uDistil-Whisper: Label-Free Data Filtering for Knowledge Distillation in Low-Data Regimes

Abdul Waheed*[1], Karima Kadaoui†[2], Bhiksha Raj‡[1,2], and Muhammad Abdul-Mageed§[3,2,4]

[1]Carnegie Mellon University
[2]MBZUAI
[3]The University of British Columbia
[4]Invertible AI

Abstract

Recent work on distilling Whisper's knowledge into small models using pseudo-labels shows promising performance while reducing the size by up to 50%. This results in small, efficient, and dedicated models. However, a critical step of distillation using pseudo-labels involves filtering high-quality predictions and using only those during training. This step requires ground truth labels to compare with and filter low-quality examples, making the process dependent on human labels. Additionally, the distillation process requires a large amount of data thereby limiting its applicability in low-resource settings. To address this, we propose a distillation framework that does not require *any* labeled data. Through experimentation, we show that our best-distilled models outperform the teacher model by 5-7 WER points and are on par with or outperform similar supervised data filtering setups. When scaling the data, our models significantly outperform all zero-shot and supervised models. Our models are also 25-50% more compute- and memory-efficient while maintaining performance equal to or better than that of the teacher model. For more details about our models, dataset, and other resources, please visit our GitHub page: https://github.com/UBC-NLP/uDistilWhisper.

1 Introduction

To democratize automatic speech recognition (ASR), significant attention has been given to multilingual models (Pratap et al., 2023; Radford et al., 2023; Communication et al., 2023; Zhang et al., 2023). These powerful systems, thanks to their massive number of parameters and training data, can simultaneously transcribe in hundreds of languages. That being said, low-resource languages tend to trail behind in performance compared to high-resource counterparts such as English (Radford et al., 2023). For instance, while OpenAI's latest ASR model Whisper-large-v3 shows the word error rate (WER) in the single digits on English and Spanish, a much lower performance is shown for East Asian and African languages[1] (Talafha et al., 2023, 2024). While the performance reported on these multilingual models is not as low for Arabic, it is important to highlight that it does not fully and accurately represent the diversity of the different varieties that characterize the Arabic language family (Abdul-Mageed et al., 2020). In addition to the disparity in performance, multilingual models are extremely compute-intensive and are thus not equally accessible to everyone. In order to alleviate this issue, knowledge distillation has been the method of choice for multiple works (Hinton et al., 2015; Yang et al., 2023; Sanh et al., 2020; Frantar et al., 2022).

Knowledge distillation has proven to be highly effective in reducing model size, thereby lowering compute and memory requirements while maintaining performance comparable to larger teacher models. For example, Gandhi et al. (2023) and Waheed et al. (2024) demonstrate its effectiveness for English and Arabic, respectively. However, a significant limitation in these works is their reliance on ground-truth labels to filter out low-quality pseudo-labels generated by the teacher model—a resource that is often scarce, particularly for low-resource languages. Additionally, Waheed et al. (2024) shows that training on unfiltered data leads to suboptimal performance. This reliance on labeled data underscores the need for an unsupervised approach to data filtering in knowledge distillation, enabling distilled models to perform better without depend-

*abdulw@cs.cmu.edu
†karima.kadaoui@mbzuai.ac.ae
‡bhiksha@cs.cmu.edu
§muhammad.mageed@ubc.ca

[1]https://github.com/openai/whisper/discussions/1762

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5750–5767
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

ing on any ground-truth labels.

In this work, we propose frameworks to distill knowledge from teacher models that do not rely on filtered pseudo-labels. Our method achieves a similar performance to supervised distillation and outperforms pseudo-label distillation without filtering. Our contributions are the following:

- We explore various unsupervised methods to filter out low-quality pseudo-labels, eliminating the need of labeled data for distillation.
- We assess the performance of our distilled models, which effectively outperform the teacher model by 5–7 WER points, and extend the existing setup to include new datasets in both Arabic and Swahili.
- We analyze the effectiveness of our best metrics in detecting low-quality pseudo-labeled examples, achieving an AUC ranging between 0.77 and 0.82 for examples with WER > 20.

We organize the paper as follows: we present a review of the existing literature in Section 2 before introducing knowledge distillation along with our methodology in Section 3. Section 4 contains particulars of our experiments and Section 5 delves into the results we obtain. We discuss these results in Section 5, conclude in Section 6, and outline our work's limitations in Section 7.

2 Related Work

Scaling speech foundation models in both data and model size has led to systems capable of handling a wide range of speech tasks (Pratap et al., 2023; Zhang et al., 2023; Radford et al., 2023; Communication et al., 2023). These models effectively transcend language barriers, demonstrating robust performance across numerous languages, including many low-resource ones (Hsu et al., 2024; Yeo et al., 2024). However, these models face two key limitations: (1) their large size (Radford et al., 2023; Communication et al., 2023), which results in high computational costs, and (2) the under-representation of low-resource languages in training, leading to suboptimal performance on these languages (Radford et al., 2023; Talafha et al., 2023). To address the first issue various efficient decoding methods (SYSTRAN; Segal-Feldman et al., 2024; Malard et al., 2023; Leviathan et al., 2023) were proposed. However, resulting models do not necessarily enjoy a reduced memory footprint.

To address the efficiency issue, knowledge distillation (Hinton et al., 2015; Gou et al., 2021) is proven to be very effective. Different approaches for knowledge distillation have been explored such as Patient Knowledge Distillation (PKD) (Sun et al., 2019), T-S learning (Manohar et al., 2018), Cross-Modal Hashing (CMH) (Hu et al., 2020), Joint Unsupervised Domain Adaptation with KD (Zhang et al., 2021) and Data-Free KD (Lopes et al., 2017) among others in the NLP and CV fields alike. Knowledge distillation has also been applied to speech recognition: Shao et al. (2023) shrink a Whisper (Radford et al., 2023) model tangibly ($\sim$ 80% of its original size), all while improving its performance; Chang et al. (2022) reduce a Hubert model to 75% of its initial size through layer-wise distillation without a significant drop in performance.

In addition, knowledge distillation has found successful applications in speech-to-text tasks as well, (Nayem et al., 2023; Hentschel et al., 2024; Tian et al., 2022) effectively reducing the memory and compute requirements. More recently, it has been used to distill multilingual models from Whisper (Ferraz et al., 2024) and strong monolingual models (Gandhi et al., 2023), as well as in low-resource settings. An example of the latter is Waheed et al. (2024) who investigate pseudo-label distillation methods (Gandhi et al., 2023) across various Modern Standard Arabic (MSA) and dialectal Arabic datasets. Their exploration demonstrate the efficacy of distillation in enhancing both efficiency and performance, showcasing that smaller dedicated models can outperform larger multilingual ones. While the pseudo-labeling approach has proven effective across diverse languages (Gandhi et al., 2023; Waheed et al., 2024), it traditionally requires labeled data for filtering low-quality pseudo-labeled examples. In this work we address this limitation by introducing an *unsupervised* framework based on data filtering methods, thereby eliminating the dependency on labeled data altogether.

