applied sciences

Article

Environment-Aware Knowledge Distillation for Improved Resource-Constrained Edge Speech Recognition

Arthur Pimentel [1,*], Heitor R. Guimarães [1], Anderson Avila [1,2] and Tiago H. Falk [1,2]

[1] Institut National de la Recherche Scientifique (INRS-EMT), Université du Québec, Montreal, QC H5A 1K6, Canada; heitor.guimaraes@inrs.ca (H.R.G.); anderson.avila@inrs.ca (A.A.); tiago.falk@inrs.ca (T.H.F.)
[2] INRS-UQO Mixed Research Unit on Cybersecurity, Gatineau, QC J8X 3X7, Canada
* Correspondence: arthur.pimentel@inrs.ca

Abstract: Recent advances in self-supervised learning have allowed automatic speech recognition (ASR) systems to achieve state-of-the-art (SOTA) word error rates (WER) while requiring only a fraction of the labeled data needed by its predecessors. Notwithstanding, while such models achieve SOTA results in matched train/test scenarios, their performance degrades substantially when tested in unseen conditions. To overcome this problem, strategies such as data augmentation and/or domain adaptation have been explored. Available models, however, are still too large to be considered for edge speech applications on resource-constrained devices; thus, model compression tools, such as knowledge distillation, are needed. In this paper, we propose three innovations on top of the existing DistilHuBERT distillation recipe: optimize the prediction heads, employ a targeted data augmentation method for different environmental scenarios, and employ a real-time environment estimator to choose between compressed models for inference. Experiments with the LibriSpeech dataset, corrupted with varying noise types and reverberation levels, show the proposed method outperforming several benchmark methods, both original and compressed, by as much as 48.4% and 89.2% in the word error reduction rate in extremely noisy and reverberant conditions, respectively, while reducing by 50% the number of parameters. Thus, the proposed method is well suited for resource-constrained edge speech recognition applications.

Keywords: automatic speech recognition; knowledge distillation; self-supervised learning; modulation spectrum; context awareness

check for updates

Citation: Pimentel, A.; Guimarães, H.R.; Avila, A.; Falk, T.H. Environment-Aware Knowledge Distillation for Improved Resource-Constrained Edge Speech Recognition. *Appl. Sci.* **2023**, *13*, 12571. https://doi.org/10.3390/app132312571

Academic Editor: Ying Shen, Cunhang Fan and Ya Li

Received: 18 October 2023
Revised: 18 November 2023
Accepted: 20 November 2023
Published: 22 November 2023

1. Introduction

Automatic speech recognition (ASR) aims to convert a continuous speech signal into a discrete text representation. While speech is one of the most efficient communication methods for humans [1], text is an important representation for machines, and a large number of techniques can be more easily applied to structure and understand the data. Recently, large deep learning models have achieved great success in ASR [2–7], with methods matching human speech recognition in a wide variety of acoustical environments [8].

Self-supervised speech representation learning (S3RL) has established itself as a driving force behind these innovations. In this learning paradigm, the aim is to extract meaningful data representations via the utilization of large unlabeled datasets by exploiting the data modality. Subsequently, the model is fine tuned with labeled data. Today, wav2vec 2.0 [2], HuBERT [3], and WavLM [4] represent the most widely deployed universal speech representations, achieving state-of-the-art results not only for ASR, but for other tasks such as speaker and emotion recognition [9].

While these models represent a breakthrough in terms of their performance, there is still a gap to bridge between ASR systems based on large models deployed on the cloud and ASR systems meant for edge deployment. Edge applications focus on bringing computing as close to the source of data as possible in order to reduce the latency and bandwidth use. This can be particularly important for speech recognition applications, where private and sensitive speaker data are typically sent over the network to be processed remotely on clusters hosting these complex models.

Existing universal S3RL methods, however, face two significant limitations in the context of edge applications: (i) their large size, which can be unfeasible for many edge applications, and (ii) their robustness, at inference time, against unseen environmental conditions, such as noise and reverberation. For example, the HuBERT model typically ranges from 95 million to 1 billion parameters, which is prohibitive on edge devices with limited storage and processing capacity. Model compression techniques, such as quantization, model pruning, and knowledge distillation, have been explored, with the former yielding the most promising outcome, as exemplified by the recent development of the "DistilHuBERT" model [10].

On the environmental robustness side, previous studies have shown that using signal-based enhancement techniques is usually insufficient, as the distortions introduced by these algorithms can also degrade the model performance (e.g., [11]). Unseen noise and reverberation levels, for example, are known to drastically reduce the accuracy of even state-of-the-art ASR systems [12] and can be highly sensitive to different environmental conditions [13,14]. While domain adaptation techniques, such as those proposed in "Robust HuBERT" [15] and "deHuBERT" [16], can alleviate this problem, the methods are not directly applicable for compression. Recent works have started to propose solutions that tackle compression and environmental robustness jointly (e.g., [17,18]). These solutions, however, have yet to be explored for ASR and have shown some sensitivity to varying environmental conditions.

In this work, we aim to fill this gap. More specifically, our overarching goal is three-fold: (1) adapt the existing DistilHuBERT representation [10] to make it better suited for ASR tasks, (2) utilize data augmentation to increase the robustness of compressed models to unseen conditions, and (3) propose a hierarchical environment-aware solution where compressed models optimized for different environmental conditions are chosen during inference time, thus making the compressed models more robust to varying environmental conditions typically seen in edge conditions.

The remainder of this work is organized as follows: Section 2 introduces the related work and necessary background. Section 3 presents the proposed methodology, while Section 4 introduces the study's experimental setup and Section 5 reports and discusses the obtained results. Lastly, Section 6 presents the conclusions.

2. Background Material

In this section, we provide a comprehensive overview of the background material needed.

2.1. Self-Supervised Speech Representation Learning

Self-supervised speech representation learning (S3RL) enables deep neural network models to capture meaningful and disentangled factors from raw speech waveforms [17]. S3RL can be seen as a special case of unsupervised learning as both schemes learn without annotations. Although conventional unsupervised methods rely on reconstruction or density estimation objectives, S3RL approaches rely on pretext tasks that exploit knowledge about the data modality used for training. Although supervised learning methods tend to learn stronger features than unsupervised learning approaches, they require costly and time-consuming work from human annotators to generate the required labels. S3RL techniques aim for the best of both worlds: training a powerful feature extractor using discriminative learning without the need for manual annotation of training examples [19]. The learned universal speech representations can then be used across numerous tasks [20].

