Interspeech 2024
1-5 September 2024, Kos, Greece

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

Dawei Liang, Alice Zhang, David Harwath, Edison Thomaz

The University of Texas at Austin, USA

dawei.liang@utexas.edu, alice.zhang@austin.utexas.edu, harwath@cs.utexas.edu, ethomaz@utexas.edu

Abstract

Acoustic scene and event classification is gaining traction in mobile health and wearable applications. Traditionally, relevant research focused on high-quality inputs (sampling rates $\geq$ 16 kHz). However, lower sampling rates (e.g., 1 kHz - 2 kHz) offer enhanced privacy and reduced power consumption, crucial for continuous mobile use. This study introduces efficient methods for optimizing pre-trained audio neural networks (PANNs) targeting low-quality audio, employing Born-Again self-distillation (BASD) and a cross-sampling-rate self-distillation (CSSD) strategy. Testing three PANNs with diverse mobile datasets reveals that both strategies boost model inference performance, yielding an absolute accuracy / F1 gain ranging from 1% to 6% compared to a baseline without distillation, while sampling at very low rates (1 kHz - 2 kHz). Notably, CSSD shows greater benefits, suggesting models trained on high-quality audio adapt better to lower resolutions, despite the shift in input quality.

Index Terms: acoustic scene and event classification, knowledge distillation, mobile health

1. Introduction

Acoustic scene and event classification, along with broader acoustic sensing, have served as a foundation for numerous human-centered applications in areas such as human activity recognition [1, 2, 3] and health monitoring [4, 5]. The advent of advanced pre-trained audio neural networks (PANNs) [6, 7, 8] utilizing extensive publicly accessible acoustic datasets like Google AudioSet [9] and ESC-50 [10] has significantly expanded the capabilities of audio classification. These datasets, mostly recorded at high sampling rates between 16 kHz and 22 kHz, have directed most audio classification and tagging research towards exploiting high-resolution audio for model development and assessment.

For real-world mobile and wearable devices, continuous audio capture and processing can sometimes pose a significant challenge. This is particularly the case for longitudinal uses, such as mobile health monitoring, where the microphone must always remain active [11]. The limited battery life of common edge devices makes processing high-frequency audio signals highly energy-intensive [12, 13]. Additionally, high-resolution audio increases the risk of revealing sensitive speech information, raising widespread privacy concerns [14, 15]. While techniques like signal reconstruction [4] or masking [16] can obscure intelligible speech, they tend to be application-specific and could add further computational demands on the device.

To overcome these hurdles, a viable strategy involves lowering the device's input sampling rate, particularly for classification targets less affected by this adjustment. This approach not only decreases power usage during signal processing, but also enhances privacy by making audio less intelligible [12]. Towards this end, this paper presents an efficient method to improve PANNs for audio classification at significantly reduced sampling rates (1 kHz to 2 kHz) through model self-distillation. Typically, knowledge distillation involves a complex "teacher" network guiding a simpler "student" network to achieve comparable performance with a reduced model size [17]. While self-distillation, where a model learns from itself, has been shown to boost model generalization in image classification [18], its application in audio classification, particularly with low-quality audio, remains under-explored. Our research examines the use of PANNs for classifying audio from low-rate samples, showing how self-distillation aids model adaptation. We find that both vanilla model self-distillation and a novel cross-sampling-rate method enhance model inference. Additionally, starting training with high-resolution audio appears to facilitate the model's adaptation to the very-low sampling rates.

2. Related Work

To the best of our knowledge, there is a scarcity of research on systematically analyzing acoustic scene and event classification using signals with low sampling rates. Ferraro *et al.* [19] recently studied the impact of varying input frequency resolutions on model performance for music tagging; however, their task differed from ours, and the sampling rates they investigated were relatively high ($\geq$ 12 kHz) compared to our intended range. Another work by Mollyn *et al.* [12] examined strategies to counteract the decline in acoustic mobile sensing performance with low-sampled signals through the addition of other modalities, such as motion data inputs on a device. In contrast, our work concentrates on leveraging advanced PANN models within the exclusive realm of audio data.

To recover speech quality lost during transmission or storage, audio super-resolution (SR) has been explored, aiming at reconstructing speech from lower frequency signals [20, 21, 22, 23]. However, while these efforts aim to address bandwidth constraints, our approach deliberately lowers audio quality to improve power usage and privacy issues on edge devices. Consequently, while audio SR typically deals with relatively high-frequency signals (e.g., $\geq$ 8 kHz), our focus is on inputs sampled at much lower rates, such as below 2 kHz.