3 Methodology

3.1 Knowledge Distillation

Knowledge distillation is a framework through which a small student model learns the behavior of a bigger teacher model (Hinton et al., 2015; Sanh et al., 2020; Kim and Rush, 2016). Gandhi et al. (2023) introduce knowledge distillation via pseudo-labeling, which generates (English) predictions from a teacher Whisper model and filters them through a WER threshold to only keep the most accurate ones. These pseudo-labels are then used to train a smaller student model. Waheed et al. (2024) build on this approach and show that a Whisper student model beats its teacher model in its average performance in an Arabic multi-dialectal setting. We follow this standard student-teacher framework to distill Whisper into small yet powerful models. The objective of the distillation process can be stated as:

$$\mathcal{L}_{KD} = \alpha_{KL}\mathcal{L}_{KL} + \alpha_{PL}\mathcal{L}_{PL} \tag{1}$$

where:

- $\mathcal{L}_{KL}$ is the Kullback-Leibler (KL) divergence loss, encouraging the student model to match the teacher's probability distribution.
- $\mathcal{L}_{PL}$ trains the student using pseudo-labels as ground truth.

The coefficients α_{KL} (0.8) and α_{PL} (1.0) balance the contributions of each loss in the overall distillation loss $\mathcal{L}_{KD}$.

3.2 Label-Free Data Filtering

Previous data filtering methods, such as those used by Gandhi et al. (2023) and Waheed et al. (2024), rely on ground-truth labels by computing the WER between teacher-generated pseudo-labels and reference transcripts, discarding examples with high error rates. However, this dependence on labeled data limits their use in low-resource languages where such labels are scarce.

Our approach introduces label-free filtering methods that assess pseudo-label quality using the teacher model's logits, synthetic speech, proxy models, and multimodal embeddings. Each method is detailed below.

3.2.1 Proxy Models

We use a pre-trained ASR model, SeamlessM4T-large-v2 (Communication et al., 2023), as a proxy to generate reference transcripts for the input speech. Then, the quality of pseudo-labels generated by the teacher model is evaluated by calculating the WER between the proxy's and the teacher's outputs, referred to as pWER. Lower pWER values indicate higher agreement and examples exceeding a defined pWER threshold are discarded.

3.2.2 Uncertainty Quantification

We leverage uncertainty in the teacher's output to filter low-quality examples, using two common metrics:

Entropy. Entropy measures uncertainty in the teacher's predicted probability distribution. High entropy suggests low confidence. It is calculated as:

$$H = -\sum_{i=1}^{N} p_i \log_2(p_i) \tag{2}$$

where p_i is the predicted probability for the i^{th} word.

Geometric Mean of Confidence Scores. The geometric mean of the confidence scores, representing the probability of each decoded token in the teacher-generated pseudo-label, is used to assess the overall confidence in the pseudo-label. It is calculated as:

$$G = \sqrt[N]{\prod_{i=1}^{N} c_i} \tag{3}$$

where c_i is the confidence score for each word. Low entropy and high geometric mean indicate better pseudo-labels.

3.2.3 Negative Log-Likelihood

We compute the negative log-likelihood (NLL) using AceGPT-7B (Huang et al., 2024) to assess the quality of pseudo-labels. Lower NLL suggests higher pseudo-label accuracy. It is calculated as:

$$NLL = -\sum_{t=1}^{T} \log(p(y_t | y_{1\ldots t-1})) \tag{4}$$

where $p(y_t|y_{1\ldots t-1})$ is the probability assigned to the t^{th} word.

3.2.4 Multimodal Embeddings

We use SONAR (Duquenne et al., 2023) to generate embeddings for input speech and pseudo-labels. The dot product is used to compute similarity, with higher scores indicating better alignment and pseudo-label quality.

3.2.5 Perceptual Evaluation of Speech Quality (PESQ)

Synthetic speech is generated from pseudo-labels using XTTS-v2[2] and compared with input speech. Similarity is assessed via PESQ, where high scores indicate higher label accuracy.

[2] https://huggingface.co/coqui/XTTS-v2

3.3 Training Data

We follow Waheed et al. (2024) for the data mixture to train our models. As a result, we randomly select segments from a diverse set of datasets, including MGB2 (Ali et al., 2016), MGB3 (Ali et al., 2017), FLEURS (Conneau et al., 2023), CommonVoice 15.0 (Ardila et al., 2020), QASR (Mubarak et al., 2021), the Arabic Speech Corpus (Halabi, 2016), and the Massive Arabic Speech Corpus (MASC) (Al-Fetyani et al., 2023). Specifically, we sample 100K and 500K segments, equivalent to approximately 100 and 500 hours of pseudo-labeled speech data, respectively. Our dataset compilation strictly includes the train splits only from each source. We filter out roughly 27% low quality examples from our training data using the metrics described in Section 3.

3.4 Teacher and Student Models

In line with previous work (Gandhi et al., 2023; Waheed et al., 2024), we utilize the *whisper-large-v2* checkpoint[3] for both pseudo-labeling and as the teacher model during training. We train two variants of students, differing in the number of layers removed from the teacher model. Following Gandhi et al. (2023) and Waheed et al. (2024), we initialize the student models with maximally spaced layers in the encoder and decoder block of the teacher model. When training on 100K segments, we refer to the models with 16-16 and 32-16 encoder-decoder blocks as *UDW-16-16*[4] and *UDW-32-16*, respectively. Similarly, we refer to the models trained on 500K segments as *UDW-16-16++* and *UDW-32-16++*.

4 Experiments

For a fair comparison and analysis, we keep our experimental setup quite identical to Waheed et al. (2024). Apart from evaluating our models on five standard benchmark datasets, we extend the evaluation to include the SADA (Alharbi et al., 2024) and Casablanca[5] (Talafha et al., 2024) datasets. We describe all the used datasets below.

[3] `https://huggingface.co/openai/whisper-large-v2`

[4] Following the format from Waheed et al. (2024), UDW refers to Unsupervised Distill Whisper.

[5] The In-House data in Waheed et al. (2024), reported in Table 1, is an earlier subset of the Casablanca dataset, used prior to the official release of the larger version.

4.1 Evaluation Dataset

FLEURS. The Few-shot Learning Evaluation of Universal Representations of Speech (FLEURS) (Conneau et al., 2023) is a multilingual speech dataset with 102 languages, each with 12 hours of speech. It supports tasks like ASR, language identification, and translation. We use the Arabic subset, which features MSA speech with Egyptian accents.

Common Voice. Common Voice (CV) (Ardila et al., 2020) is a volunteer-driven multilingual dataset with 124 languages and 31,176 hours of speech. The Arabic subset includes 156 hours of primarily MSA speech.

Multi-Genre Broadcast. The MGB dataset (Ali et al., 2016, 2017, 2019) consists of MGB2 (1,200 hours of MSA-dominant Aljazeera Arabic speech), MGB3 (six hours of Egyptian dialectal speech), and MGB5 (14 hours of Moroccan dialectal speech).

SADA. The Saudi Audio Dataset for Arabic (SADA) (Alharbi et al., 2024) includes 668 hours of Arabic speech (435 labeled), featuring Saudi dialects, MSA, and speech from the Levant, Yemen, and Egypt. Statistics are provided in Appendix B 2.1.

Casablanca. Casablanca (Talafha et al., 2024) is a dialectal Arabic dataset that contains 48 hours of speech from various TV shows. It covers eight different dialects and is annotated by native speakers. The results on this dataset are separately reported in Table 4.

4.2 Baselines

We employ several monolingual and multilingual speech recognition models on different Arabic varieties, including standard and accented MSA, as well as various dialects. The models are grouped as follows:

4.2.1 Supervised Models and Commercial Systems.

We evaluate three baselines, namely Wav2Vec2-XLS-R (Conneau et al., 2020; Babu et al., 2022) trained on CV8.0, HuBERT (Hsu et al., 2021) trained on both MGB-3 (Ali et al., 2017) and a 5.5-hour Egyptian Arabic corpus, and a fine-tuned Whisper-large-v2 model trained on CV11.0 and MGB-2. We use these models off-the-shelf from their respective checkpoints.