In the context of S3RL, "upstream models" are designed for pre-training and learning general representations from unlabeled data; in essence, they serve as "feature generators" for downstream tasks. The goal is to capture high-level features and patterns in the input speech data. These models are trained on a large dataset without task-specific labels. In turn, the "downstream models" are task-specific models fine tuned on labeled data for a particular speech-related task, such as speech recognition or emotion classification. Downstream models are trained on a smaller dataset with task-specific labels. The pre-trained upstream model is thus fine tuned on this data to adapt its learned representations to the specific requirements of the target task.

While several universal representations exist (e.g., wav2vec, wav2vec 2.0, wavLM, and HuBERT), here we will focus on the HuBERT representation, as several variants have been proposed in the literature recently to either make it more robust to environment noise or to build more robust compressed versions for edge applications. This allows for several benchmark methods to be used for comparisons with the proposed method. It is important to emphasize that while the results reported herein will be based on a HuBERT representation, the proposed environment-aware method is applicable to any speech representation.

2.2. Hubert and Variants

The HuBERT model, as introduced in [3], represents a speech pre-training technique designed to learn effective speech representations by means of masked feature reconstruction. This architecture comprises two major components: the encoder and the context network. The encoder network, denoted as f, operates as a mapping function from the raw waveform sample $x_i \in \mathcal{X}$ to an intermediate feature representation $z_i \in \mathcal{Z}$. It is structured as a 1D convolutional network, encompassing seven layers with 512 feature maps. The kernel and stride sizes for each layer vary accordingly. Following this convolutional network, a GELU non-linear activation function [21] is applied. The context network, represented by the mapping function $g : \mathcal{Z} \mapsto \mathcal{C}$, serves a crucial role in the HuBERT architecture. Its fundamental principle revolves around mapping diverse features into a unified context vector via multiple Transformer encoders, thereby capturing long-range information. The interested reader can refer to the first figure of [3,22] for more details on the HuBERT model and the Transformer architecture, respectively.

Different versions of the HuBERT model have been developed, including the HuBERT *Base*, *Large*, and *X-Large* versions. The main difference between them is the size of their respective Transformer networks, comprising 12, 24, and 48 layers, respectively. Furthermore, it is worth noting that the HuBERT Base model was trained on a dataset encompassing 960 h of the LibriSpeech dataset [23], while the Large and X-Large versions were trained on a substantially larger dataset consisting of 60,000 h from the LibriLight dataset [24].

2.3. Making Hubert Robust to Noise

While larger versions of HuBERT have shown improved ASR accuracy in noisy conditions [17], for edge applications such as large footprint models, they can be problematic, and smaller models can be more sensitive to noise. To this end, the authors in [15] proposed the so-called *Robust HuBERT* method, where domain adversarial training was used to make the system more robust to environmental factors. More specifically, a domain discriminator is responsible for classifying the source of the distortions applied to the utterance. The HuBERT Base model was continually trained on more diverse data, with a probability of receiving speech data corrupted by varying types and levels of noise, including Gaussian noise and recorded noises taken from the MUSAN database [25]. Moreover, another technique to increase robustness is that of noise disentangling, as showed in [16]. The authors propose *deHuBERT*, a novel self-supervised training framework that encourages noise invariance in HuBERT's embedded contextual representations by introducing a second embedding from different noise-augmented signals, using a shared CNN encoder and a new pair of auxiliary loss functions.

2.4. Distilled Versions of Hubert

As mentioned above, one emerging topic in speech processing is that of edge ASR. For these applications, smaller models are needed; thus, knowledge distillation has been explored recently. The recent work in [10], for example, proposed the so-called *DistilHuBERT* model, where hidden representations from a HuBERT model were directly distilled to reduce model size and inference time.

More specifically, DistilHuBERT proposes a novel teacher–student framework for speech representation learning via multi-task knowledge distillation. The model consists of a CNN feature extractor and a small Transformer encoder. The idea is to learn to generate multiple teacher's hidden representations from shared representations. This is performed by predicting the teacher's hidden representations with separate prediction heads. This objective is a multi-task learning paradigm and encourages the Transformer encoder to produce compact representations for multiple prediction heads. DistilHuBERT then takes a frozen pre-trained HuBERT Base model and uses only three prediction heads to respectively predict the 4th, 8th, and 12th HuBERT hidden layers' output. Before pre-training, the student is initialized with the teacher's parameters. Then, the student's prediction heads learn to generate the teacher's hidden representations by minimizing the loss function $\mathcal{L}^{(l)} = \mathcal{L}^{(4)} + \mathcal{L}^{(8)} + \mathcal{L}^{(12)}$. After training, the heads are removed, as the multi-task learning paradigm forces the DistilHuBERT model to learn representations containing rich information. These layers, however, are chosen to maximize the performance across various speech-related tasks and may not be optimal for ASR. In our experiments, we explore how different prediction heads can improve the model performance.

One other method proposed recently for distilled universal representations is the one described in [17], called *RobustDistiller*. This is a recipe which combines data augmentation and multi-task denoising learning, which are incorporated into the knowledge distillation process. In the data augmentation step, an online contamination of the data is performed during the distillation process. In particular, the student model receives the noisy data as input, but the network's target is to reconstruct the clean representations of the teacher model. Thus, in the multi-task denoising learning step, beyond learning to reconstruct the teacher's representations, an additional enhancement head is responsible for rebuilding the clean speech waveform from the learned representations. In contrast to the usual enhancement techniques, the objective is to enforce the upstream model to carry enough information about the speech itself and not the noise components. While RobustDistiller was shown to outperform DistilHuBERT on keyword spotting, intent classification, and emotion recognition tasks, with noise and/or reverberation, it still showed some sensitivity to different environmental conditions.

Lastly, we explore if the distillation of a teacher model built to be robust also results in a robust compressed model. This is an important step to decide if robustness needs to be implemented at the teacher level or during the distillation process, as proposed herein. To this end, we apply the same distillation process used in DistilHuBERT, but applied it to the *Robust HuBERT* teacher described in Section 2.3. Henceforth, we term this model *DistilRobustHuBERT*.

3. Proposed Model

As mentioned previously, our proposed model incorporates three innovations to tackle the issues with the existing systems, as highlighted above. In the sections to follow, these three innovations are described.

3.1. Innovation #1: Modifying Prediction Heads

In an attempt to maximize the performance across the different tasks of the SUPERB [26] benchmark, the original DistilHuBERT recipe utilized as the prediction heads the 4th, 8th and 12th Transformer layers of the HuBERT Base teacher model [10]. These layers were shown to achieve improved accuracy across different tasks, but were not necessarily optimal for ASR. As such, our first innovation is to adapt the DistilHuBERT recipe to optimize the prediction heads for the ASR task.