Knowledge distillation is a technique where knowledge from a larger "teacher" model is transferred to a smaller "student" model [17], often for the purpose of model compression. Subsequent studies have explored the concept of self-distillation, where a model distills knowledge from itself to enhance its own generalization capabilities, a method found to be effective in computer vision [18, 24]. Although recent re-

10.21437/Interspeech.2024-2285

search has applied self-distillation to speech-related tasks like fake speech detection [25], its application to audio classification, particularly at lower sampling rates, has not been extensively investigated. We demonstrate that self-distillation, especially with a novel setup across input resolutions, effectively enhances three tested PANNs using these low-sampled inputs while obviating the need for an external teacher model architecture. This represents a novel aspect of our research.

3. Model Self-Distillation Setup

Conventionally, knowledge distillation in neural networks involves two models with differing complexities, both trained on the same data. The more complex model captures detailed data representations, with its outputs and internal states serving to guide the simpler model. Diverging from this, Furlanello *et al.* [18] later introduced a method where knowledge is transferred between generations of models with the same capacity. In this approach, once a model stabilizes, a new "student" with the same or similar design is trained not only to predict accurately but also to emulate the predecessor's predictions, creating what are known as Born-Again Networks (BANs). The underlying idea is that a model distilling knowledge from its previous iteration can possibly uncover subtle details missed in earlier training phases, such as specific instances where the original model's confidence varies. A benefit of this self-distillation process is the elimination of the need for a separate, more complex teacher model, streamlining the learning process.

For acoustic classifiers initially optimized on high-frequency audio datasets such as PANNs, the optimal learning potential of these models may not be fully harnessed for significantly lower-sampled audio due to discrepancies in signal frequency resolutions. Hence, self-distillation could enhance their adaptation to such low-sampled data. Moreover, by initially training models on high-frequency data, they can more effectively discern the full spectrum of audio patterns, which could in turn enhance their subsequent iterations trained on partially-sampled inputs.

3.1. Proposed Pipelines

Figure 1 depicts the pipelines according to our assumptions. The first pipeline follows the Born-Again self-distillation (BASD) framework. Initially, an acoustic model is developed in three steps. Firstly, the acoustic neural network is trained using low-quality audio signals as the initial iteration. Following this, the initial iteration serves as a fixed teacher model, and a second iteration is trained using both ground truth signals and outputs from the first iteration. The output logits from the teacher model's final fully-connected layer are utilized as the teacher knowledge. Finally, the second iteration is deployed for inference, termed as fine-grained generation. In this BASD approach, both teacher and student models receive the same low-sampled audio input.

The second pipeline also comprises three steps. However, the initial iteration of the acoustic neural network is trained using high frequency resolution audio signals, e.g., 16 kHz, instead of low-quality ones. Subsequently, the teacher and student networks are trained using synchronized pairwise signals of high and low quality, respectively. Despite requiring different quality inputs during training, this approach is deemed reasonable and scalable since the objective is to adapt the student model to low-sampled audio during deployment. These pairwise signals can be easily prepared by collecting high-

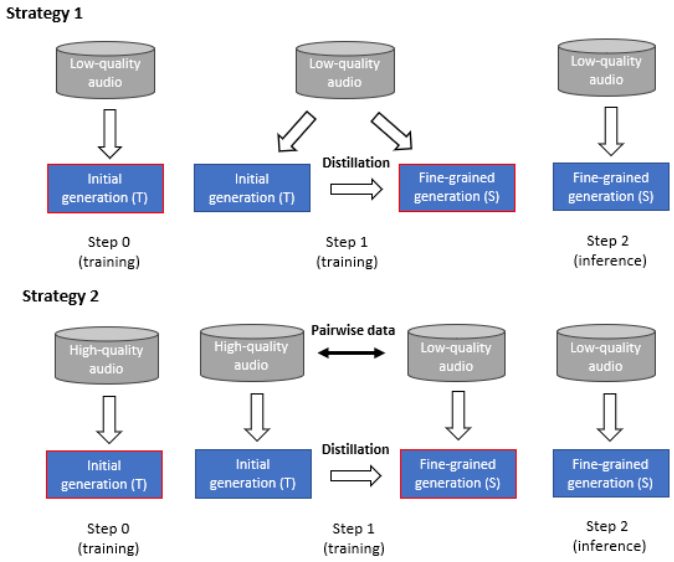

Figure 1: *Our pipelines of vanilla Born-Again self-distillation (BASD, strategy 1) and cross-sampling-rate self-distillation (CSSD, strategy 2). The terms "initial" and "fine-grained" generations denote different stages of the same model. The generations highlighted in red are the ones being trained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