4.2.2 Zero-Shot Models

Whisper. Whisper (Radford et al., 2023) is a versatile multilingual speech model designed for both speech recognition and translation, including Arabic. We assess four Whisper variants: *whisper-small* (W-S), *whisper-medium* (W-M), *whisper-large-v2* (W-L-v2), and *whisper-large-v3* (W-L-v3), using their default decoding parameters with a maximum sequence length of 225 tokens.

SeamlessM4T. SeamlessM4T models excel at generating high-quality transcripts across languages (Communication et al., 2023), but evaluations on them typically focus on English. To bridge this gap, we test three of its versions – *seamless-m4t-medium* (SM4T-M), *seamless-m4t-v1-large* (SM4T-L-v1), and *seamless-m4t-v2-large* (SM4T-L-v2) – under a zero-shot setting on Arabic ASR, using the default parameters from the model's inference pipeline.

4.2.3 Distilled Models

To compare our results with an equivalent supervised setup, two distilled model variants of different sizes are evaluated. We extend the evaluation to SADA's *test* and *valid* splits, using 16-16 and 32-16 models trained with a WER threshold of 80% on both 100K and 500K segments. Additionally, we include models trained without filtering as a lower bound. For a fair comparison with supervised data filtering, we follow an experimental setup identical to prior work (Waheed et al., 2024).

4.3 Experimental Setup

During evaluations, we adhere to the default decoding parameters specified by the original model implementations unless specified otherwise. Throughout our experiments, a maximum sequence length of 225 is maintained and WER and CER are used as the primary metrics for the quantitative evaluation of the models.

For distillation, roughly 27% of the top segments are selected based on the metrics, to match the amount of data used in a supervised setup when $\lambda = 80\%$. We do not conduct hyperparameter search due to computational constraints, and instead apply the configuration detailed in Gandhi et al. (2023). The distillation process and its key parameters are detailed in Table 13 in Appendix C.

5 Results and Discussion

We report the baseline results from Waheed et al. (2024) and evaluate all models on SADA's test and validation splits. Our main results can be found in Table 1, and the average across different evaluation setups in Table 2. Results on the top five varieties[6] of SADA are provided in Table 9. The reported results are computed on normalized text that includes removing diacritics. Results on the text with no normalization can be found in Appendix 3.3.

While larger variants of both SeamlessM4T and Whisper perform well on standard benchmark setups like Common Voice and Fleurs, we find that they perform poorly on dialectal speech. For example, the best zero-shot model (SeamlessM4T-large-v2) has a 7.6% WER on Fleurs compared to 65.0% on the SADA test set as shown in Table 1. This underscores the inadequate evaluation and limited generalization capability of zero-shot models in unseen and challenging settings. The best-distilled model DW-32-16++ outperforms all other models including its teacher (large-v2), the best Whisper version (large-v3), and the best baseline (SeamlessM4T-large-v2).

When it comes to our distilled models, we observe that the two best filtering measures, *sonar-sim* and *proxy-ref*, are comparable to or surpass the equivalent supervised setups. For instance, DW-32-16 trained on 100K examples with a filtering threshold of 80 achieves a WER of 35.3% on the benchmark test split, 66.6% on the SADA test split, and 61.1% averaged on five novel dialects. By comparison, our best-performing model in a similar configuration, UDW-32-16-sonar, achieves 35.5% WER on the benchmark test split, 58.7% on the SADA test, and 60.6% on five dialects. This reflects close to 3% improvement in average score compared to the supervised setup (51.6 vs 54.3) and more than a 7% improvement over the no-filter setup, as detailed in Table 2. In addition to that, our experiments show that the smaller student model (UDW-16-16) performs better with the *proxy-ref* method. More specifically, when trained on 100K examples, it matches the results of supervised setups (58.5% vs 57.6%) and shows over 6% improvement compared to setups where pseudo-labels are not filtered (64.2%).

Scaling the Data. We scale our dataset from 100K to 500K segments and we train two models, UDW-16-16 and UDW-32-16, employing *proxy-ref* and *sonar-sim* as filtering criteria. As shown in Table 2, our UDW-32-16++ model, trained with 500K seg-

[6] We refer labels in SADA as varieties as not all labels are necessarily dialects.

Model	Size	CV15.0	MGB2	MGB3	MGB5	Fleurs	In-house Data					SADA
							ALG	JOR	PAL	UAE	YEM	
Baselines												
Amazon	–	–	–	–	–	–	83.6	45.5	52.4	58.8	64.7	–
XLS-R	0.96	89.7	97.6	98.7	99.5	94.9	99.7	99.1	99.1	99.4	99.5	99.5
HuBERT	0.31	55.2	49.6	**25.2**	92.4	34.9	96.8	65.2	73.8	83.0	90.5	75.6
W-FT	1.5	35.8	**15.3**	48.9	101.4	9.8	115.5	67.8	69.6	105.9	107.1	92.3
MMS-all	1.0	106.4	39.3	75.3	89.7	23.8	100.2	89.8	99.9	100.1	100.2	77.6
SM4T-M	1.2	16.3	19.5	41.4	83.8	8.7	81.1	46.3	55.2	59.8	68.9	65.2
SM4T-L-v1	2.3	19.8	21.8	44.4	89.9	11.1	87.9	50.7	57.5	61.8	72.2	64.9
SM4T-L-v2	2.3	**11.3**	17.3	36.2	89.1	**7.6**	92.1	**41.5**	49.5	55.9	69.7	65.0
W-S	0.24	40.3	46.8	81.4	226.5	28.2	130.7	68.6	73.8	97.8	107.1	139.9
W-M	0.77	29.8	33.1	64.3	127.7	16.4	103.7	50.5	58.7	82.5	86.8	99.3
W-L-v2	1.5	19.6	26.5	53.0	99.2	11.4	106.4	42.3	51.1	63.8	77.3	69.8
W-L-v3	1.5	15.8	15.9	35.7	79.8	9.7	101.9	43.6	53.4	63.4	76.1	67.2
DW-16-16	0.80	22.1	26.0	50.5	82.4	18.8	83.0	50.4	61.0	64.6	72.7	66.7
DW-32-16	1.12	18.8	21.1	43.8	78.9	14.2	79.5	44.4	55.0	58.1	68.5	72.3
DW-16-16++	0.80	19.2	23.0	47.2	79.0	15.0	79.0	46.7	56.4	60.4	69.1	69.9
DW-32-16++	1.12	17.1	19.7	40.7	**76.6**	11.1	74.6	41.6	51.4	53.5	**63.5**	60.3
No-filter												
– DW-16-16	0.80	22.8	26.1	54.1	95.1	17.6	93.4	51.8	64.7	68.0	78.3	78.4
– DW-32-16	1.12	21.2	22.8	51.3	90.5	14.9	87.3	47.6	57.1	63.9	73.0	70.7
Ours												
UDW-16-16	0.80											
– nll		23.6	26.4	54.4	92.3	18.5	82.4	63.3	53.0	72.0	93.4	84.3
– pesq		24.3	28.3	54.2	93.7	19.7	81.5	64.9	52.9	69.7	87.0	79.7
– entropy		23.8	27.4	57.3	94.8	18.2	90.1	54.8	64.3	73.7	96.4	92.0
– conf		23.5	27.2	52.7	87.8	17.7	89.2	53.1	63.4	69.0	79.1	78.6
– proxy		22.5	25.5	49.4	84.8	17.6	74.5	61.1	50.5	65.4	84.1	68.6
– sonar		24.1	28.3	55.1	85.6	20.4	86.9	56.4	65.4	70.7	76.5	74.9
UDW-32-16	1.12											
– nll		18.9	22.6	48.1	94.1	13.3	75.4	55.2	44.7	63.7	83.8	73.4
– pesq		21.7	24.2	49.0	87.0	16.1	88.8	47.5	58.5	67.0	75.5	73.3
– entropy		19.1	21.8	46.9	87.1	13.3	91.0	46.7	58.2	64.9	76.4	68.3
– conf		20.3	22.8	46.6	83.4	14.5	**73.4**	57.3	61.1	**47.5**	85.7	66.2
– proxy		19.0	21.5	44.3	82.6	14.2	80.4	45.5	56.0	61.6	69.5	64.6
– sonar		17.8	21.1	45.3	80.4	13.1	79.0	44.3	54.3	58.8	66.8	58.7
UDW-16-16++	0.80											
– proxy		15.5	23.3	46.8	84.6	14.2	82.9	46.4	57.0	61.1	70.8	60.2
– sonar		18.9	22.8	47.2	82.0	14.9	81.5	47.7	57.4	62.1	69.2	59.4
UDW-32-16-++	1.12											
– proxy		18.1	20.8	43.1	82.3	12.6	80.8	43.1	54.4	57.7	68.5	**55.1**
– sonar		17.0	21.9	44.1	79.8	12.8	78.2	54.2	**44.4**	58.2	66.6	61.6

