [7]. We show that the Spearman’s rank correlation coefficient (SRCC) between a 1% sample produced by Aura and the entire mixture of noisy speech is 0.91. This is an improvement of 26% compared to a 1% random sampling of noisy speech. We also apply Aura to augment a current benchmark test set for noise suppressors [4]. The resulting test set is more challenging to noise suppressors with a predicted absolute decline of 0.36 (a relative decline of 100%) in DNSMOS P.835 [7]. Moreover, the Aura-based test set captures more diverse audio scenarios with an increase of 43% in diversity (as measured by $\chi^2$—distance, and an addition of 73% out-of-distribution (OOD))
samples compared to the current benchmark test set [4].

**Related work.** As part of the Deep Noise Suppression (DNS) challenge [4], Reddy et al. created an extensive open-sourced noisy speech test set with about 20 noise types, 15 languages, 200+ speakers, emotions, and singing. However, the resulting DNS benchmark test set does not cover all the scenarios that customers who use real-time communication platforms experience. These gaps motivate additional effort to identify challenging scenarios where DNS models underperform.

Active learning is the common paradigm to select which data points to label from a pool of unlabeled data [8] when the labeling budget is limited. Many active learning heuristics are uncertainty-based and try to sample in priority hard examples [9]. Kossen et al. show that heuristics in active learning lead to biased estimates of model performances because they do not preclude selecting redundant/outlier scenarios [10]. On the other hand, Sener et al. propose a diversity-based solution identifying a set of data points that are the most representative of the entire data set [11]. This paper integrates diversity and uncertainty-based techniques to a pipeline that creates a test set for noise suppression models while operating in an ‘ears-off’ environment.

Our ‘ears-off’ pipeline protects customers from privacy leakages, including membership attacks [12] and the extraction of identifiable information [13]. A growing body of work has developed differentially private learning algorithms that do not expose private customer data [14]. Federated learning is a partial solution to the ‘ears-off’ problem since it allows training models on decentralized devices without the need for a central server to access private data. Latif et al. and Guliani et al. present applications of the federated learning paradigm to speech-related tasks [15][16]. However, these methods rely on an availability of labels at client endpoints.

We obtain an ‘ears-off’ pipeline by transferring knowledge from models trained on open-source data. Self-supervision uses surrogate tasks to learn representations or embeddings that are useful to real-world downstream tasks. The method has been successful in image-based tasks (e.g., [17]), and audio/video (e.g., [18]). In this paper, our pretext task is to label the type of background noise in noisy speech. Moreover, we use DNSMOS P.835 [7] to accurately measure the quality of speech without the need for clean reference. DNSMOS P.835 has a 0.98 correlation with human ratings for noise suppression models and a 0.74 correlation per clip. This gives the flexibility to estimate speech quality before and after noise suppression without listening to the file.

### 2. SOLUTION

Our objective is to estimate a performance metric ρ of noise suppression on a target data set that represents real-world conditions. For each speech clip, the P.835 protocol generates a MOS for signal quality SIG, background noise BAK, and overall quality OVRL. We derive two performance metrics ρ from MOS: (i) differential MOS (DMOS) between after and before denoising; (ii) stack ranking of competing noise suppressors according to their average DMOS calculated on the target data $D_t$.

The constraints are that only a small subset $S$ of files can be sampled out of all the noisy speech in $D_t$; and, that audio files in $D_t$ cannot be used to fit a model that could potentially encode identifiable information in its parameters. The restriction $|S| \ll |D_t|$ limits the size of the test set and allows rapid testing of models during development. We want the sampling estimate of the performance metric $\rho$ to have a small error $\varepsilon$ compared to the expected value of $\rho$ on the target data $D_t$. To minimize $\varepsilon$, we trade-off bias and variance [19]. A random sample of $D_t$ audio files would result in zero bias, but high variance. On the other hand, probability-proportional-to-size sampling (PPS) would minimize the variance of the estimator by sampling audio files with a probability proportional to $\rho$ [20]. One shortcoming of PPS sampling performed solely on $\rho$ is that it does not consider the diversity of scenarios.

