Distilled Non-Semantic Speech Embeddings with Binary Neural Networks for Low-Resource Devices

Aaqib Saeed and Harlin Lee, *Member*

***Abstract*—This work introduces BRILLsson, a novel binary neural network-based representation learning model for a broad range of non-semantic speech tasks. We train the model with knowledge distillation from a large and real-valued TRILLsson model with only a fraction of the dataset used to train TRILLsson. The resulting BRILLsson models are only 2MB in size with a latency less than 8ms, making them suitable for deployment in low-resource devices such as wearables. We evaluate BRILLsson on eight benchmark tasks (including but not limited to spoken language identification, emotion recognition, heath condition diagnosis, and keyword spotting), and demonstrate that our proposed ultra-light and low-latency models perform as well as large-scale models.**

***Index Terms*—speech representations, knowledge distillation, paralinguistic tasks, binary neural networks, digital health, internet-of-things**

I. Introduction

Representation learning takes advantage of large amounts of unlabeled data to learn features that can be used for a variety of downstream signal processing and machine learning tasks. This has demonstrated especially impressive performance in speech and audio processing [1], [2], [3], as the ubiquity of smart phones, watches, and home appliances has made it easy and inexpensive to collect a wealth of unlabeled audio signals. In particular, there is growing interest in using representation learning to build general-purpose models for non-semantic speech tasks, which are problems related to human speech other than its meaning, such as spoken language identification [4], emotion recognition [5], heath condition diagnosis [6], keyword spotting [7] and more.

However, the large sizes of the trained models and the amount of computational resources that are required to run them on newly acquired data have stalled the real-world deployment of existing models for non-semantic speech applications. These assumptions are critical in mobile computing, edge computing, internet-of-things (IoT), and tiny machine learning (tinyML) settings, which are where many paralinguistic speech applications actually take place, e.g., in small wearable for healthcare or with voice-controlled artificial intelligence (AI) assistants in smart devices. Although models such as FRILL [8] and TRILLsson [3] were recently proposed to reduce the complexity and size of the deep learning models, a large gap still remains between *highly effective* (i.e., accurate) and *highly efficient* (i.e., light enough to be run on devices as small as smartwatches) representation learning models for paralinguistic speech tasks.

To this end, we design and evaluate BRILLsson, binary neural networks (BNNs) [9] that are small and fast enough to be deployed in devices with limited memory and computational resources. BNNs have weights of only +1 or -1, which make them ideal compact architectures, especially in conjunction with co-designed machine learning hardware that one may see in modern IoT applications. Furthermore, we employ knowledge distillation [10], [11] from TRILLsson to BRILLsson, and show that distillation can be achieved using data that is slightly unrelated or smaller than the one used to train the original model, which is beneficial when original data is not available or so large that it requires extensive computing power. Finally, we illustrate that BRILLsson achieves performance on many non-semantic speech benchmark and other tasks that is comparable to that of much larger models.

In summary, our main contributions are:

- Develop and open-source BRILLsson [1], ultra-light and fast models for representation learning that are suitable for low-resource devices. BRILLsson's size is only 2MB, and its latency is less than 8ms.
- Perform successful knowledge transfer via embedding distillation from a large-scale real-valued model (i.e., EfficientNet-v2) to binary neural networks. While similar approaches have been explored in image classification [12] and speech separation [13], ours is the first in the context of general-purpose representation learning for non-semantic speech.
- Demonstrate that despite their compressed size, our BNNs perform comparably to TRILL, FRILL, and TRILLsson on eight different benchmark datasets.
- Our models are ideal for continuous on-device inference for privacy-preserving health monitoring (e.g., coughing, sneezing) due to its low-computational footprint.