Table 1: WER (↓) after normalization and removing diacritics. All baseline distilled models (dw-) are trained with a filtering threshold of 80 if not specified. Best results are shown in **bold**. Second best results are underlined. We report the score on the test split of each dataset. Abbreviations. **W** - Whisper, **FT** - Finetuned, **M** - Medium, **L** - Large, **S** - Small, **DW** - Distill Whisper, **UDW** - Unsupervised Distill Whisper, **nll** - negative log likelihood, **conf** - confidence score.

ments, performs comparably with its supervised counterpart DW-32-16++. For instance, UDW-32-16++ with *proxy-ref* filtering achieves a WER of 51.4%, similar to DW-32-16++'s 50.0%. Notably, our smaller student model (UDW-16-16++) trained on 500K segments results in superior performance compared to models trained under supervised filtering. For example, UDW-16-16++ with *sonar-sim* gives 53.9% WER compared to 56.2 for DW-16-16++. These results showcase the efficacy of our data filtering methods.

Generalization to Unseen Dialects. We evaluate our models under various unseen conditions, including five novel dialects that our models have not encountered before. Additionally, we compute category-wise [7] error rates for both the SADA test

[7]By *categories* we refer to dialectal and non-dialectal labels in the SADA data.

Model	Bench Test	Bench Valid	SADA2022 Test	SADA2022 Valid	IH
Baselines					
HuBERT	51.4	54.4	75.6	73.7	81.9
W-FT	42.2	47.4	92.3	81.7	93.2
SM4T-v1	37.4	39.5	64.9	60.3	66.0
SM4T-v2	<u>32.3</u>	<u>34.7</u>	65.0	63.2	61.7
W-M	54.3	59.6	99.3	91.9	76.4
W-L-v2	41.9	46.8	69.8	66.5	68.2
W-L-v3	**31.4**	**33.0**	67.2	58.4	67.7
DW-16-16	40.0	42.2	66.7	65.2	66.3
DW-32-16	35.3	37.7	66.9	64.5	61.1
DW-16-16++	36.7	38.8	69.9	64.7	62.3
DW-32-16++	33.0	34.9	60.3	57.1	**56.9**
No-Filter					
– DW-16-16	43.1	46.5	78.4	72.5	71.2
– DW-32-16	40.1	43.6	70.7	66.0	65.8
Ours					
UDW-16-16					
– proxy	40.0	42.3	68.6	65.4	67.1
– sonar	42.7	44.9	74.9	71.2	71.2
UDW-32-16					
– proxy	36.3	39.1	64.6	58.7	62.6
– sonar	35.5	37.8	<u>58.7</u>	56.4	60.6
UDW-16-16++					
– proxy	36.9	41.4	64.1	60.2	63.6
– sonar	37.2	39.3	61.7	59.4	63.6
UDW-32-16++					
– proxy	35.1	37.2	58.9	<u>55.1</u>	<u>60.3</u>
– sonar	35.4	37.7	**57.9**	**54.7**	60.9

Table 2: Average WER across different evaluation datasets. Bench: CV15.0, FLEURS and the three MGBs. IH: In-House data. Best results are shown in **bold**. Second best results are <u>underlined</u>. WER scores are reported after normalization and removing diacritics.

Model	NJD	MTOS	KHLJ	HJZ	UNK
Baselines					
W-FT	106.4	77.0	117.8	84.2	139.9
SM4T-v1	56.5	74.0	61.9	54.6	69.6
SM4T-v2	52.6	76.2	58.2	53.3	74.7
W-M	88.2	107.9	102.8	86.7	134.0
W-L-v2	56.6	74.0	80.2	63.4	98.2
W-L-v3	53.1	73.7	70.5	61.7	89.0
DW-16-16	58.2	74.9	65.1	58.2	68.4
DW-32-16	58.7	74.9	66.7	58.7	70.1
DW-16-16++	58.8	81.9	62.1	57.5	77.5
DW-32-16++	51.7	67.8	<u>57.5</u>	52.6	67.5
No-Filter					
– DW-16-16	61.6	90.4	72.9	68.5	88.8
– DW-32-16	58.2	82.3	67.3	60.4	76.6
Ours					
UDW-16-16					
– proxy	60.2	75.6	65.7	62.1	75.6
– sonar	64.6	<u>59.9</u>	84.6	72.1	66.6
UDW-32-16					
– proxy	53.4	75.7	60.0	51.4	72.8
– sonar	**48.8**	**52.6**	67.5	58.3	**48.9**
UDW-16-16++					
– proxy	53.4	73.9	61.9	53.3	68.6
– sonar	53.7	69.4	61.4	53.7	64.6
UDW-32-16++					
– proxy	<u>48.9</u>	66.7	60.6	**48.7**	66.0
– sonar	50.5	65.3	**57.3**	<u>49.1</u>	<u>61.8</u>

Table 3: WER results on top five dialects/categories on the test set of the SADA data. NJD: Najdi. MTOS: More than one speaker. KHLJ: Khaleeji. HJZ: Hijazi. UNK: Unknown. Best results are shown in **bold**. Second best results are <u>underlined</u>. WER scores are reported after normalization and removing diacritics.

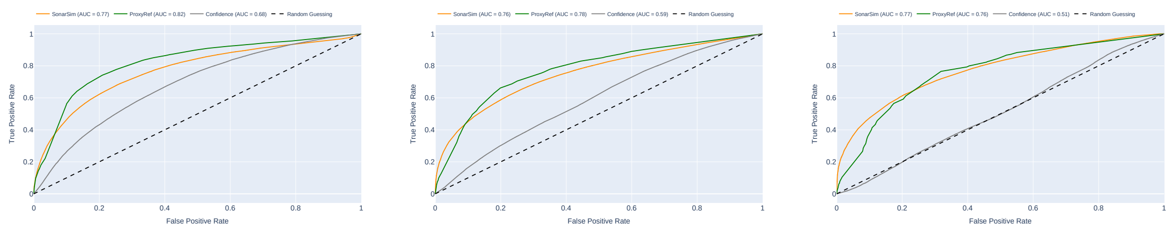
and validation splits. We present results on the top five categories. As detailed in Table 9, our best model UDW-32-16++ consistently outperforms all others when averaged across these dialects.