According to the SUPERB instructions, after pre-training, the hidden states of different upstream layers are weighted, summed, and fed to the task-specific layers, with the weights from each layer changing depending on the downstream task. A larger weight indicates a greater contribution of the corresponding layer. We analyzed the weights from each layer of the HuBERT model and found that layers 8, 9, and 10 provided the greatest contribution for the ASR task; thus, our proposed method will rely on these three layers instead. Figure 1 shows the weight analysis of the HuBERT base model, fine tuned for ASR.

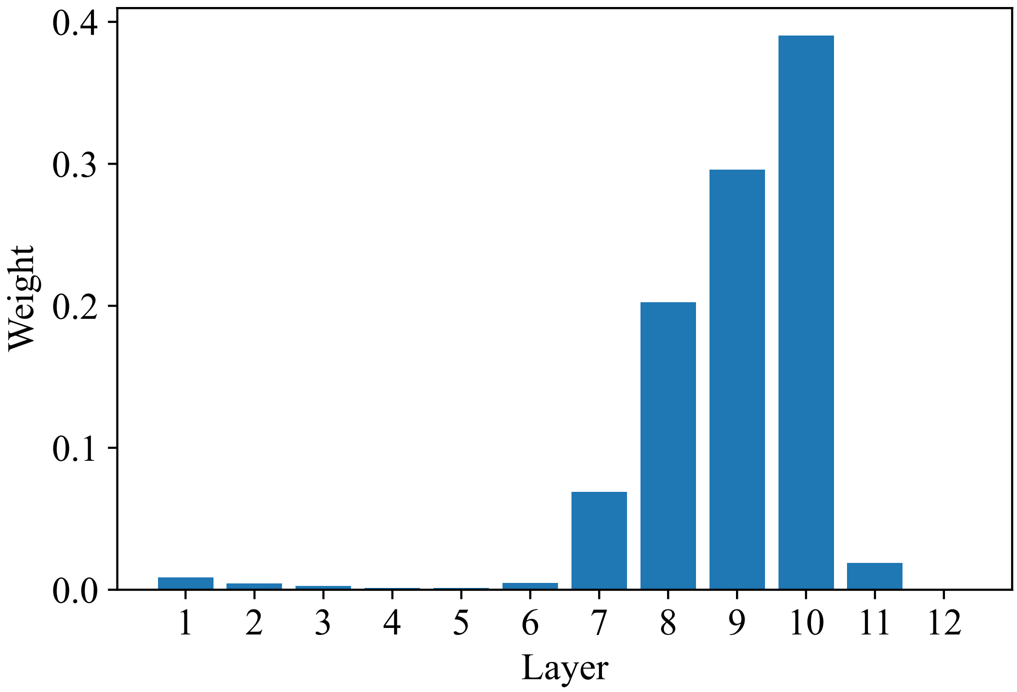

Figure 1. Weight analysis of the HuBERT model. The x-axis corresponds to each of the Transformer layers of the context network.

3.2. Innovation #2: Data Augmentation

Earlier S3RL models commonly relied on learning features from large unlabeled audiobook data, such as LibriSpeech [23] or LibriLight [24]. However, even though these models can learn fundamental characteristics from speech signals, real-world deployment data often involves diverse channel conditions and environmental noises that harm the system performance, a problem known as domain shift. To tackle this issue, data augmentation has been proposed as a technique to improve data diversity and reduce model bias [4,15,27]. Here, we adapt the RobustDistiller recipe proposed by [17], as shown in Figure 2. In particular, aiming to reduce computational requirements, we do not utilize the speech enhancement head present in the original work. At training time, given a batch of clean speech utterances, we sample one action to be applied to each utterance in the batch: (i) no changes are made to the training utterance; (ii) contaminate the utterance with either additive noise with a signal-to-noise ratio randomly chosen from $[0, 30]$ dB or convolve the speech waveform with a randomly selected room impulse response. The sampling probabilities of scenarios (i) and (ii) are 30% and 70%, respectively.

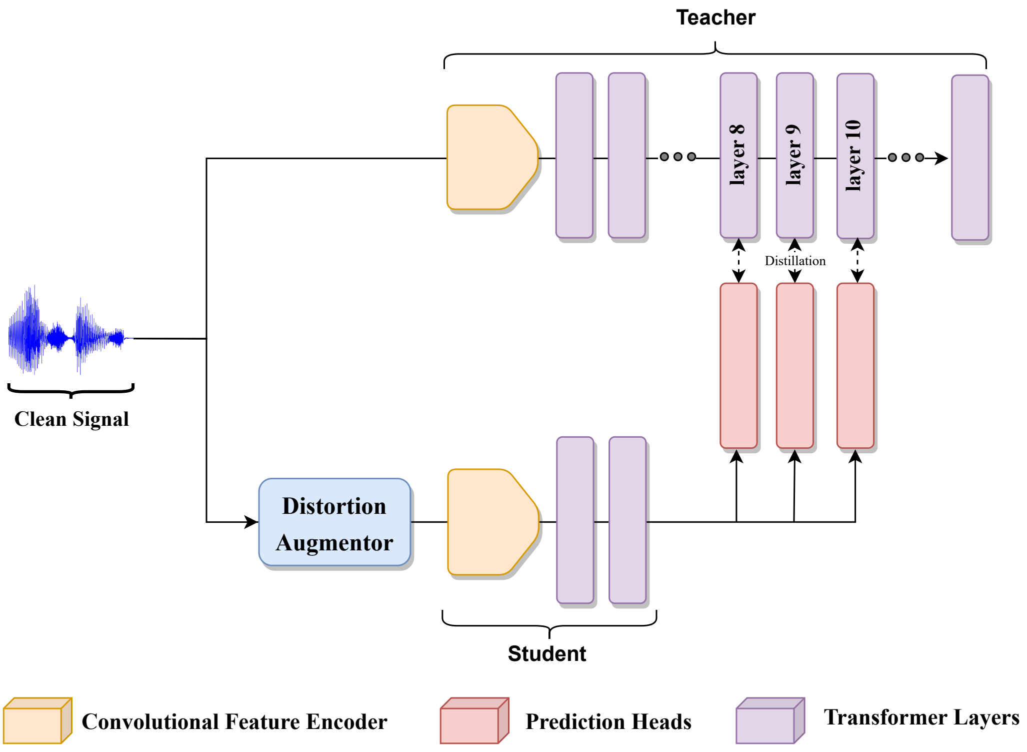

Figure 2. Block diagram of the adapted RobustDistiller pipeline without the enhancement head.