II. Methods

Our objective is to create extremely compact general-purpose audio models that 1) generate informative embeddings (or representations) for a broad range of audio recognition tasks, and 2) can run efficiently on-device with low latency for low-resource devices, e.g., wearables without constant connection to the cloud. We use knowledge distillation to transfer learned representations of a large-scale pre-trained

A. Saeed (aaqib.saeed@philips.com) is with Philips Research, Eindhoven, The Netherlands. Philips infrastructure is not used in any manner for this project, neither for data processing nor for model development, training or evaluation.

H. Lee (harlin@math.ucla.edu) is with the Department of Mathematics, University of California, Los Angeles, USA. Her work is supported by NSF DMS-1952339.

[1]Will be released on Github upon publication.

teacher model to smaller BNN-based student models that are otherwise difficult to pre-train with self-supervised learning (or a similar strategy) due to their limited capacity. We would like to highlight that while distillation has been successfully leveraged for knowledge transfer for non-semantic speech before, to the best of our knowledge, this work is the first attempt at utilizing it for ultra-compact binary neural networks. Figure 1 provides high-level illustration of our approach, and the following subsections describe its essential building blocks.

A. Embedding Distillation

Distillation is a technique to create a model with smaller size and less computational load without sacrificing its effectiveness [10]. It transfers information from a large supervising model —the teacher ($\mathcal{F}_t$)— to a relatively small model —the student ($\mathcal{F}_s$)— with the goal of compression for efficient inference. The teacher is generally a fixed pre-trained network learned with a massive amount of high-quality data, in other words a privileged model. In contrast, the student is a low capacity network that is guided to imitate the output of the teacher. The information-rich signal from the teacher enables a compact student network to learn important aspects of the data that would otherwise be missed when solely minimizing a task-specific objective.

In the seminal paper, Hinton et al. [10] proposed to use softened class probabilities from the teacher to provide supervision, which acts as targets for the student model to optimize for. Here, as our teacher model provides 1024-dimensional embeddings, we instead leverage mean-squared-error loss and a linear layer ($\mathcal{F}_r$) of the same dimensionality on top of student model for distillation purpose [11]. We use one-second audio clips as inputs to teacher-student models to get outputs of 1024 dimensions, on top of which the following loss function is computed:

$$\mathcal{L}(\theta_s, \theta_r) = \frac{1}{2} \parallel \mathcal{F}_t(\mathbf{x}; \theta_t) - \mathcal{F}_r(\mathcal{F}_s(\mathbf{x}; \theta_s); \theta_r) \parallel_2^2 . \quad (1)$$

$\mathbf{x}$ is the training data, θ represents the respective model parameters, and $\mathcal{F}_r$ is the linear layer model representing a regressor function to match teacher's dimensionality that is discarded after distillation. We use a batch size of 512 and a fixed learning rate of 0.001 with an Adam optimizer [14] to train for approximately 234K steps with a single NVIDIA RTX3090 GPU.

B. Teacher: Large EfficientNet-V2 Model

For our teacher model $\mathcal{F}_t$, we use EfficientNet-v2 [15] from TRILLsson [3], which has achieved exceptional performance on several Non-Semantic Speech Benchmark [2] (NOSS) and other related tasks. This model was trained using a combination of two large-scale speech datasets (Speech AudioSet 4.9K and Libri-light 53K) and the CAP12 model (with non-public 606M parameters) via the teacher-student distillation framework. We choose version three of the EfficientNet-v2 model with 21.5M parameters and access it directly from

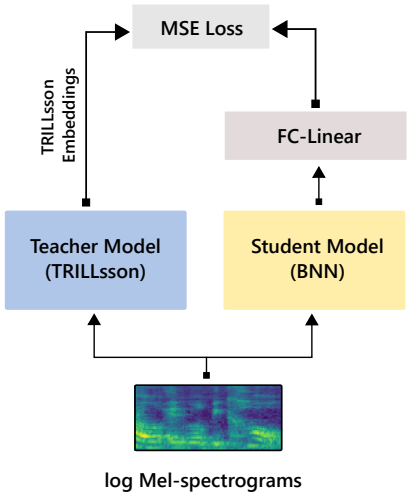

Fig. 1. Illustration of distilling binary neural models for non-semantic speech.