Furthermore, UDW-32-16++, demonstrates superior performance over DW-32-16++ in the top five SADA categories. For instance, when using *proxy-ref* as a filtering measure, UDW-32-16++ achieves 58.06% WER, compared to DW-32-16++'s 59.42% averaged across top (with most utterances) five categories in the SADA *test* split. This demonstrates our ability to (1) distill smaller models from larger Whisper models, (2) maintain or improve performance, and (3) reduce model size without relying on labeled data.

Effectiveness of Unsupervised Metrics to Filter Low-Quality Pesudo Labels. We investigate the effectiveness of two of our best metrics for filtering low-quality pseudo-labels, specifically targeting instances with a WER higher than 80%, 40%, and 20%. To assess their efficacy, we calculate the area under the curve (AUC) (as shown in Figure 1) for detecting low-quality examples. The results indicate that *sonar-sim* achieves an AUC of 0.77 for detecting examples with a WER > 80, demonstrating reasonably high discriminative power in identifying low-quality labels. The *proxy-ref* metric shows a slightly better performance, with an AUC of 0.82, indicating robust capability in distinguishing between high and low-quality pseudo-labels. In contrast, the confidence-based measure yielded an AUC of 0.68, which falls behind the other measures' discriminative power. These findings highlight sonar embeddings and the proxy reference-based measure as promising tools for improving the quality of pseudo-labels in scenarios where ground truth data is unavailable.

Model	ALG	EGY	JOR	MAU	MOR	PAL	UAE	YEM	AVG
Baselines									
SM4T-v2	94.3	**52.3**	**39.2**	88.9	91.0	**49.0**	**54.7**	**62.4**	**67.3**
W-L-v2	89.9	58.1	43.2	108.3	101.1	51.9	60.8	81.4	75.6
DW-16-16	85.7	65.3	51.1	88.5	86.2	60.2	64.1	69.3	71.9
DW-32-16	86.6	65.1	51.0	89.2	87.1	59.9	63.0	69.8	72.1
No-Filter									
– DW-32-16	93.2	66.6	48.1	95.3	86.0	58.3	63.8	68.8	73.1
Ours									
UDW-16-16									
– proxy	85.9	67.1	51.0	88.0	87.4	61.3	63.5	70.8	72.6
– sonar	88.4	69.8	57.3	91.3	90.9	65.7	68.4	72.8	76.2
UDW-32-16									
– proxy	85.6	61.8	46.2	87.9	**84.5**	57.8	59.1	64.8	69.1
– sonar	**82.1**	58.8	45.5	**86.0**	85.5	53.3	57.2	63.3	**67.3**

Table 4: Results on the Casablanca dataset. Best results are shown in bold. Second best results are underlined. WER (↓) scores are reported after normalization and removing diacritics. We report the score on the test split of each dataset.



(a) WER > 80 (b) WER > 40 (c) WER > 20

Figure 1: Area under the curve (AUC) for detecting low-quality examples (WER > 20%, 40%, 80%). The Y-axis represents the true positive rate (TPR), and the X-axis represents the false positive rate (FPR).

Evaluation	Dataset	Baselines W-L-v2	Baselines DW-16-16	Baselines DW-32-16	Ours UDW-16-16$_{pr}$	Ours UDW-32-16$_{pr}$
IID	OpenBible	101.3	59.1	**58.8**	59.2	58.9
	CommonVoice17	117.1	82.9	**69.8**	75.6	70.4
	ALFAA	217.1	78.2	74.4	76.8	**73.8**
OOD	DVoice	214.6	124.4	**110.2**	110.7	114.9
	AMMI-LigAikuma	**46.7**	60.1	51.8	60.4	52.2
	Fleurs	54.6	60.9	**51.6**	58.9	51.8

Table 5: WER (↓) results on the Swahili datasets. pr: using the proxy filtering method. Best results are shown in **bold**. Second best results are underlined. WER scores are reported after normalization and removing diacritics.

5.1 Experiments on Other Language.

To further validate the effectiveness of our approach, we conduct experiments on Swahili, a low-resource language. We collect over 100 hours of labeled speech data from a variety of sources, namely OpenBible (Meyer et al., 2022), CommonVoice (Swahili subset) (Ardila et al., 2020), ALFAA[8], DVoice (Gauthier et al., 2016), AMMI-LIGAikuma[9], and FLEURS (Swahili subset) (Conneau et al., 2023).

We distill two models, UDW-16-16 and UDW-32-16, using our best filtering method: *proxy-ref*. The training data includes the train splits of OpenBible, CommonVoice, and ALFAA, and we evaluate the models on their respective test splits. We also test the models on three out-of-distribution (OOD) datasets: DVoice, AMMI-LigAikum, and FLEURS, which were not included in the training data.

We compare our distilled models to the teacher model and evaluate the performance of our unsupervised approach. The results show that our unsupervised distillation models perform on par with, or better than the supervised setup. Additionally, our distilled models outperform the teacher model by a significant margin on both familiar (IID) and novel (OOD) datasets, demonstrating the utility of our approach in extremely low-resource settings. Specifically, the UDW-32-16 model achieves a WER/CER of 58.86/14.13% on the IID OpenBible dataset, compared to the teacher model's 101.33/44.43%. On the OOD dataset FLEURS, UDW-32-16 attains a WER/CER of 51.82/14.88, significantly outperforming the teacher model's 54.61/14.81. Across various datasets, our distilled models consistently outperform the teacher, with UDW-32-16 showing the best results overall. Table 5 presents the WER and CER scores for the different models and datasets.

These findings highlight the strength of our unsupervised data filtering approach, particularly in low-resource scenarios, where labeled data is scarce but the distilled models still perform robustly.

6 Conclusion

In this study, we explore methods for distilling large Whisper models into smaller, more efficient ones without relying on labeled data. Our filtering techniques bridge a gap in prior research and facilitate the creation of compact and effective speech recognition models for limited label settings. We show through a comprehensive evaluation that our models outperform both their teacher model and those using supervised distillation. Our evaluation spans a diverse range of Arabic varieties, demonstrating their generalization to linguistic diversity and their competitive performance with SOTA models twice their size. Applying our approach to Swahili datasets further validates its effectiveness for different languages. Notably, our model-based filtering methods (proxy and sonar) demonstrate superior robustness across linguistic variations. Moving forward, we aim to explore model-free approaches to further enhance the efficacy of model distillation, while including extremely low-resource languages and domains.

7 Limitations

In this study, we distill small Whisper models from relatively large ones via pseudo-labeling and unsupervised data filtering. Our distilled models are computationally efficient and maintain a performance similar to or better than the base teacher model and models trained in a supervised data filtering setup. Unlike Waheed et al. (2024); Gandhi et al. (2023), our approach does not utilize any labeled data in the distillation process, making it directly applicable in data-scarce settings. However, despite these advantages, we acknowledge several limitations in our work, which we outline below.

Efficiency. Our distilled models achieve 25-50% compute efficiency relative to their larger counterparts while maintaining comparable performance. However, the training of these models requires significant computational resources.