3.3. Innovation #3: Environment Awareness

Recent work has shown that data augmentation during the knowledge distillation process can improve the performance of compressed models in mismatched domain scenarios without compromising the model's size [18]. However, there are still limitations on how much robustness a model can acquire from the distillation process. The latest system described in [17], for example, showed some sensitivity to varying environmental conditions. Here, we propose a third innovation to overcome this issue, namely the use of environment awareness.

More specifically, different compressed models are obtained, each optimized for a distinct environmental condition (e.g., high signal to noise or highly reverberant room). During inference, the best model is selected and used for ASR. This hierarchical approach allows for each compressed model to act as an expert for a given environment scenario. While increasing the number of models used reduces the overall compression gains seen with distillation, here we explore the use of only two models, thus still achieving some compression relative to the original teacher model. Moreover, as only one model is used during inference, the gains in inference time are not affected by relying on two models. Figure 3a,b depicts the two models we explore here: one that characterizes the noise levels in the environment, and another that characterizes the reverberation levels, respectively.

As seen from the figures, for real-time inference, a noise/reverberation level selector is needed. As we assume that access to clean reference signals are not available, a "blind" measure is needed. Here, we explore with a signal-to-noise ratio (SNR) estimator called Waveform Amplitude Distribution Analysis (WADA), described in [28]. For the reverberation time (RT60) estimation, we rely on the speech-to-reverberation modulation energy ratio (SRMR) metric, followed by a mapping from the SRMR metric to the reverberation time (RT60) via a support vector regressor, as described in [29].

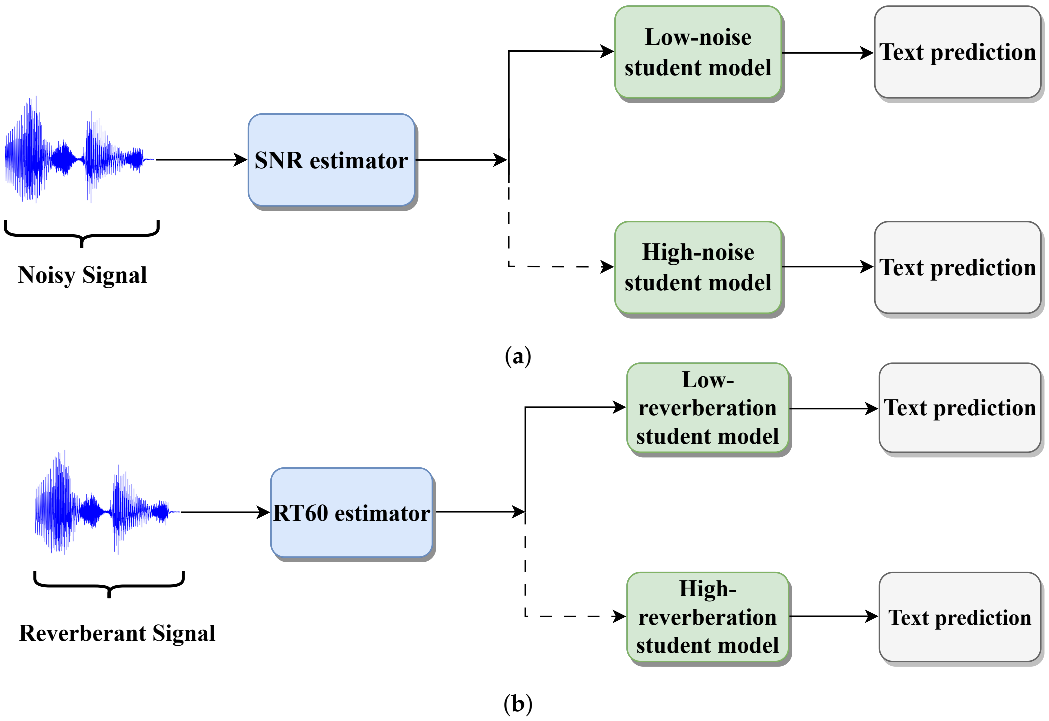

Figure 3. Diagram of the proposed Environment-Aware DistilHuBERT pipelines for (**a**) noisy and (**b**) reverberant environments.

4. Experimental Setup

In this section, we describe the databases used, the benchmark methods explored, figures-of-merit, as well as the blind noise and reverberation time estimators used.

4.1. Datasets

To train our models, we use the LibriSpeech corpus [23] as the dataset of clean speech utterances. It consists of 960 h of audiobook recordings with a 16 kHz sampling rate derived from the LibriVox project. As we are interested in performing data augmentation and evaluating environmental awareness, we use noise signals present in the MUSAN [25] and UrbanSound8K [30] datasets to corrupt the signals from the LibriSpeech dataset during the training stage. These datasets contain approximately 15 h of recordings in a wide variety of categories. UrbanSound8K contains 8732 labeled sound excerpts of urban sounds from 10 distinct classes, such as engine idling, car horn, and siren. We removed the children playing and street music categories to focus on non-speech-like noise sources in this analysis. Moreover, we use the noise portion of MUSAN, which contains 929 files of assorted noise types, including office-like noises, as well as ambient sounds, such as car idling, thunder, wind, footsteps, and rain. This portion of the dataset does not include recordings with intelligible speech. However, some recordings include crowd and babble noises with indistinct voices. All of the utterances are resampled to 16 kHz.

A room impulse response (RIR) dataset is also used, namely the Big Impulse Response Dataset (BIRD) [31], comprising simulated room impulse responses corresponding to rooms of various sizes and absorption coefficients, with RT60 values ranging from 140 ms to 1 s. The train set used for reverberation consists of approximately 35,000 simulated RIRs sampled from the BIRD dataset. Half of these samples have a lower reverberation time (RT60 smaller than 500 ms) and the other half have a higher reverberation time (RT60 greater than 500 ms).

At test time, combined with the LibriSpeech test set, two additional datasets are used to test the model performance under unseen conditions. The first is the Acoustic Scene Classification from the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge of 2020, namely DCASE2020 [32]. The dataset consists of 64 h of audio recordings in 10 acoustic scenes, recorded with four different recording devices in 12 different cities. For reverberation, we use a different subset of the BIRD dataset comprising approximately 4000 simulated room impulse responses. Again, the signals are equally split between low and high RT60 values.