TensorFlow Hub (TFHub)[2] during training. This model has a front-end based on log-magnitude Mel spectrograms with 80 bins ranging from 125Hz to 7500Hz, uses window length of 25ms and hop length of 10ms, and is initially trained with frame width of 2s of audio.

There are larger models available within the TRILLsson family, but we choose the EfficientNet variant as it is both high-performing and less computationally demanding to accommodate a modest compute budget (e.g., hardware with a single GPU system). This allows us to train, or distill, longer in a short period of time with a large batch size, and demonstrate our central point that we can extract compact binary models purely with knowledge distillation. Further, EfficientNet is a mobile friendly architecture discovered by neural architecture search for image classification tasks with a large capacity. However, we do note that having access to more computing resources may allow one to leverage even bigger models with better supervision, which we leave for future work.

C. Student: Light-weight Binary Neural Networks

Our student models $\mathcal{F}_s$ are binary neural network (BNN) architectures with single-bit weights and activations. Because neurons in BNNs can have only two possible states, BNNs provide extreme compression and speed-up gains compared to real-valued artificial neural networks. These fast and energy efficient BNNs are well-suited for deployment on low-resource devices with limited memory and battery power. Specifically, we use binary convolutional networks: a binary DenseNet-28 [16] and MeliusNet [17] with around 4.5M and 6.4M parameters, respectively. Their sizes are less than 2MB in floating-point format, and 1.03MB for DenseNet-28 and 1.25MB for MeliusNet in quantized form, which are several folds smaller compared to models utilized in [2], [3], [8]. We use the Larq [18] framework for the implementation of BNNs.

Our student models are paired with an audio processing front-end based on log-magnitude Mel spectrograms, and can directly consume raw audio waveform. Our front-end uses

[2]https://tfhub.dev

TABLE I
OVERVIEW OF THE DOWNSTREAM EVALUATION DATASETS INVOLVING NON-SEMANTIC SPEECH.

Dataset	Task	Samples	Classes
MUSAN [21]	Speech, Music, and Noise	2,016	3
ESC-50 (HS) [22]	Human Sounds	400	10
Voxforge [4]	Spoken Language	176,436	6
SpeechCommands [7]	Commands Recognition	100,503	12
CREMA-D [5]	Emotion Detection	7,438	6
MSWC-EN [23]	Keyword Spotting	96,099	31
MSWC-ES [23]		28,039	20
Vocalsound [6]	Health Condition Monitoring	21,024	6

window size of 25ms, hop size of 10 ms, and 64 Mel-spaced frequency bins in the range of 60Hz to 7800Hz for 98 frames, corresponding to 980ms. These inputs to the BNNs are then mapped to latent vectors of size 576 for DenseNet-28 and 512 for MeliusNet. A GlobalMaxPooling layer on top reduces the size of the bottleneck layer by computing the maximum of all pixels in each output feature map. Finally, these compact output embeddings are provided to classifiers for downstream tasks. We note that to match TRILLsson's embedding dimensions, we add an additional binary fully-connected layer with 1024 hidden units $\mathcal{F}_r$, which is discarded after the training phase.

III. EXPERIMENTAL SETUP

A. Distillation Dataset

We perform knowledge distillation with open-source Libri-light dataset [19], which is derived from public audio books in the LibriVox project. It is the largest publicly available, unlabeled semi-supervised audio dataset to date. From this, we use a medium subset of the dataset with around 5193 hours of speech (approximately 321GB in size) due to our modest compute budget. We split each audio clip into non-overlapping one-second segments for training, resulting in around 12M examples. It is important to note that TRILLsson models are trained with 58K hours of speech data originating from both Audioset and Libri-light. Also, TRILLsson is trained with a teacher of massive scale, i.e., CAP12, that is in turn trained on 900K hours of audio from YT-U data [20].