Our main approach relies heavily on a robust reference model to serve as a proxy for filtering lower-quality pseudo labels. Specifically, we utilize SeamlessM4T-large-v2, a state-of-the-art model with 2.3 billion parameters, to generate proxy references which is then used to filter out low-quality data points. For similarity-based measures, we use SONAR (Duquenne et al., 2023) to generate multimodal embeddings from speech and pseudo labels. These embeddings provide contextual similarity which is then utilized to discard low-quality pseudo labels. We use AceGPT (7B), to compute the log-likelihood of the pseudo labels which is leveraged to filter out low-quality examples.

Although these measures allow attaining a performance on par or better than the supervised setup, it's important to highlight that each of these methodologies entails additional computational overhead.

Multilinguality. We use SeamlessM4T-large-v2 for generating proxy references, SONAR for gen-

[8] https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR/SWAHILI

[9] https://github.com/besacier/AMMIcourse

erating multimodal embeddings, AceGPT (7B) for computing log-likelihood, and XTTS-v2 for generating synthetic speech. The multilingual capabilities of these models are crucial for effectively applying our techniques to a wide range of languages and dialects. However, a significant limitation of our approach is that it is constrained to languages supported by these models. This dependency restricts our ability to extend our distillation process to languages beyond the scope of the models' multilingual capacities.

Evaluation. Arabic is a linguistically rich and complex language with over 400 million speakers (Abdul-Mageed et al., 2021, 2024), resulting in its wide range of varieties and dialects. We evaluate all the models on eleven different datasets representing different varieties, including five novel dialects collected and curated by native speakers and never seen before by any models. However, our varieties do not cover all Arabic-speaking regions. We aim to address this in future work by covering more varieties and dialects.

Distillation Training Data. We distilled four variants of student models using 100K and 500K segments of which approximately 25% are filtered. We see improvement going from 100K (≈100 hours) to 500K (≈500 hours) segments. As Gandhi et al. (2023) shows, going over 1,000 hours results in a better model, we aim to study how distillation can be done under a low resource setting which is why we do not scale the data. Additionally, we also keep the WER threshold high (80) so that we remain close to a setting where no labeled data is available (even for filtering). It would be interesting, however, to see how distilled models may perform on unfiltered data in low-resource setting.

Nature of Speech Data. Despite putting together a never-seen dataset of under-represented Arabic dialects, we realize that sourcing our data from television series renders its nature distant from speech spoken *in the wild*. This type of content tends to be more "theatrical" and involves different elements such as background music and laughing tracks that do not accurately reflect regular conversational Arabic. Consequently, this could fail to accurately portray the performance of these models on real speech.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[10] and UBC Advanced Research Computing-Sockeye.[11]

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2023. Masc: Massive arabic speech corpus. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1006–1013.

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim, Raghad Aloraini, Raneem Alnajim, Ranya Alkahtani, Renad Almuasaad, Sara Alrasheed, Shaykhah Alsubaie, and Yaser Alonaizan. 2024. Sada: Saudi audio dataset for arabic. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284.

[10]https://alliancecan.ca
[11]https://arc.ubc.ca/ubc-arc-sockeye

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019. The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282.

Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. 2022. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7087–7091.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady ElSahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Peng Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Shang-Wen Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, M.L. Ramadan, Abinesh Ramakrishnan, Anna Sun, Ke M. Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bo Yu, Pierre Yves Andrews, Can Balioglu, Marta Ruiz Costa-jussà, Onur Çelebi, Maha Elbayad, Cynthia Gao, Francisco Guzm'an, Justine T. Kao, Ann Lee, Alexandre Mourachko, Juan Miguel Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual&multimodal machine translation.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations.

Thomas Palmeira Ferraz, Marcely Zanon Boito, Caroline Brun, and Vassilina Nikoulina. 2024. Multilingual distilwhisper: Efficient distillation of multi-task speech models via language-specific experts. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10716–10720.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *Preprint*, arXiv:2311.00430.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof. In *International Conference on Language Resources and Evaluation*.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Nawar Halabi. 2016. *Modern standard Arabic phonetics for speech synthesis*. Ph.D. thesis, University of Southampton.

Michael Hentschel, Yuta Nishikawa, Tatsuya Komatsu, and Yusuke Fujita. 2024. Keep decoding parallel with effective knowledge distillation from language models to end-to-end speech recognisers. *Preprint*, arXiv:2401.11700.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Ming-Hao Hsu, Kuan Po Huang, and Hung yi Lee. 2024. Meta-whisper: Speech-based meta-icl for asr on low-resource languages. *Preprint*, arXiv:2409.10429.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Hengtong Hu, Lingxi Xie, Richang Hong, and Qi Tian. 2020. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3123–3132.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. Acegpt, localizing large language models in arabic. *Preprint*, arXiv:2309.12053.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. *Preprint*, arXiv:2211.17192.

Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. 2017. Data-free knowledge distillation for deep neural networks. *CoRR*, abs/1710.07535.

Hugo Malard, Salah Zaiem, and Robin Algayres. 2023. Big model only for hard audios: Sample dependent whisper model selection for efficient inferences. *Preprint*, arXiv:2309.12712.

Vimal Manohar, Pegah Ghahremani, Daniel Povey, and Sanjeev Khudanpur. 2018. A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 250–257.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Samuel Olanrewaju, Jesujoba Alabi, and Shamsuddeen Muhammad. 2022. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus. *Preprint*, arXiv:2207.03546.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. Qasr: Qcri aljazeera speech resource – a large scale annotated arabic speech corpus. *Preprint*, arXiv:2106.13000.

Khandokar Md. Nayem, Ran Xue, Ching-Yun (Frannie) Chang, and Akshaya Vishnu Kudlu Shanbhogue. 2023. Knowledge distillation on joint task end-to-end speech translation. In *Interspeech 2023*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Yael Segal-Feldman, Aviv Shamsian, Aviv Navon, Gill Hetz, and Joseph Keshet. 2024. Whisper in medusa's ear: Multi-head efficient decoding for transformer-based asr. *Preprint*, arXiv:2409.15869.

Hang Shao, Wei Wang, Bei Liu, Xun Gong, Haoyu Wang, and Yanmin Qian. 2023. Whisper-kdq: A lightweight whisper via guided knowledge distillation and quantization for efficient asr. *Preprint*, arXiv:2305.10788.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *Preprint*, arXiv:1908.09355.

SYSTRAN. faster-whisper: Faster whisper transcription with ctranslate2.

Bashar Talafha, Karima Kadaoui, Samar Magdy, Mariem Habiboullah, Chafei Chafei, Ahmed El-Shangiti, Hiba Zayed, Mohamedou Tourad, Rahaf Alhamouri, Rwaa Assi, et al. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758.

Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-shot benchmarking of whisper on diverse arabic speech recognition. In *Interspeech 2023*, pages 5092–5096.

Sanli Tian, Keqi Deng, Zehan Li, Lingxuan Ye, Gaofeng Cheng, Ta Li, and Yonghong Yan. 2022. Knowledge distillation for ctc-based speech recognition via consistent acoustic representation learning. In *Interspeech 2022*, pages 2633–2637.

Abdul Waheed, Karima Kadaoui, and Muhammad Abdul-Mageed. 2024. To distill or not to distill? on the robustness of robust knowledge distillation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12603–12621, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoyu Yang, Qiujia Li, Chao Zhang, and Philip C. Woodland. 2023. Knowledge distillation from multiple foundation models for end-to-end speech recognition. *Preprint*, arXiv:2303.10917.

Jeong Hun Yeo, Minsu Kim, Shinji Watanabe, and Yong Man Ro. 2024. Visual speech recognition for languages with limited labeled data using automatic labels from whisper. *Preprint*, arXiv:2309.08535.