4.2. Pre-Training

To gauge the benefits of our proposed methodology, different experiments are performed. First, we selected the HuBERT Base as our Teacher model for the distillation process. Then, we implement our proposed models. The model referred to as *Robust DistilHuBERT* is a proposed variation of the DistilHuBERT model with only the first two innovations being implemented and used to gauge the added benefits of noise awareness. This model is either robust to noise from 0 to 30 dB or to reverberation from 140 ms to 1 s. Meanwhile, our proposed model with all three modifications implemented is referred to as *Noise-Aware DistilHuBERT* or *Reverb-Aware DistilHuBERT*. They are composed of the pipeline depicted in Figure 3a and Figure 3b, respectively. Additionally, HuBERT Large [3] HuBERT Base [3], Robust HuBERT [15], DistilHuBERT [10], DistilRobustHuBERT, and RobustDistiller [17] are used as benchmark models.

Throughout all experiments, we use a single NVidia A100 GPU to train the upstream models and fine tune the downstream models. The execution time for both our robust distillation method and the fine-tuning step in the ASR downstream task is approximately 30 h each, for a total of 60 h. The training process utilizes the AdamW optimizer with a batch size of 24 utterances, for 200,000 iterations. After 14,000 updates, the learning rate decays from 2×10^{-4} to zero.

4.3. Evaluation Metric

In speech recognition, the word error rate (WER) is a common metric for evaluating the model performance [33]. The WER is defined as

$$WER = \frac{S+D+I}{N} \times 100 = \frac{S+D+I}{S+D+C} \times 100\,, \tag{1}$$

where S, D, I, and C are the number of substitutions, deletions, insertions, and correct words in the estimated sequence, respectively, and N is the number of words in the true sequence $(N = S + D + C)$. The WER is based on the Levenshtein distance and a smaller value signifies a closer approximation between the estimated word sequence and the ground truth transcription [34]. It is important to notice that, while the WER is usually presented as a value typically between 0 and 100, it does not represent a true percentage. While a WER of zero means a perfect estimation, this metric is not constrained to an upper bound, and a sequence with more insertions than correct words will have a WER greater than 100.

4.4. Noise Estimator

In this work, we utilize the WADA SNR estimator, an algorithm initially proposed by [28]. This algorithm provides a direct mapping from a signal to an estimated SNR value. It assumes that the amplitude distribution of clean speech can be approximated via the Gamma distribution with a shaping parameter of 0.4 and that the signal is corrupted with additive noise with Gaussian distribution. Even with these assumptions, experiments from the original authors show that the algorithm performs well when tested for three different types of noise: additive white Gaussian noise, musical segments from the DARPA HUB 4 Broadcast News database, and noise from a single interfering speaker. The noise signals were artificially added to the speech signals at different SNRs ranging from −10 dB to 30 dB. We chose to use this algorithm due to its computational efficiency and ease of use.

4.5. Reverberation Time Estimator

In order to estimate the reverberation time of a signal without knowing the properties of the room in which it was recorded, we follow the method described by [29,35]. The authors investigate the use of temporal dynamics' information for blind measurement of a room's acoustical parameters. Long-term dynamics' information is obtained by means of spectral analysis of temporal envelopes of speech, a process commonly termed as modulation spectrum processing. In our work, we rely on the speech-to-reverberation modulation energy ratio metric (SRMR) described in [29] with the source code available at https://github.com/jfsantos/SRMRpy (accessed on 18 October 2023). SRMR is inversely proportional to RT60, with lower values indicating higher reverberation levels. Additionally, we use a support vector machine (SVM) algorithm to provide a mapping between the inverse of the SRMR and the RT60.

5. Experimental Results and Discussion

In this section, we describe the obtained results and discuss them in light of the existing literature.

5.1. Accuracy of SNR and RT60 Estimators

First, we need to validate if the proposed noise and reverberation level estimators are accurate. To this end, clean speech samples taken from the dev-clean set of the LibriSpeech dataset were corrupted with additive noise randomly sampled from the DCASE dataset. The SNR is uniformly sampled from 0 to 30 dB. A total of 2800 noisy test files are used for this experiment. Figure 4 shows the scatterplot between estimated and true SNR values. An overall correlation of 82.5% is achieved. It is important to remember, however, that as mentioned previously, here we are employing two compressed models, one optimized on low SNR levels ranging from 0 to 10 dB and another for high SNR levels (10–30 dB). Using the WADA algorithm to detect signals within these two classes results in an overall accuracy of 94.2%, suggesting that the model can be accurate enough for deployment.

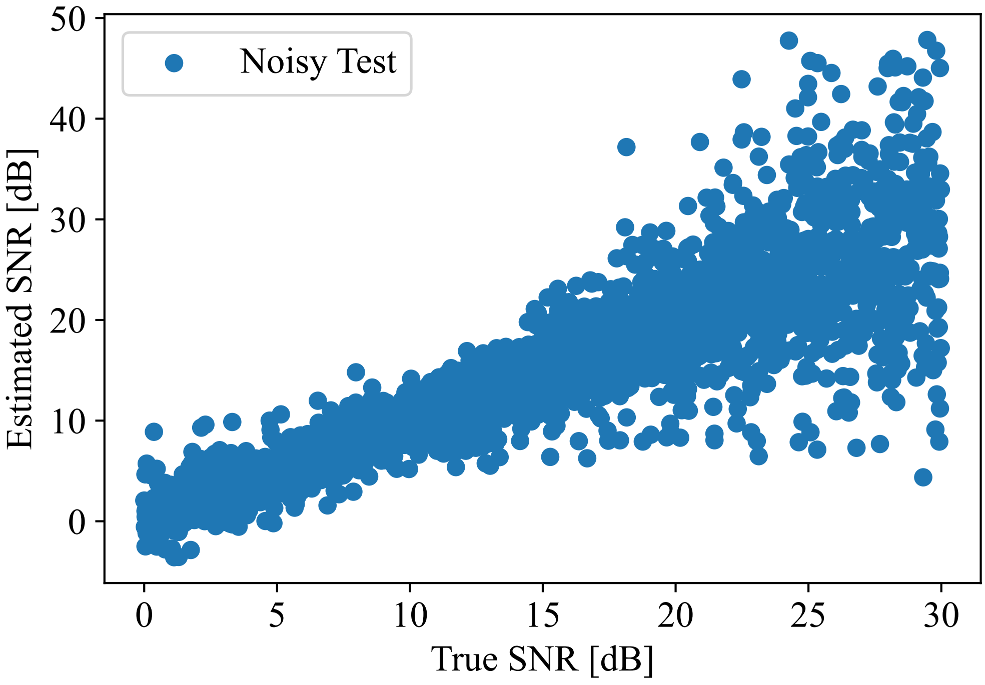

Figure 4. Scatterplot of estimated and true SNRs for noisy test signals using the WADA algorithm.