resolution signals during training and down-sampling them to the target sampling rates. Once the student model is trained, it can be deployed for inference based solely on low-sampled signals, akin to BASD. As the teacher and student models in this setup receive audio signals of differing types, distinct from a typical knowledge distillation setup, we term this approach as cross-sampling-rate self-distillation (CSSD) in our paper.

3.2. Distillation Loss

In both pipelines, we adhere to the conventional knowledge distillation setup to calculate the student loss during training. The total student loss during training is determined as a weighted average of cross-entropy and the Kullback-Leibler (KL) divergence loss with respect to the teacher logits [17]:

$$\mathcal{L} = \alpha * \mathcal{L}_{CE}(y_s, y) + (1 - \alpha) * T^2 \mathcal{L}_{KL}(q_s, q_t) \quad (1)$$

where $\mathcal{L}_{CE}$ and $\mathcal{L}_{KL}$ denote the cross-entropy loss and the KL-divergence loss, respectively. $\mathcal{L}_{CE}$ measures the disparity between the student output, y_s, and the hard labels of sound classes, y. $\mathcal{L}_{KL}$ compares a smoothed version of the student output, q_s, with the smoothed teacher output, q_t. Two hyperparameters, α and T, control the influence of each loss component and the level of smoothness, respectively. The smoothed outputs are calculated as follows:

$$q_s = ln(\frac{exp(y_s/T)}{\sum_j exp(y_{s_j}/T)}) \qquad q_t = \frac{exp(y_t/T)}{\sum_j exp(y_{t_j}/T)} \quad (2)$$

where y_t denotes the teacher output and j is the target index.

4. Experimental Setup

4.1. Data

We utilized three datasets for our study. The first two datasets include the public ESC-50 Environmental Sound Classification

Table 1: *Comparison of audio classification results (accuracy for ESC / TAU and F1 for user data) with / without model self-distillation for very-low-sampled audio models. "Raw": models trained solely on 16 kHz audio; "Fine-tuned": models fine-tuned with low-sampled audio. Only results for the best distillation method (CSSD) vs fine-tuned baselines are presented for user data, as this aligns with our deployment strategy, and observations for raw models and BASD have been demonstrated using public data.*

		ESC-50				TAU-2019-Mobile				User	
		Raw	**Fine-tuned**	**BASD**	**CSSD**	**Raw**	**Fine-tuned**	**BASD**	**CSSD**	**Fine-tuned**	**CSSD**
2 kHz	CNN14	0.245	0.655	0.704	**0.719**	0.094	0.526	0.531	**0.555**	0.655	**0.700**
	ResNet38	0.455	0.645	0.668	**0.686**	0.333	0.524	0.514	**0.536**	0.661	**0.678**
	MBNetV2	0.306	0.623	0.652	**0.661**	0.184	0.532	0.555	**0.561**	0.675	**0.685**
1 kHz	CNN14	0.073	0.451	**0.502**	0.498	0.044	0.455	0.464	**0.474**	0.528	**0.570**
	ResNet38	0.205	0.432	0.466	**0.493**	0.200	0.436	0.429	**0.448**	0.570	**0.593**
	MBNetV2	0.109	0.444	0.479	**0.485**	0.101	0.449	0.481	**0.491**	0.561	**0.580**

dataset [10] and the TAU Urban Acoustic Scenes 2019 Mobile dataset [26]. ESC-50 comprises 2,000 crowd-sourced 5-second audio recordings across 50 balanced semantic sound categories commonly encountered in mobile applications, encompassing various human and contextual sounds. On the other hand, TAU-Mobile is notably larger, containing 46 hours of audio divided into approximately 16K 10-second segments across 10 acoustic scene categories captured by various mobile devices.

Alongside the public datasets, we augmented our analysis with proprietary user audio data. The user data includes sound classes relevant to everyday human activities, collected from multiple participants' homes using commercial smartphones in real-life settings with IRB approval. This dataset comprises 8.9 hours of 16 kHz mono audio, segmented every 5s, and manually annotated across 16 activity sound categories including *bathing, flushing toilet, brushing teeth, shaving, frying food, chopping, heating food, boiling water, using blender, television, playing music, vacuum cleaning, washing hands, speech, strolling*, and an additional *null* category. These datasets capture real-world mobile sensing scenarios, a primary focus of our work.