Bo Zhang, Xiaoming Zhang, Yun Liu, Lei Cheng, and Zhoujun Li. 2021. Matching distributions between model and data: Cross-domain knowledge distillation for unsupervised domain adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5423–5433, Online. Association for Computational Linguistics.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *Preprint*, arXiv:2303.01037.

A Appendix

B Dataset

2.1 SADA Dataset

Dialect	Test(S/D)	Valid (S/D)
Najdi	1703/2.0709	2249/3.3155
MTOS	1320/4.8044	1048/3.82
Khaliji	1150/1.1308	679/0.6317
Hijazi	809/1.1202	528/0.6423
Unknown	762/0.8325	489/0.4861
NA	167/0.1341	2/0.0004
MSA	157/0.5406	54/0.1682
Egyptian	96/0.0865	45/0.0524
Shamali	18/0.0243	-
Yemeni	7/0.0052	23/0.0349
Levantine	-	19/0.0137
Total	6189/10.75	5136/9.17

Table 6: SADA stats. S is the number of segments and D is the duration (in hours). **MTOS** - More than one speaker.

C Experiments

3.1 CER Results

3.2 Training Parameters

3.3 Results

D Efficacy

Model	Size	CV15.0	MGB2	MGB3	MGB5	Fleurs	In-house Data					SADA
							ALG	JOR	PAL	UAE	YEM	
Baselines												
Amazon	-/-	-/-	-/-	-/-	-/-	-/-	70.2	25.6	29.0	40.8	43.5	-/-
XLS-R	0.96	39.4	53.1	61.6	68.0	43.9	67.0	61.4	61.1	64.6	63.6	68.3
HuBERT	0.31	18.9	17.3	**9.5**	45.5	10.9	44.3	23.3	27.9	36.7	38.8	34.5
W-FT	1.5	21.9	8.1	26.9	62.3	3.4	69.6	37.2	35.4	69.1	64.8	65.7
MMS-all	1.0	80.9	13.4	34.6	45.9	6.3	78.0	55.4	75.1	78.1	76.6	38.0
SM4T-M	1.2	5.7	9.0	21.7	46.6	3.6	39.7	15.9	20.1	24.7	29.5	39.3
SM4T-L-v1	2.3	7.3	10.5	22.6	52.1	5.1	47.8	18.8	23.1	27.4	32.5	37.8
SM4T-L-v2	2.3	**3.5**	8.7	18.6	53.7	4.0	52.0	14.6	17.2	23.3	30.7	41.8
W-S	0.24	16.4	24.7	51.9	164.8	8.7	84.7	32.9	36.3	59.7	66.7	103.6
W-M	0.77	13.2	18.5	39.5	88.3	5.1	69.9	21.1	24.7	52.6	52.0	74.1
W-L-v2	1.5	7.8	15.3	33.0	68.9	3.6	71.7	17.0	22.3	38.2	45.5	51.2
W-L-v3	1.5	5.2	**7.6**	17.3	44.6	3.2	65.4	16.3	22.7	32.7	38.9	45.6
DW-16-16	0.80	7.2	10.8	25.1	43.3	6.6	38.5	18.2	23.3	27.7	31.6	38.9
DW-32-16	1.12	5.9	8.9	21.4	**40.4**	4.8	33.4	14.7	19.5	22.8	28.1	49.2
DW-16-16++	0.80	6.2	10.2	24.8	42.6	5.2	39.0	17.2	21.6	26.8	31.5	40.6
DW-32-16++	1.12	5.5	8.8	20.3	40.6	**3.1**	**33.3**	**13.4**	18.8	21.1	**26.8**	35.8
No-filter												
– DW-16-16	0.80	7.6	11.2	29.7	59.1	6.0	51.6	20.2	27.3	34.0	38.8	49.6
– DW-32-16	1.12	7.3	10.4	30.8	58.8	4.9	63.2	20.0	24.9	35.6	50.9	53.6
Ours												
UDW-16-16	0.80											
– nll		8.15	11.26	27.98	55.25	6.26	41.4	25.7	20.52	35.96	49.2	55.0
– pesq		8.41	12.11	27.69	54.88	6.89	40.34	27.41	20.16	32.55	44.17	50.1
– entropy		8.1	12.17	31.24	56.64	6.4	48	22.81	27.67	37.85	52.56	61.8
– conf		7.83	11.87	27.85	50.73	6.12	43.94	20.29	25.52	31.75	39.27	49.2
– proxy		7.48	11.39	26.36	49.97	7.5	42.15	23.69	19.66	30.93	41.94	46.2
– sonar		8.04	11.86	28.66	49.21	7.06	43.48	22.61	27.43	32.89	36.13	45.6
UDW-32-16	1.12											
– nll		6.24	10.12	25.39	55.53	4.47	35.85	20.88	16.04	30.49	41.38	46.1
– pesq		7.5	10.6	26.4	51.0	5.3	43.2	17.2	22.7	30.9	36.4	48.1
– entropy		6.53	10.34	28.71	66.87	4.34	84.02	21.07	31.08	44.23	54.25	52.3
– conf		6.46	9.79	23.42	48.61	4.73	36.08	21.99	25.82	**17.7**	42.82	41.3
– proxy		6.17	9.87	22.45	46.05	6.23	36.11	15.6	20.88	25.69	29.49	41.9
– sonar		5.62	8.98	23.4	46.97	4.41	37	15.11	19.56	24.29	28.24	35.6
UDW-16-16++	0.80											
– proxy		4.8	10.4	24.3	48.0	4.8	41.6	16.3	21.1	27.5	33.0	35.3
– sonar		6.1	9.8	24.2	47.1	5.0	38.9	17.2	21.4	26.2	29.5	33.96
UDW-32-16-++	1.12											
– proxy		5.8	9.2	21.1	44.1	4.2	37.1	14.5	20.1	23.1	29.2	**31.4**
– sonar		5.3	9.9	22.4	44.2	3.9	34.6	19.2	**15.1**	23.2	27.0	33.4

Table 7: CER (↓) scores after normalization and removing diacritics. All baseline distilled models (DW-) are trained with a filtering threshold of 80 if not specified. Best results are shown in **bold**. Second best results are underlined. We report the score on the test split of each dataset. Abbreviations. **W** - Whisper, **FT** - Finetuned, **M** - Medium, **L** - Large, **S** - Small, **U** - Unsupervised, **D** - Distil, **nll** - negative log likelihood, **conf** - confidence score.