Figure 5, in turn, shows the scatterplot between the inverse of the SRMR and the true RT60. The signals used in this ablation study are speech samples taken again from the dev-clean set of the LibriSpeech dataset convolved with RIR signals sampled from the BIRD dataset. As can be seen, an overall correlation of 82.2% is achieved. Again, as the proposed solution utilizes models optimized on low reverberation (RT60 smaller than 500 ms) and high reverberation (RT60 grater than 500 ms) levels, only a binary classifier is needed. Unlike the previous experiment that maps the signal directly to the estimated quantity, this algorithm maps the signal to the inverse of the SRMR. Thus, we use an SVM to map between this quantity and the desired RT60. We perform a 90/10% train/test split on the data and use an SVM to perform the classification. Experiments using a vanilla SVM with a radial basis function (RBF) kernel classifier resulted in an 88.9% classification accuracy, indicating again that the model can be accurate enough for deployment.

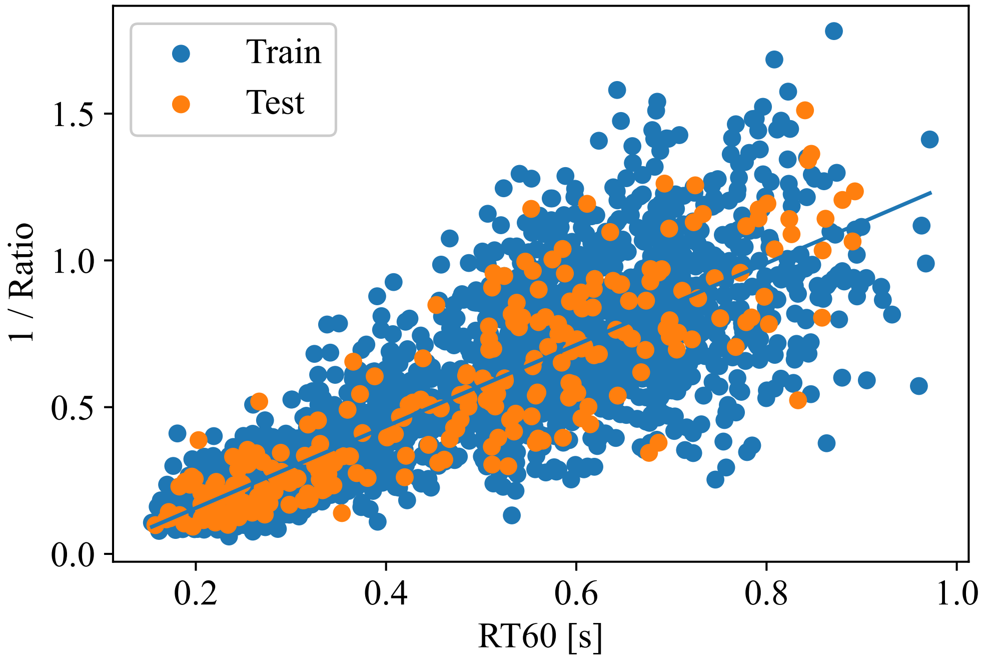

Figure 5. Scatterplot of inverse SRMR and true RT60.

5.2. Proposed System Performance

Table 1 compares the proposed methods to the six benchmark algorithms in terms of the number of model parameters, the number of multiply–accumulate operations (MACs), and the WER achieved on the clean and noisy test files. The noisy signals have been split into three categories: indoor, outdoor, and transportation noise types, where the SNR was randomly sampled between 0 and 30 dB. As can be seen, the first two innovations (row "Robust DistilHuBERT") already provide substantial improvement relative to the original DistilHubert model. Overall, relative gains of 33.04%, 31.16%, and 22.40% are obtained for the indoor, outdoor, and transportation noise types, respectively. By incorporating all three proposed innovations (row "Noise-aware DistilHuBERT"), a slight decrease in WER is achieved. It can also be observed that the proposed solutions result in compressed models that achieve similar WER across noise-type conditions, suggesting improved robustness to the ambient factors and applicability to the edge conditions.

To further explore the benefits of the proposed context-awareness solution, we next focus on the lower SNR conditions, known to be the most impactful for ASR. Table 2 further reports the achieved accuracy for the benchmarks and proposed solutions for only the noisy files corrupted by additive noise ranging from 0 to 10 dB SNR. As can be seen, the gains achieved with the proposed noise-aware solution outperforms the Robust DistilHuBERT model by 6.19%, 4.24%, and 4.62% for the indoor, outdoor, and transportation noise types, respectively. Relative to the original DistilHuBERT, these relative gains are of 48.42%, 46.42%, and 37.85%, respectively. It is important to emphasize that while the noise-aware solution does require the storage of double the number of parameters relative to DistilHuBERT and RobustDistiller, inference time and computation requirements remain the same, as only one of the two models is used at a time. As such, the achieved gains can still be useful for edge applications involving very noisy conditions.

Table 1. Performance comparison across different clean and noisy conditions with SNR between 0 and 30 dB.

Model	#Params (M)	MACs ($\times 10^9$)	Clean	Noise Type (WER)		
				Indoor	Outdoor	Transport
HuBERT Large [3]	300	4324	3.62	8.64	7.58	5.23
HuBERT Base [3]	95	1669	6.43	11.80	10.82	8.89
Robust HuBERT [15]	95	1669	6.75	9.38	8.85	8.04
DistilHuBERT [10]	24	785	13.29	26.21	23.94	19.64
DistilRobustHuBERT	24	785	12.70	24.41	21.92	18.36
RobustDistiller [17]	24	785	14.03	19.93	18.53	17.22
Robust DistilHuBERT	24	785	12.74	17.55	16.48	15.24
Noise-Aware DistilHuBERT	48	785	12.44	17.33	16.39	15.09

Table 2. Performance comparison across different noisy conditions with SNR between 0 and 10 dB.