4.2. Model Configuration

Our study employed pre-trained CNN14 [6], ResNet38 [27], and MobileNetV2 [28] for audio classification. Prior to fine-tuning on our dataset, all models were initialized with pre-training weights from AudioSet [6], and their final output layers were replaced with custom fully-connected layers tailored to our target classes. For standard fine-tuning without model distillation, we utilized cross-entropy loss. For fine-tuning with BASD or CSSD, we employed the aforementioned knowledge distillation loss. The learning rates for the models were set to 1×10^{-3} and 1×10^{-4} for ESC-50 / user data and TAU-Mobile data, respectively. We utilized the Adam optimizer [29] with a Beta value of (0.9, 0.999). The mini-batch size was 64. Models were developed using PyTorch [30]. The α and T were determined using a grid search, detailed in the supplementary page.

4.3. Data Preparation

We down-sampled both the public and user datasets to our target sampling rates of 1 kHz and 2 kHz to simulate low-quality audio capture. For standard model fine-tuning and BASD, only the down-sampled data was utilized for both development and evaluation. In contrast, for CSSD, both high-quality and low-quality audio pairs were used for student model development, with the teacher and student models receiving high-quality and low-quality inputs, respectively.

To accommodate PANNs, log mel spectrogram features were necessary. Aligned with the original PANNs, which were initialized with model weights from 16 kHz data, we extracted log mel spectrogram features at 16 kHz for the low-quality audio by up-sampling. However, it is important to note that the high-frequency information was lost during down-sampling, resulting in log mel features devoid of high-frequency details, even though they were extracted at 16 kHz. No additional post-processing was applied to these features.

We employed the 5-fold evaluation split provided by the ESC-50 dataset to assess our model's performance on this dataset. For the TAU-Mobile dataset, we adhered to the data splitting scheme outlined in the DCASE 2019 challenge [26]. Additionally, we randomly allocated 25% of the training subset for hyper-parameter tuning and model validation. The user data underwent a 5-fold evaluation procedure as well, where each fold contained only non-overlapping participants. Throughout the 5-fold evaluations, the hyper-parameters, such as α and T, remained consistent across folds. To ensure fair comparison, we maintained a fixed random seed throughout the development of both baseline and self-distilled models across all cross-validation folds, ensuring reproducibility of results.

5. Results

5.1. Overall Results on the Public Datasets

Table 1 presents a comparison of macro-averaged class accuracy values between baseline models fine-tuned on the public datasets without model self-distillation and models developed with distillation. Additionally, we include the accuracy of models trained solely with 16 kHz audio and applied to inference at low sampling rates, denoted as "raw". From the table, it is evident that fine-tuning the models on the target low-sampled audio was crucial for enabling the adoption of PANNs with low-quality audio. Also, BASD and CSSD consistently surpassed baseline fine-tuned models without distillation across the tested low-sampled scenarios from ESC-50 and TAU-Mobile. The absolute accuracy gains ranged from 1% to 6%, typically around 3% to 5%, and remained consistent across varying complexities of PANNs, from the complex CNN14 to the much more lightweight MobileNet. This underscores the effectiveness of self-distillation in enhancing the model's learning capabilities.

Furthermore, CSSD consistently outperformed BASD across nearly all tested scenarios, suggesting the advantage and stability of initializing models with high-quality data to guide subsequent fine-grained generations, despite the input mismatch between teacher and student generations. Further discussions on the impact of model generations on inference will be provided in section *Ablation Study*.

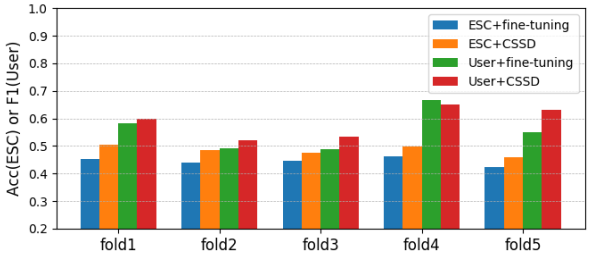

Figure 2: *5-fold performance variations: Fine-tuned vs. CSSD with MobileNetV2 at 2 kHz.*

5.2. Results on User Data

To comprehensively assess model performance on our user data, we used the F1 score, which is the harmonic mean of precision and recall, providing a single measure of both metrics in classification. Given the highly unbalanced nature of our user data, we utilized macro-averaged class F1 scores across all folds.