B. Downstream Speech Sensing Tasks and Evaluation

We evaluate the effectiveness of our method on a broad range of tasks varying from spoken language identification, keyword spotting, accent recognition, identifying emotion, to health condition monitoring. Table I provides an overview and key characteristics of the 8 datasets. We use MUSAN [21] to evaluate the detection of music, speech and noise in audio clips. Voxforge [4] is used for identifying spoken English, Spanish, French, German, Russian, and Italian. We use SpeechCommands [7] with 12 classes for spoken commands and CREMA-D with 6 classes (anger, disgust, fear, happy/joy, neutral, sad) for emotion recognition. For human sounds task, we utilize 10 classes subset from ESC-50 [22], same as FRILL [8]. We also use microsets from multilingual word corpus (MSWC) for keyword spotting in English (EN) and Spanish (ES), each task with 31 and 20 classes, respectively. Lastly, Vocalsound [6] contains audio recording for detection of laughter, sighs, coughs, throat clearing, sneezes, and sniffs.

We follow standard train and test splits of the datasets except for MUSAN, where we randomly split the data into training (80%) and test (20%) sets. In case of ESC-50 (HS), following FRILL, we use first four folds as training set and the last fold as a test set. We evaluate the quality of learned representations with a linear classifier trained on top frozen feature extractor or encoder in a similar manner as prior work [2], [3]. The classifier is trained with a batch size of 64 (except for CREMA-D, where we use batch size of 32 due to relatively small size of the datasets) with learning rate of 0.001 with Adam optimizer [14] for 100 epochs. We use a randomly selected one-second segment from each audio clip in the training set, and evaluate the performance on the entire audio clip during testing.

C. Latency Benchmarking of Binary Neural Networks

We use Larq Compute Engine [18] for latency benchmarking of our BNN models. To align with prior work, we create a float-32 TFLite format model and run it for 150 runs in a single thread to get an averaged inference time on a device equipped with Snapdragon 855.

D. Baseline Models for Comparison

We compare our approach with five methods: TRILL, TRILL-Distilled, FRILL, teacher model TRILLsson3 (EfficientNet-V2), and TRILLsson1 (ResNet-50). TRILL [2] is a *TRIpLet-Loss Network* that is pre-trained with large amount of speech data from Audioset. It has shown to learn powerful representations for non-semantic speech tasks and achieved state-of-the-art performance on some of them when it was published in 2020. It uses a ResNet-50 network architecture, and its layer 19 has shown to provide the most useful features with dimensionality of 12288. TRILL-Distilled [2] is a smaller MobileNet-based model with 2048-dimensional embeddings that is trained with distillation to predict TRILL's embeddings. Along a similar line, FRILL [8] uses a MobileNetV3 model that is designed to be a fast variant of TRILL specifically focusing on mobile devices. It is trained with distillation to mimic the TRILL layer 19 representations. In our work, we use *Small 2.0 GAP* model with 2048-dimensional embeddings due to its best performance. Finally, we compare against the teacher model $\mathcal{F}_t$, as well as a smaller model in TRILLsson family, i.e., a ResNet-50 model number one with 5M parameters.

We access all the baseline models from TFHub and use them as frozen feature extractors. We use default audio front-end that comes along with the model from TFHub. Similar to our method, we only add a linear classification layer for evaluating performance on downstream tasks as explained in Section III-B. In cases where baseline models provide predictions per-time step, we average them to compute final prediction. Furthermore, we train and evaluate baseline models for tasks and datasets that were not presented in the prior work to establish fair comparison; for the rest, we use accuracy

TABLE II
GENERALIZATION PERFORMANCE OF BRILLSSON ON A RANGE OF NON-SEMANTIC SPEECH REPRESENTATIONS TASKS. WE TRAIN A LINEAR CLASSIFIER ON TOP OF DISTILLED FROZEN MODELS TO ASSESS THE QUALITY OF LEARNED EMBEDDINGS. RN-50 IS RESNET-50, EN-V2 IS EFFICIENTNET-V2, DN IS DENSENET-28, AND MN IS MELIUSNET.