Model	Noise Type (WER)		
	Indoor	Outdoor	Transport
HuBERT Large [3]	17.76	14.96	8.26
HuBERT Base [3]	20.77	18.36	12.95
Robust HuBERT [15]	14.64	12.37	9.40
DistilHuBERT [10]	45.56	39.64	29.59
DistilRobustHuBERT	41.80	36.35	27.04
RobustDistiller [17]	28.86	24.99	20.99
Robust DistilHuBERT	25.05	22.18	19.28
Noise-Aware DistilHuBERT	23.50	21.24	18.39

Lastly, Table 3 compares the WERs achieved for the benchmark and proposed solutions for speech signals under reverberant conditions. As can be seen, reverberation is a more challenging distortion to tackle and the WER of the benchmark distilled model is severely degraded. Implementing the first two innovations (Robust DistilHuBERT) allows the WER to be reduced by 66.52% overall and by 88.87% and 55.27% for the high RT60 and low RT60 conditions, respectively, relative to DistilHuBERT. The proposed environment-aware solution further decreases WER by an extra 2.55%,

2.76%, and 2.87%, respectively. Interestingly, the proposed solutions with only 24 M or 48 M parameters already significantly improve ASR accuracy relative to HuBERT Large with 300 M parameters. For the high reverberation level conditions, for example, the proposed environment-aware solution improves on HuBERT Large by 70.75% while requiring roughly one-sixth of the number of parameters.

The results reported above have relied on predicted SNR/RT60 estimates. We have also explored the use of an "oracle" system in which the true SNR/RT60 values are used (i.e., assuming perfect classification), but this did not result in any significant improvement, suggesting that the few errors made by the classifiers had minimal impact on the overall recognition accuracy. Overall, the obtained findings show the benefits of the proposed solutions for both noisy and reverberant settings, with the greatest gains seen in extremely low SNR conditions and highly reverberant settings, thus making the proposed models ideal for edge applications.

Table 3. Performance comparison across different clean and reverberant conditions with RT60 from 140 ms to 1 s.

Model	**Clean**	**Reverberation Time (WER)**		
		All	**High**	**Low**
HuBERT Large [3]	3.62	79.99	239.42	22.46
HuBERT Base [3]	6.43	77.15	145.59	36.02
Robust HuBERT [15]	6.75	58.36	89.13	30.29
DistilHuBERT [10]	13.29	156.98	648.02	74.04
DistilRobustHuBERT	12.70	77.14	97.48	59.61
RobustDistiller [17]	14.03	67.54	90.86	43.61
Robust DistilHuBERT	13.78	52.56	72.02	33.12
Reverb-Aware DistilHuBERT	13.21	51.22	70.03	32.17

5.3. To Distill or Not to Distill (A Robust Model)

Several works have proposed to make large speech models more robust to environmental factors via adversarial training [15], disentanglement [16], or data augmentation [4], to name a few methods. The results reported herein suggest, however, that applying conventional distillation methods to robust teacher models does not guarantee that the resulting compressed model will retain robustness to its fullest. Results from Table 1, for example, show the WER achieved from the DistilRobustHuBERT student model being roughly 2.5 times higher than that achieved by the original Robust HuBERT teacher. In the high-noise conditions, from Table 2, WER went up almost three times with the compressed student versions. While distilling from a noise-robust teacher (DistilRobustHuBERT) showed some improvement over distilling from an original teacher (DistilHuBERT), the obtained results are still far from those achieved with the proposed method. These findings suggest that adding robustness to the distillation process, as proposed herein, is more important than finding a robust latent representation in which distillation can be applied to.

5.4. Study Limitations

The proposed study is not without limitations. While the HuBERT model was used because several robust and compressed variants have been published in the literature recently (and thus can be used as benchmarks), more recently larger models, such as WavLM (with up to 300 M parameters) [4] or Whisper (up to 1.5 B parameters) [7], have been proposed and shown to be more robust to environmental noise. As such, the findings here are to be considered a lower bound on what could be achieved with environment awareness. Future work should explore the proposed distillation recipe with these emerging models. Moreover, here we have only tested the performance of the model on noise-only and reverberation-only conditions. In realistic settings, their combined effects may be present. In such scenarios, it may be possible to utilize the SNR and RT60 predictors together and see which distortion condition is most severe. This analysis is left for future work.

6. Conclusions

In this paper, environmental awareness is proposed as a method of improving the robustness of compressed universal representations to be used for speech recognition. In particular, we propose three innovations on top of the existing DistilHuBERT distillation recipe: (1) optimize the prediction heads, (2) employ a targeted data augmentation method for different environmental scenarios, and (3) employ a real-time SNR or RT60 estimator to choose the best compressed model for inference. We perform extensive experiments and compare against six benchmark models. When evaluated under noisy conditions (SNR between 0 and 30 dB), the proposed models outperform the benchmarks of comparable size (i.e., DistilHuBERT and RobustDistiller) by as much as 33.04%. The gains are more substantial in very noisy conditions (SNR between 0 and 10 dB), where gains of up to 48.42%

were seen as relative to DistilHuBERT. In turn, for high reverberation levels (RT60 greater than 500 ms), the proposed model was shown to outperform even the teacher models with a 2–6 times greater number of parameters (i.e., HuBERT base and large models) by as much as 89.19%. Lastly, we show the advantages of developing an environment-robust distillation process relative to just compressing robust latent representations from large speech models. Overall, these findings suggest the proposed method can be better suited for edge speech applications under varying environmental conditions.

Author Contributions: Conceptualization, A.P. and T.H.F.; methodology, all authors; software, A.P. and H.R.G.; validation, A.P. and H.R.G.; formal analysis, A.P.; resources, T.H.F.; writing, all authors; supervision, A.A. and T.H.F.; project administration, T.H.F.; funding acquisition, T.H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), grant number RGPIN-2021-03246.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used are publicly available. LibriSpeech available at https://www.openslr.org/12, MUSAN available at https://www.openslr.org/17/, UrbanSound8K available at https://urbansounddataset.weebly.com/urbansound8k.html, BIRD available at https://github.com/FrancoisGrondin/BIRD, DCASE2020 available at https://dcase.community/challenge2020/index (accessed on 19 November 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SOTA	State-of-the-art
WER	Word error rates
ASR	Automatic speech recognition
S3RL	Self-supervised speech representation learning
SNR	Signal-to-noise ratio
WADA	Waveform Amplitude Distribution Analysis
SRMR	speech-to-reverberation modulation energy ratio
BIRD	Big Impulse Response Dataset
DCASE	Detection and Classification of Acoustic Scenes and Events
RT60	Reverberation time
SVM	Support vector machine
RBF	Radial basis function
MAC	Multiply–accumulate

References

1. O'shaughnessy, D. *Speech Communications: Human and Machine (IEEE)*; Universities Press: New York, NY, USA, 1987.
2. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
3. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
4. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
5. Ao, J.; Wang, R.; Zhou, L.; Wang, C.; Ren, S.; Wu, Y.; Liu, S.; Ko, T.; Li, Q.; Zhang, Y.; et al. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. *arXiv* **2021**, arXiv:2110.07205.
6. Babu, A.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q.; Goyal, N.; Singh, K.; von Platen, P.; Saraf, Y.; Pino, J.; et al. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv* **2021**, arXiv:2111.09296.
7. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv* **2022**, arXiv:2212.04356.