Table 1 also illustrates the comparison between fine-tuned baseline models and models developed with CSSD. It is evident that CSSD consistently enhances the performance of all tested model architectures at both sampling rates, aligning with our observations from public datasets. Despite the initial generations (16 kHz) of the three models achieving F1 scores of 0.789, 0.695, and 0.732, respectively, this analysis emphasizes CSSD's advantages in improving model learning capabilities over sole fine-tuning at low sampling rates, with added benefits of low-cost and less sensitive audio for mobile device processing.

5.3. Ablation Study

Fold-level Variations: We conducted 5-fold evaluation for both the ESC-50 and our user data to assess model performance. Figure 2 illustrates fold-level variations in performance for MobileNetV2, providing insight into model generalization ability. The results show that CSSD outperforms baseline fine-tuned models without distillation across almost all folds, and this is consistent for both 1 kHz and 2 kHz and on both datasets, indicating improved generalization across different data splits.

Individual Class Performance: In addition to overall model performance, we examined changes in individual class accuracy. Table 2 displays sample human sound classes from ESC-50 and our user data alongside the average class accuracy scores over five folds for both the fine-tuned baseline models on low-sampled audio and models developed with CSSD, using MobileNetV2 as an example. Despite the overall performance improvement brought by CSSD, the impact on individual classes is less straightforward. Additionally, for the same model derived from the same dataset, the class performance at 2 kHz does not consistently outperform its performance at 1 kHz. This suggests that the model training and distillation process is a complex, global optimization process, and the benefits of CSSD may not always translate to individual classes.

Influence of Student Generations: We explored the potential of additional self-distillation to enhance the student model's performance and the impact of self-distillation from low to high-quality audio. Interestingly, when adding an extra generation to the BASD process for MobileNetV2 at 1 kHz, we observed a notable decrease in accuracy / F1 values for the new generation across all datasets: 0.221 for ESC-50, 0.483 for TAU-Mobile, and 0.295 for our user data. This suggests that increasing the number of student generations may not guarantee better distillation outcomes. Additionally, CSSD appeared more effective for distillation from high-quality to low-quality audio. For example, using MobileNetV2 at 1 kHz for the initial generation and 2 kHz for the fine-grained generation yielded inference accuracy / F1 scores of 0.635, 0.369, and 0.659 for ESC-50, TAU-Mobile, and user data, respectively. This suggests that initial model generations developed with high-quality audio (e.g., 16 kHz) may better capture sound patterns for guiding subsequent generations. Further research is needed to explore these phenomena.

Table 2: *Accuracy change of sample classes before/after CSSD.*

	1 kHz	**2 kHz**
coughing (ESC)	0.63 / 0.78	0.78 / 0.73
crying_baby (ESC)	0.93 / 0.83	0.75 / 0.88
snoring (ESC)	0.53 / 0.63	0.83 / 0.80
strolling (user)	0.94 / 0.89	0.84 / 0.92
vacuum cleaning (user)	0.49 / 0.47	0.68 / 0.75

Table 3: *Power (300 mAh battery) and privacy gains with lower sampling rates. Higher metric values indicate greater gains.*

	16 kHz	**2 kHz**	**1 kHz**
Battery duration (hr)	50.6	58.4	59.6
WER (%)	5.4	55.6	96.5

Power and Privacy Benefits: Table 3 showcases the benefits to power consumption and privacy in sampling audio at a lower frequency. We measured the total current draw of an Ambiq Apollo3 microcontroller performing fast Fourier transforms on outputs from a Knowles SPH064 microphone [31] at individual sampling rates. The current consumption decreased from 5.93 mA to 5.03 mA as the sampling rate decreased from 16 kHz to 1kHz. For an edge device such as a smartwatch with a battery capacity of 300 mAh, these savings could extend the device's battery life by almost 10 hours. Additionally, we conducted a speech intelligibility study with 10 native US English speakers transcribing 9 audio clips (3 clips at each sampling rate) from the LibriSpeech dataset [32]. We observed significant increases in the word error rate (WER) as the sampling rate decreased, indicating reduced audio intelligibility. These findings align with prior research [12], supporting our hypothesis that low-sampled audio enhances privacy by making audio less intelligible.