To trade-off variance and bias, we propose Aura, the pipeline shown in Figure 1. Aura learns a partition of the target data into $k$ clusters in an embedding space obtained from a pre-trained feature extractor. Then, it applies PPS to each cluster. Aura samples audio files in two steps. First, it samples files in all the clusters of the embedding space to capture the diversity represented in the target data. Second, within each cluster, it favors audio files with the largest $\rho$ to reduce the variance of the performance metric $\rho$.

**Test set for noise suppression models.** The first application of Aura is to sample $|S|$ audio files in the target data $D_t$ to form a test set for noise suppression models. In each of the $k$ clusters of the embedding space, we sample $|S|/k$ files with a probability inversely proportional to their predicted DMOS. Note that a smaller DMOS indicates a more challenging (or noisier) speech scenario.

**Model ranking.** A model evaluation task in noise suppression involves comparing the speech quality produced by different models on the same set of audio files. Given $N$ noise suppressors, we adjust our sampling procedure in the follow-
ing manner: Within each cluster, an audio file is selected with a probability proportional to the variance across models of the predicted DMOS for that file. In this application, the sampling performance $\epsilon$ is defined as the SRCC between the ranking of $N$ models obtained on the sample $S$ and the one obtained with the entire target data $D_t$.

3. EXPERIMENTS

3.1. Pipeline Components

Choice of feature extractor. Our feature extractor is constructed from VGGish, a model pre-trained on a 100M YouTube videos dataset [6]. It generates a 128 dimension embedding for each audio clip. VGGish embeddings are trained to identifying the type of foreground sound in an audio clip. In the context of noise suppression, we aim to detect the noise in the background of a speech clip. Therefore, we fine-tune the feature extractor by using its embedding as inputs to two fully connected layers and by classifying the type of background noise.

We follow Reddy et al. [4] to synthesize 1.5 million instances of 10 second noisy speech containing 527 categories of noise sounds from AudioSet [21]. We add each noise sound to the background of speech clips randomly drawn from the bank of clean speech clips created by [4]. Our background noise type classifier trained on this synthetic data achieved a mean average precision (MAP) of 0.33 and area-under-the-curve (AUC) of 0.96 on a hold-out sample. It provides us with a feature extractor that maps noisy speech to a 128 dimension embedding space that is semantically aligned with the type of background noise. Using kmeans++ on the embedding space, we partition 1.5 million noisy speech into 256 clusters [23]. We validated the quality of the clusters using subjective listening tests. The raters were presented clips from 24 random clusters and asked to report whether audio files in the cluster share a common audio property. We found that 80% of the clusters share a common background noise.

Choice of MOS predictor. We use a pre-trained model, DNSMOS P.835 to predict subjective speech quality [7].

3.2. Dataset

We will apply Aura to two target test sets.

Augmented DNS challenge test set. We create a target dataset of potential noisy speech candidates to add to the INTERSPEECH 2021 DNS Challenge test set [4]. We mix sounds from the balanced AudioSet data [21] with clean speech clips [23]. We use segments from AudioSet as background noise and follow [23] to generate noisy speech with signal-to-noise ratio (SNR) between -5 dB and 5 dB. The resulting target data combines 1.7K files from the DNS challenge test set with 22K of 10 seconds clips from the newly synthesized noisy speech that covers 527 noise types with at least 59 clips per class. The pool of noisy speech candidates is at least partially out-of-distribution compared to [4], which covers 120 noise types. Moreover, it does not overlap the dataset of audio clips used to fine-tune our feature extractor and train DNSMOS P.835 and thus, reproduces the conditions of an “ears-off” environment.

Augmented DNS challenge test set + clean speech. To simulate real-world conditions, for each noisy speech in the previous target data, we randomly draw 10 clean speech clips from [4]. Clean speech presents a challenge to stack rank models in development because it does not allow to discriminate the performance of noise suppressors.

3.3. Augmenting DNS challenge test set

We evaluate the following: From the pool of noisy speech candidates, can Aura generate a new test set of the same size (1.7K) as the INTERSPEECH 2021 benchmark test set [4], but with audio scenarios that are more challenging and diverse?