8. Spille, C.; Kollmeier, B.; Meyer, B.T. Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Comput. Speech Lang.* **2018**, *52*, 123–140. [CrossRef]
9. Feng, T.H.; Dong, A.; Yeh, C.F.; Yang, S.w.; Lin, T.Q.; Shi, J.; Chang, K.W.; Huang, Z.; Wu, H.; Chang, X.; et al. Superb @ SLT 2022: Challenge on Generalization and Efficiency of Self-Supervised Speech Representation Learning. In Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 9–12 January 2023; pp. 1096–1103. [CrossRef]
10. Chang, H.J.; Yang, S.W.; Lee, H.Y. DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 7087–7091.
11. Kshirsagar, S.; Pendyala, A.; Falk, T.H. Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions. *Front. Comput. Sci.* **2023**, *5*, 1039261. [CrossRef]
12. Zhang, P.; Huang, Y.; Yang, C.; Jiang, W. Estimate the noise effect on automatic speech recognition accuracy for mandarin by an approach associating articulation index. *Appl. Acoust.* **2023**, *203*, 109217. [CrossRef]
13. Pimentel, A.; Guimarães, H.; Avila, A.R.; Rezagholizadeh, M.; Falk, T.H. On the Impact of Quantization and Pruning of Self-Supervised Speech Models for Downstream Speech Recognition Tasks "In-the-Wild". *arXiv* **2023**, arXiv:2309.14462.
14. Li, S.; Yerebakan, M.O.; Luo, Y.; Amaba, B.; Swope, W.; Hu, B. The Effect of Different Occupational Background Noises on Voice Recognition Accuracy. *J. Comput. Inf. Sci. Eng.* **2022**, *22*, 050905. [CrossRef]
15. Huang, K.P.; Fu, Y.K.; Zhang, Y.; Lee, H.Y. Improving Distortion Robustness of Self-supervised Speech Processing Tasks with Domain Adaptation. *Proc. Interspeech* **2022**, 2193–2197. [CrossRef]
16. Ng, D.; Zhang, R.; Yip, J.Q.; Yang, Z.; Ni, J.; Zhang, C.; Ma, Y.; Ni, C.; Chng, E.S.; Ma, B. De'hubert: Disentangling Noise in a Self-Supervised Model for Robust Speech Recognition. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–10 June 2023; pp. 1–5. [CrossRef]
17. Guimarães, H.R.; Pimentel, A.; Avila, A.R.; Rezagholizadeh, M.; Chen, B.; Falk, T.H. Robustdistiller: Compressing Universal Speech Representations for Enhanced Environment Robustness. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–10 June 2023; pp. 1–5. [CrossRef]
18. Huang, K.P.; Fu, Y.K.; Hsu, T.Y.; Gutierrez, F.R.; Wang, F.L.; Tseng, L.H.; Zhang, Y.; Lee, H.y. Improving Generalizability of Distilled Self-Supervised Speech Processing Models Under Distorted Settings. In Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 9–12 January 2023; pp. 1112–1119. [CrossRef]
19. Ericsson, L.; Gouk, H.; Loy, C.C.; Hospedales, T.M. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Process. Mag.* **2022**, *39*, 42–62. [CrossRef]
20. Mohamed, A.; Lee, H.y.; Borgholt, L.; Havtorn, J.D.; Edin, J.; Igel, C.; Kirchhoff, K.; Li, S.W.; Livescu, K.; Maaløe, L.; et al. Self-Supervised Speech Representation Learning: A Review. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1179–1210. [CrossRef]
21. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
23. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Queensland, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 5206–5210.
24. Kahn, J.; Rivière, M.; Zheng, W.; Kharitonov, E.; Xu, Q.; Mazaré, P.E.; Karadayi, J.; Liptchinsky, V.; Collobert, R.; Fuegen, C.; et al. Libri-light: A benchmark for ASR with limited or no supervision. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7669–7673.
25. Snyder, D.; Chen, G.; Povey, D. MUSAN: A Music, Speech, and Noise Corpus. *arXiv* **2015**, arXiv:1510.08484.
26. Yang, S.-W.; Chi, P.H.; Chuang, Y.S.; Lai, C.I.J.; Lakhotia, K.; Lin, Y.Y.; Liu, A.T.; Shi, J.; Chang, X.; Lin, G.T.; et al. Superb: Speech Processing Universal performance Benchmark. *Proc. Interspeech* **2021**, 1194–1198.
27. Hsu, W.N.; Sriram, A.; Baevski, A.; Likhomanenko, T.; Xu, Q.; Pratap, V.; Kahn, J.; Lee, A.; Collobert, R.; Synnaeve, G.; et al. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv* **2021**, arXiv:2104.01027.
28. Kim, C.; Stern, R. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In Proceedings of the Ninth Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008; pp. 2598–2601. [CrossRef]
29. Falk, T.H.; Chan, W.Y. Temporal Dynamics for Blind Measurement of Room Acoustical Parameters. *IEEE Trans. Instrum. Meas.* **2010**, *59*, 978–989. [CrossRef]
30. Salamon, J.; Jacoby, C.; Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. In Proceedings of the 22nd ACM International Conference on Multimedia (ACM-MM'14), Orlando, FL, USA, 3 November 2014; pp. 1041–1044.
31. Grondin, F.; Lauzon, J.S.; Michaud, S.; Ravanelli, M.; Michaud, F. BIRD: Big Impulse Response Dataset. *arXiv* **2020**, arXiv:2010.09930.
32. Mesaros, A.; Heittola, T.; Virtanen, T. A multi-device dataset for urban acoustic scene classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey, UK, 19–20 November 2018; pp. 9–13.
33. Huang, X.; Baker, J.; Reddy, R. A Historical Perspective of Speech Recognition. *Commun. ACM* **2014**, *57*, 94–103. [CrossRef]

34. von Neumann, T.; Boeddeker, C.; Kinoshita, K.; Delcroix, M.; Haeb-Umbach, R. On Word Error Rate Definitions and Their Efficient Computation for Multi-Speaker Speech Recognition Systems. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–10 June 2023; pp. 1–5. [CrossRef]
35. Falk, T.; Yuan, H.; Chan, W.Y. Spectro-temporal processing for blind estimation of reverberation time and single-ended quality measurement of reverberant speech. In Proceedings of the Eighth Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 27–31 August 2007; Volume 2; pp. 514–517. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.