Method	MUSAN	ESC-50 (HS)	Voxforge	SpeechCommands	CREMA-D	MSWC-EN	MSWC-ES	Vocalsound	Size (MB)	Latency (ms)
TRILL	98.2	86.4	84.5	81.9	66.2	81.3	88.0	88.2	98.1	275.3
TRILL-Distilled	98.5	87.9	80.0	80.2	70.2	74.4	87.9	85.8	107.1	22.5
FRILL	98.2	86.4	76.9	79.7	70.9	79.1	87.6	86.7	38.5	8.5
TRILLsson (RN-50)	98.5	60.0	98.6	91.2	81.3	91.4	94.5	87.2	22.0	-
TRILLsson (EN-v2)	98.7	87.5	99.2	93.2	83.2	87.2	93.9	89.0	99.0	
BRILLsson (DN)	93.0	85.0	70.8	88.7	65.3	87.6	88.6	80.2	2.0	6.4
BRILLsson (MN)	91.5	80.0	70.1	89.2	63.8	88.5	89.1	83.2	2.1	7.6
BRILLsson (T)	90.5	73.7	73.0	88.5	54.6	87.2	88.4	78.4	0.65	6.1

score as reported in [3], [8]. The baseline latency values when available are taken from FRILL [8]. Importantly, unlike previous works that trained multiple linear classifiers with different techniques, we only train and evaluate a logistic regressor implemented with a linear dense layer. An exhaustive search over classification methods may yield further improvement.

IV. RESULTS AND DISCUSSION

We evaluate the performance of our BRILLsson approach, and contextualize how well binary models generalize on a broad range of speech sensing tasks as compared to large-scale models. Table II presents results of BRILLsson along with five large-scale baselines models. First, we notice that the teacher model TRILLsson (EN-v2) overall perform well in comparison to other floating-point based models, which highlights the usefulness of this model as teacher for distillation. Our DenseNet (DN) based binary model demonstrate excellent performance on all considered tasks even with its small size. Note that DN is only 2MB including the audio front-end, whereas the teacher model has size of around 99MB. Similarly, our MeliusNet (MN) has similar or slightly better performance than DN, in particular on Vocalsound where it achieves accuracy of 83.2%. This is only 2% less than the acuracy of TRILL-Distilled. Interestingly, BRILLsson has superior generalization on keyword spotting tasks, achieving 89.2% accuracy on SpeechCommands and 88.5% on MSWC-EN. We once again emphasize that our BRILLsson models have latency of less than 8ms with size of merely around 2MB. Furthermore, we add a linear classifier after *batch-normalization-12* intermediate layer in DN model to experiment with creating an even smaller model, labeled BRILLsson (T) in Table II. The resulting model has size of 0.65MB and latency of 6.1ms. Interestingly, on Voxforge the tiny model achieves 73.0% while having only 1.4M parameters. These results demonstrate the usefulness of representations learned with BNNs and that a single BNN model can be used as a feature extractor on low-resource devices for multiple downstream tasks.

Along similar lines, we evaluate the quality of representations from intermediate layers of the distilled model using MN on SpeechCommands. For each intermediate layer, a classification head is added on top and trained in the same manner as previous experiments. The rest of the model is fixed during this phase. Then, we convert each model to TFLite format, evaluate its accuracy and latency, and report the results in Figure 2. We see that conversion to TFLite format does not result in a significant performance degradation. Also, we observe a trade-off between accuracy and speed, as expected; for instance, the model built on the layer *section-2-transition-pw* has the highest accuracy of 86.1%, but has relatively high latency of 7.3ms.