Protocol. We measure the diversity of a test set as the $\chi^2$ distance between the distribution of audio segments across the embedding clusters and a uniform distribution between clusters. We normalize the value by calculating $\chi^2$ of the contingency table over the percentage of data points in each cluster rather than raw frequency. The lower the $\chi^2$ distance, the more audio properties encoded in the embedding space the resulting test set covers.

Results. Table 1 compares DMOS and $\chi^2$ for the benchmark test set (top row) and the Aura-generated test set (bottom row). The proposed test set replaces 73% of noisy speech in the benchmark test set with a new noisy speech from our synthetic data. Aura forms a test set with clips for which noise suppressors are more likely to degrade the signal and overall

| Test set                        | SIG | BAK | OVRL | $\chi^2$ | OOD |
|---------------------------------|-----|-----|------|----------|-----|
| DNS Challenge                   | 0.24| 0.99| 0.18 | 498      | 0%  |
| (no cluster)                    | ±0.01|±0.01|±0.01 | (< 0.001)|     |
| Proposed                        | -0.57|0.91|-0.32 | 3843     | 69% |
| (no cluster)                    | ±0.01|±0.01|±0.001| (< 0.001)|    |
| Proposed + clustering           | -0.47|0.89|--0.18| 287      | 73% |
| (no cluster)                    | ±0.01|±0.01|±0.01 | (≥ 0.2)  |     |

Table 1. Aura-based DNS test set. Sample provided by Aura on the new noisy data set increases difficulty level (lower DMOS) and diversity (lower $\chi^2$). The evaluation set is compared to the INTERSPEECH test set [4], ± represents 95% confidence interval and P-values are in parenthesis. Low p-values indicate sample does not cover the clusters formed by the target data; large p-values mean almost uniform coverage.
Table 2. SRCC between the ranking of 28 NS models obtained from a 1% Aura sample with the model ranking obtained from using the entire data. Diversity is measured using $\chi^2$ distance (lower is better). 95% conf interval depicted using ±.

| Sampling method | SIG SRCC (±0.02) | BAK SRCC (±0.01) | OVRL SRCC (±0.02) | $\chi^2$ ($\times10^3$) |
|-----------------|------------------|------------------|-------------------|------------------------|
| Random          | 0.58 ± 0.02      | 0.72 ± 0.01      | 0.72 ± 0.02       | 9 ± 0.01               |
| Diversity       | 0.72 ± 0.02      | 0.89 ± 0.01      | 0.82 ± 0.01       | 4 ± 0.02               |
| Variance        | 0.80 ± 0.01      | 0.93 ± 0.01      | 0.88 ± 0.01       | 7 ± 0.02               |
| Aura            | 0.84 ± 0.01      | 0.93 ± 0.01      | 0.91 ± 0.01       | 1 ± 0.01               |

Fig. 2. Top-10 of out-of-distribution (green) and in-distribution (blue) noise categories in Aura-based samples. In-distribution relates to categories in INTERSPEECH 2021 test set.

Fig. 3. SRCC of Aura-based rank estimates as a function of sample size.

3.4. Model ranking

We evaluate Aura’s sampling efficiency to accurately stack rank noise suppression models with few samples $S$ from the target data $D_t$. The objective is to estimate a model ranking that has a high SRCC with the ranking that would be obtained on the entire data. Protocol. For each speech, we run 28 noise suppression models from [4] and on the DMOS predicted by our DNSMOS P.835 model as opposed to human ratings. We bootstrap the sampling 200 times and report the mean and standard deviation of the resulting rank correlation coefficients. We compare Aura’s sampling performance to three alternative strategies: (i) Random, which draws randomly 1% of data; (ii) Diversity, which samples uniformly across embedding clusters; (iii) Variance, which samples proportionally to the variance of predicted DMOS across the 28 models.

Results. Table 2 shows the SRCC for ranking based on signal, background and overall DMOS. Aura’s sampling leads to a 26% improvement over random sampling for signal-based ranking. We also observe that the 95% confidence interval of the ranking obtained from Aura-based samples is narrower than the one obtained by random sampling. Compared to alternative approaches, Aura generates the sample with the lowest $\chi^2$ distance, which indicates a better coverage of audio scenarios. On the other hand, Random has the highest $\chi^2$ distance because it mostly samples clean speech. Figure 2 shows that even for larger samples, Aura still outperforms Random in terms of rank correlation.

