Review of Paper: Improving Audio Classification with Low-Sampled Microphone Input

Shreyansh Pathak , D24CSA006

April 12, 2025

1 Title of the paper

"Improving Audio Classification with Low-Sampled Microphone Input: An Empirical Study Using Model Self-Distillation"

2 Summary of the paper

This paper presents methods for optimizing pre-trained audio neural networks (PANNs) for low-quality audio inputs with sampling rates between 1-2 kHz, addressing privacy concerns and power consumption issues in mobile health and wearable applications. The authors introduce two self-distillation approaches: Born-Again Self-Distillation (BASD) and Cross-Sampling-Rate Self-Distillation (CSSD), comparing their effectiveness across three model architectures (CNN14, ResNet38, and MobileNetV2) and three datasets (ESC-50, TAU Urban Acoustic Scenes 2019 Mobile, and a proprietary user dataset). Results show both approaches improve inference performance with low-sampled audio by 1-6% compared to the baseline, with CSSD consistently outperforming BASD. The study also quantifies the privacy and power consumption benefits of using lower sampling rates.

3 Main Architecture

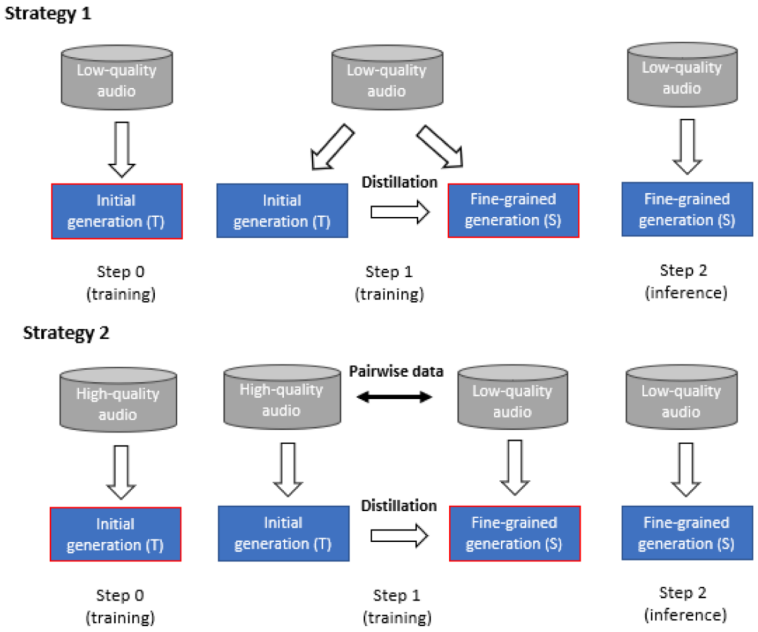

Figure 1: Pipelines of Born-Again self-distillation (BASD, strategy 1) and cross-sampling-rate self-distillation (CSSD, strategy 2). The terms "initial" and "fine-grained" generations denote different stages of the same model. The generations highlighted in red are the ones being trained per step.

4 Strengths of the paper

1. **Novel Application of Self-Distillation:** The paper introduces a cross-sampling-rate self-distillation approach (CSSD) specifically designed for audio classification with very low sampling rates, which has not been extensively explored in prior work.

2. **Comprehensive Evaluation:** The study tests the proposed methods across multiple pre-trained architectures (CNN14, ResNet38, MobileNetV2) and diverse datasets, providing robust evidence of effectiveness.

3. **Practical Benefits Quantification:** The paper quantifies both power consumption reduction and privacy enhancement through reduced audio intelligibility, directly addressing real-world concerns for mobile applications.

4. **Efficient Implementation:** The proposed methods improve model performance without requiring additional teacher architectures or model complexity, making them practical for deployment on resource-constrained devices.

5. **Consistent Performance Gains:** Both self-distillation approaches consistently outperform baseline fine-tuning across all tested models and sampling rates, demonstrating the robustness of the proposed techniques.

5 Weaknesses of the paper

1. **Limited Sampling Rate Range:** The study focuses only on 1-2 kHz sampling rates, without investigating the optimal sampling rate or exploring the performance gradient across a wider range of sampling rates.

2. **Lack of Statistical Significance Analysis:** While performance improvements are reported, the paper does not include statistical significance tests to confirm the robustness of the observed differences.

3. **Incomplete Explanation of Student Generation Effects:** The ablation study reveals that adding more student generations can decrease performance, but the paper does not fully investigate or explain this counterintuitive result.

4. **Limited Theoretical Foundation:** The paper lacks an in-depth theoretical analysis of why CSSD outperforms BASD, relying primarily on empirical observations rather than explaining the underlying learning dynamics.

5. **No Comparison with Other Knowledge Transfer Methods:** The study does not compare the proposed self-distillation methods with other knowledge transfer techniques that might be applicable in this context.

6 Minor Questions/Minor Weakness

1. How robust are the performance improvements across different hyperparameter settings? The paper mentions grid search but provides limited details about sensitivity to different α and T values.

2. The paper does not discuss potential trade-offs between privacy, power consumption, and classification performance at different sampling rates, which would be valuable for application-specific deployment decisions.

3. It would be interesting to see how these self-distillation techniques compare to specialized architectures designed specifically for low-quality audio inputs.

7 Suggestions as a Reviewer

To strengthen this paper, I recommend adding a theoretical analysis explaining why CSSD outperforms BASD, potentially exploring the information transfer between high and low sampling rates. The authors should conduct statistical significance tests to verify the robustness of performance differences. Including experiments with a wider range of sampling rates (e.g., 500 Hz, 4 kHz, 8 kHz) would help identify optimal operating points for different applications. A comparison with other knowledge distillation techniques would provide broader context for the contribution. Finally, investigating the diminishing returns with additional student generations could yield insights into the limitations of self-distillation for this task and potentially lead to improved distillation strategies.

8 Rating and Justification

I would rate this paper as **8/10**. The work addresses an important practical issue with novel and effective solutions, demonstrates consistent improvements across multiple models and datasets, and provides quantifiable benefits for real-world deployment. However, the lack of theoretical foundation and limited exploration of alternative approaches prevent it from achieving a higher score.

9 Bonus Assignment Section

The paper shows results on ESC-50 , TAU-19 and User dataset on models like CNN14,MobileNetV2 and ResNet, out of which I was able to find only ESC-50 publicly available.

The complete code of this paper isn't available , I implemented a few things by myself , I was able to produce results on ESC-50 dataset as it is publicly available on CNN14 and MobileNetV2 models.

I ran the experiment 3 times and recorded the average of those runs. Here is the comparitive table of original paper's result and my results on ESC-50:

article booktabs

		Original				My Results			
	Model	Raw	Fine-tuned	BASD	CSSD	raw	finetuned	BASD	CSSD
*2 kHz	CNN14	0.245	0.655	0.704	0.719	0.247	0.661	0.704	0.725
	MBNetV2	0.306	0.623	0.652	0.661	0.307	0.624	0.652	0.665
*1 kHz	CNN14	0.073	0.451	0.502	0.498	0.078	0.455	0.508	0.503
	MBNetV2	0.109	0.444	0.479	0.485	0.111	0.444	0.486	0.487

Table 1: Comparison of Original and My Results on ESC-50

I also tried to do the dora fine tuning but ran into errors , but I have uploaded the dora.py file for reference.