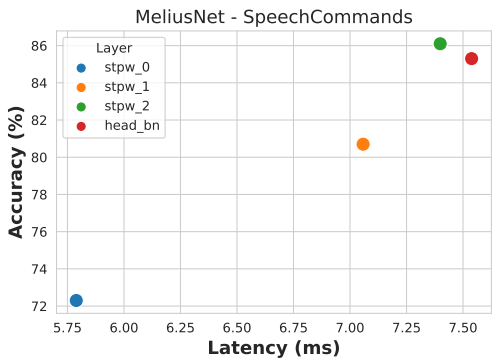

Fig. 2. Performance of intermediate layers' representations on SpeechCommands for MeliusNet TFLite format model. Each point corresponds to the accuaracy and latency of a model where we added a classifier after an intermediate layer. For example, a model with classifier after *section-1-transition-pw* layer achieves 81% accuracy with latency of 7ms. *stpw* is an abbreviation of *section-transition-pw*.

V. CONCLUSIONS

We have designed, developed, and publicly released BRILLsson— an extremely compact, fast, and flexible model for non-semantic speech representation learning. We used embedding distillation to transfer knowledge from an existing TRILLsson model to small binary neural networks. Our approach significantly reduced the model size while keeping the performance on par with large-scale real-valued counterparts, which is valuable for low-resource devices. While this work focused on utilizing existing neural architectures, we would like to explore neural architecture search methods in the future to design even more light-weight BNN models that are suitable for micro-controllers.

REFERENCES

[1] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3875–3879.

[2] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. de Chaumont Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards Learning a Universal Non-Semantic Representation of Speech," in *Proc. Interspeech 2020*, 2020, pp. 140–144.

[3] J. Shor and S. Venugopalan, "Trillsson: Distilled universal paralinguistic speech representations," *arXiv preprint arXiv:2203.00236*, 2022.

[4] K. MacLean, "Voxforge," *Ken MacLean.[Online]. Available: http://www. voxforge. org/home.[Acedido em 2012]*, 2018.

[5] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[6] Y. Gong, J. Yu, and J. Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 151–155.

[7] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[8] J. Peplinski, J. Shor, S. Joglekar, J. Garrison, and S. Patel, "Frill: A non-semantic speech embedding for mobile devices," *arXiv preprint arXiv:2011.04609*, 2020.

[9] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," *Advances in neural information processing systems*, vol. 28, 2015.

[10] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[11] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[12] S. Leroux, B. Vankeirsbilck, T. Verbelen, P. Simoens, and B. Dhoedt, "Training binary neural networks with knowledge transfer," *Neurocomputing*, vol. 396, pp. 534–541, 2020.

[13] X. Chen, G. Liu, J. Shi, J. Xu, and B. Xu, "Distilled binary neural network for monaural speech separation," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[15] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 096–10 106.

[16] J. Bethge, H. Yang, M. Bornstein, and C. Meinel, "Back to simplicity: How to train accurate bnns from scratch?" *arXiv preprint arXiv:1906.08637*, 2019.

[17] J. Bethge, C. Bartz, H. Yang, Y. Chen, and C. Meinel, "Meliusnet: Can binary neural networks achieve mobilenet-level accuracy?" *arXiv preprint arXiv:2001.05936*, 2020.

[18] L. Geiger and P. Team, "Larq: An open-source library for training binarized neural networks," *Journal of Open Source Software*, vol. 5, no. 45, p. 1746, Jan. 2020.

[19] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[20] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang *et al.*, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2109.13226*, 2021.

[21] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[22] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, 2015, pp. 1015–1018.

[23] M. Mazumder, S. Chitlangia, C. Banbury, Y. Kang, J. M. Ciro, K. Achorn, D. Galvez, M. Sabini, P. Mattson, D. Kanter *et al.*, "Multilingual spoken words corpus," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.