4. CONCLUSION

Aura presents an end-to-end system for improving the test set used to evaluate deep noise suppression models. Aura is designed to work for ‘ears-off’ customer workloads to preserve privacy. Aura detects out-of-distribution samples by relying on an objective quality metric and a pre-trained feature extractor. Although targeted for noise suppression, the method is generic and can be extended to other media (i.e., audio/video/image) given the successful development of a powerful pre-trained feature extractor. This work is an important step toward balancing customer privacy and measuring model’s performance in real-world scenarios.
5. REFERENCES

[1] Yariv Ephraim and David Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE transactions on acoustics, speech, and signal processing, vol. 33, no. 2, pp. 443–445, 1985.

[2] Thomas Lotter and Peter Vary, “Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model,” EURASIP Journal on Advances in Signal Processing, vol. 2005, no. 7, pp. 1–17, 2005.

[3] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2014.

[4] Chandan KA Reddy, Harishchandra Dubey, Kazuhide Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Achter, and Sridam Srinivasan, “Interspeech 2021 Deep Noise Suppression Challenge,” Interspeech, 2021.

[5] Ziyi Xu, Maximilian Strake, and Tim Fingscheidt, “Deep noise suppression with non-intrusive pesqnet supervision enabling the use of real training data,” arXiv preprint arXiv:2103.17088, 2021.

[6] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., “Cnn architectures for large-scale audio classification,” in 2017 IEEE ICASSP. IEEE, 2017, pp. 131–135.

[7] Chandan Reddy, Vishak Gopal, and Ross Cutler, “DNS-MOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” arXiv preprint arXiv:2101.11665, 2021.

[8] Burr Settles, “Active learning literature survey,” 2009.

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani, “Deep bayesian active learning with image data,” in PMLR, 2017, pp. 1183–1192.

[10] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth, “Active testing: Sample-efficient model evaluation,” arXiv preprint arXiv:2103.05331, 2021.

[11] Ozan Sener and Silvio Savarese, “Active learning for convolutional neural networks: A core-set approach,” in ICLR, 2018.

[12] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, “Membership inference attacks against machine learning models,” in 2017 IEEE Symposium on Security and Privacy (SP), 2017, pp. 3–18.

[13] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al., “Extracting training data from large language models,” arXiv preprint arXiv:2012.07805, 2020.

[14] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, “Deep learning with differential privacy,” in Proceedings of the 2016 ACM SIGSAC, 2016, pp. 308–318.

[15] Siddique Latif, Sara Khalifa, Rajib Rana, and Raja Jur-Jur, “Federated learning for speech emotion recognition applications,” in 2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN). IEEE, 2020, pp. 341–342.

[16] Dhruv Guliani, Françoise Beaufays, and Giovanni Motta, “Training speech recognition models with federated learning: A quality/cost framework,” in ICASSP 2021. IEEE, 2021, pp. 3080–3084.

[17] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer, “Revisiting self-supervised visual representation learning,” in Proceedings of the IEEE/CVF, 2019, pp. 1920–1929.

[18] Sascha Grollmisch, Estefanía Cano, Christian Kehling, and Michael Taenzer, “Analyzing the potential of pre-trained embeddings for audio classification tasks,” in EUSIPCO 2020. IEEE, 2021, pp. 790–794.

[19] Corinna Cortes, Yishay Mansour, and Mehryar Mohri, “Learning bounds for importance weighting,” in Nips. Citeseer, 2010, vol. 10, pp. 442–450.

[20] Henrik Imberg, Johan Jonasson, and Marina Axelsen-Fisk, “Optimal sampling in unbiased active learning,” in AISTATS. PMLR, 2020, pp. 559–569.

[21] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in Proc. IEEE ICASSP 2017, New Orleans, LA, 2017.

[22] David Arthur and Sergei Vassilvitskii, “k-means++: The advantages of careful seeding.” Tech. Rep., Stanford, 2006.

[23] Chandan KA Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sridam Srinivasan, “ICASSP 2021 Deep Noise Suppression Challenge,” in ICASSP 2021. IEEE, 2021, pp. 6623–6627.