6. Conclusion and Future Work

In this study, we delved into audio classification using very low sampling frequencies, targeting battery and privacy concerns for long-term mobile applications. We explored Born-Again model self-distillation (BASD) and a novel cross-sampling-rate self-distillation (CSSD) strategy for developing pre-trained audio neural networks on low-quality audio without additional teacher architecture. Our experiments, conducted with three models across real-world datasets, showcased enhanced model inference performance. CSSD particularly stood out, yielding accuracy / F1 gains up to 6%. While constrained by space limits and we could not extend our analysis with more scenarios, our aim is to introduce a novel method to adapting pre-trained models to low-quality inputs with minimal additional model development, a direction not widely explored in the literature. Future enhancements could involve ensemble methods for distillation.

7. References

[1] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "Soundsense: scalable sound sensing for people-centric applications on mobile phones," in *MobiSys 2009*.

[2] N. D. Lane, P. Georgiev, and L. Qendro, "Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *UbiComp 2015*.

[3] D. Liang and E. Thomaz, "Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos," *IMWUT*, vol. 3, no. 1, pp. 1–18, 2019.

[4] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, "Accurate and privacy preserving cough sensing using a low-cost microphone," in *UbiComp 2011*.

[5] E. Thomaz, C. Zhang, I. Essa, and G. D. Abowd, "Inferring meal eating activities in real world settings from ambient sounds: A feasibility study," in *IUI 2015*.

[6] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM TASLP*, vol. 28, pp. 2880–2894, 2020.

[7] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," in *INTERSPEECH 2021*.

[8] T. Pellegrini, I. Khalfaoui-Hassani, E. Labbé, and T. Masquelier, "Adapting a convnext model to audio classification on audioset," in *INTERSPEECH 2023*.

[9] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP 2017*.

[10] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *ACM MM 2015*.

[11] B. Little *et al.*, "Deep learning-based automated speech detection as a marker of social functioning in late-life depression," *Psychological medicine*, vol. 51, no. 9, pp. 1441–1450, 2021.

[12] V. Mollyn, K. Ahuja, D. Verma, C. Harrison, and M. Goel, "Samosa: Sensing activities with motion and subsampled audio," *IMWUT*, vol. 6, no. 3, pp. 1–19, 2022.

[13] D. Liang, A. Zhang, and E. Thomaz, "Automated face-to-face conversation detection on a commodity smartwatch with acoustic sensing," *IMWUT*, vol. 7, no. 3, pp. 1–29, 2023.

[14] P. P. Zarazaga *et al.*, "Sound privacy: A conversational speech corpus for quantifying the experience of privacy." in *Interspeech 2019*.

[15] D. Liang, W. Song, and E. Thomaz, "Characterizing the effect of audio degradation on privacy perception and inference performance in audio-based human activity recognition," in *MobileHCI 2020*.

[16] D. Liaqat, E. Nemati, M. Rahman, and J. Kuang, "A method for preserving privacy during audio recordings by filtering speech," in *LSC 2017*. IEEE, 2017, pp. 79–82.

[17] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv 2015*.

[18] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *ICML 2018*.

[19] A. Ferraro, D. Bogdanov, X. S. Jay, H. Jeon, and J. Yoon, "How low can you go? reducing frequency and time resolution in current cnn architectures for music auto-tagging," in *EUSIPCO 2020*.

[20] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super resolution using neural networks," *arXiv 2017*.

[21] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson, "Time-frequency networks for audio super-resolution," in *ICASSP 2018*.

[22] H. Wang and D. Wang, "Towards robust speech super-resolution," *IEEE/ACM ASLP*, vol. 29, pp. 2058–2066, 2021.

[23] Y. Li, M. Tagliasacchi, O. Rybakov, V. Ungureanu, and D. Roblek, "Real-time speech frequency bandwidth extension," in *ICASSP 2021*.

[24] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE PAMI*, vol. 44, no. 6, pp. 3048–3068, 2021.

[25] J. Xue, C. Fan, J. Yi, C. Wang, Z. Wen, D. Zhang, and Z. Lv, "Learning from yourself: A self-distillation method for fake speech detection," in *ICASSP 2023*.

[26] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *DCASE 2018*.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR 2016*.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR 2018*.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015*.

[30] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS 2019*.

[31] *Digital Zero-Height SiSonicTM Microphone With Multiple Performance Modes*, Knowles Electronics, 2014, rev. A.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP 2015*.