Model	Split	NJD	MTOS	KHLJ	HJZ	UNK
Baslines						
W-FT	Test	77.5	51.8	85.4	61.5	112.2
	Valid	52.6	41.1	100.3	89.7	107.6
SM4T-v1	Test	30.9	46.0	32.2	29.0	39.4
	Valid	28.1	44.0	31.5	30.9	35.2
SM4T-v2	Test	31.1	53.1	30.4	32.0	45.1
	Valid	30.7	53.7	35.3	30.3	34.4
W-M	Test	65.8	79.3	77.0	59.7	122.2
	Valid	56.9	75.1	62.9	52.0	106.5
W-L-v2	Test	39.9	57.4	54.4	39.6	80.7
	Valid	41.4	55.4	44.9	43.6	67.1
W-L-v3	Test	31.6	53.7	44.1	38.6	61.3
	Valid	30.2	47.7	39.2	27.2	49.2
DW-16-16	Test	30.8	47.6	32.7	30.7	39.8
	Valid	31.4	44.7	35.2	32.8	39.8
DW-32-16	Test	35.8	60.1	38.7	34.1	44.5
	Valid	34.8	54.1	37.6	38.2	40.1
DW-16-16++	Test	30.9	50.7	31.5	31.0	46.8
	Valid	29.8	43.8	31.8	33.0	41.0
DW-32-16++	Test	28.3	43.1	29.4	28.6	41.3
	Valid	27.3	38.3	34.5	28.1	43.0
No-Filter						
– DW-16-16	Test	34.8	59.7	41.4	42.3	63.0
	Valid	38.9	53.4	41.9	37.7	54.3
– DW-32-16	Test	42.8	63.9	47.0	45.9	63.2
	Valid	35.2	54.9	43.3	36.5	49.6
Ours						
UDW-16-16						
– proxy	Test	35.5	55.6	38.9	39.6	52.0
	Valid	34.0	50.9	39.1	37.1	41.2
– sonar	Test	35.8	30.3	55.7	38.8	36.8
	Valid	35.9	39.3	52.4	38.7	36.7
UDW-32-16						
– proxy	Test	31.1	54.0	32.1	30.8	46.0
	Valid	29.3	44.3	29.1	28.6	36.6
– sonar	Test	25.4	**23.6**	45.9	29.9	**25.5**
	Valid	26.0	26.7	44.1	30.3	29.5
UDW-16-16++						
– proxy	Test	29.7	48.8	33.8	29.6	42.8
	Valid	27.8	42.0	34.3	32.2	41.7
– sonar	Test	28.4	43.3	30.8	27.5	37.0
	Valid	27.5	40.3	32.4	30.7	35.8
UDW-32-16++						
– proxy	Test	**25.3**	41.3	31.0	**24.6**	38.2
	Valid	**25.3**	37.4	30.1	25.6	37.4
– sonar	Test	26.3	40.8	**28.3**	24.8	34.5
	Valid	**25.3**	37.0	30.2	29.8	34.2

Table 8: CER (↓) results on top five dialects/categories in SADA data. Best results are shown in **bold**. Second best results are underlined. The scores are reported after normalization and removing diacritics.

Model	NJD	MTOS	KHLJ	HJZ	UNK
Baselines					
W-FT	77.1	63.4	139.4	119.1	140.3
SM4T-v1	51.9	68.7	61.7	54.2	62.3
SM4T-v2	52.2	75.8	65.1	51.1	59.8
W-M	80.4	102.8	89.5	72.9	127.7
W-L-v2	60.9	72.9	67.7	64.5	68.0
W-L-v3	49.3	65.5	67.5	**46.5**	67.7
DW-16-16	59.4	70.6	66.2	61.1	69.9
DW-32-16	58.3	69.7	67.5	62.7	68.3
DW-16-16++	56.8	72.0	62.3	60.2	75.2
DW-32-16++	50.3	61.8	62.3	53.7	66.4
No-Filter					
– DW-16-16	64.8	80.3	71.4	65.5	77.0
– DW-32-16	57.9	73.7	68.6	56.8	72.3
Ours					
UDW-16-16					
– proxy	59.3	70.7	66.5	61.4	68.7
– sonar	64.8	67.5	78.1	69.9	65.3
UDW-32-16					
– proxy	51.0	65.9	58.1	53.7	64.9
– sonar	**49.2**	**51.6**	62.5	58.9	**52.6**
UDW-16-16++					
– proxy	52.7	66.6	62.0	55.7	67.3
– sonar	53.6	64.4	62.3	55.7	63.7
UDW-32-16++					
– proxy	49.0	60.4	**57.6**	50.7	60.7
– sonar	49.3	58.8	58.1	53.6	61.1

Table 9: WER (↓) results on top five dialects/categories on the validation set of the SADA data. Best results are shown in **bold**. Second best results are underlined. WER scores are reported after normalization and removing diacritics.

Model	Bench Test	Bench Valid	SADA2022 Test	SADA2022 Valid	IH
Baselines					
HuBERT	20.4	22.7	34.5	31.9	34.2
W-FT	24.5	28.6	65.7	56.2	55.2
SM4T-v1	19.5	21.2	37.8	35.6	29.9
SM4T-v2	17.7	19.4	41.8	40.8	27.6
W-M	32.9	38.0	74.1	66.7	44.1
W-L-v2	25.7	29.7	51.2	48.5	38.9
W-L-v3	**15.6**	**17.1**	45.6	39.3	35.2
DW-16-16	18.6	20.5	38.9	37.8	27.8
DW-32-16	16.3	<u>17.9</u>	47.3	43.9	<u>23.7</u>
DW-16-16++	17.8	19.4	40.6	36.6	27.2
DW-32-16++	<u>15.7</u>	**17.1**	35.8	33.4	**22.7**
No-Filter					
– DW-16-16	22.7	25.1	49.6	45.9	34.4
– DW-32-16	22.5	25.8	53.6	45.0	38.9
Ours					
UDW-16-16					
– proxy	20.5	22.1	46.2	42.2	31.7
– sonar	21.0	22.8	45.6	43.5	32.5
UDW-32-16					
– proxy	18.2	19.6	41.9	36.1	25.6
– sonar	17.9	19.7	35.6	34.7	24.8
UDW-16-16++					
– proxy	18.4	21.2	39.2	35.3	27.9
– sonar	18.4	19.9	35.4	34.0	26.6
UDW-32-16++					
– proxy	17.1	18.7	<u>33.6</u>	<u>31.4</u>	23.8
– sonar	16.9	18.5	**33.1**	**31.3**	24.8

Table 10: Average CER (↓) across different evaluation datasets. Bench: CV15.0, FLEURS and the three MGBs. Best results are shown in **bold**. Second best results are <u>underlined</u>. The scores are reported after normalization and removing diacritics.

Evaluation	Dataset	Baselines			Ours	
		W-L-v2	DW-16-16	DW-32-16	UDW-16-16$_{pr}$	UDW-32-16$_{pr}$
IID	OpenBible	44.4	<u>14.0</u>	**13.8**	14.0	14.1
	CommonVoice17	60.1	35.0	**24.8**	29.2	<u>25.4</u>
	ALFAA	143.2	28.2	**25.7**	27.2	<u>26.5</u>
OOD	DVoice	144.6	74.1	<u>62.6</u>	**62.4**	69.1
	AMMI-LigAikuma	**13.0**	18.0	<u>14.4</u>	18.5	14.4
	Fleurs	<u>14.8</u>	18.9	**14.8**	18.5	14.9

Table 11: CER (↓) results on the Swahili datasets. pr: using the proxy filtering method. Best results are shown in **bold**. Second best results are <u>underlined</u>. WER scores are reported after normalization and removing diacritics

Model	ALG	EGY	JOR	MAU	MOR	PAL	UAE	YEM	AVG
Baselines									
SM4T-v2	53.48	**26.12**	**13.15**	52.20	54.96	**18.20**	**22.71**	**27.07**	34.44
W-L-v2	58.63	30.28	20.37	79.66	63.21	25.70	38.06	51.49	46.83
DW-16-16	<u>40.08</u>	31.80	19.11	<u>49.83</u>	<u>42.16</u>	24.10	26.99	30.53	33.64
DW-32-16	44.45	32.80	19.27	49.95	43.46	26.43	26.26	34.03	35.12
No-Filter									
– DW-32-16	61.50	43.52	18.41	64.19	51.36	29.44	36.97	41.75	43.95
Ours									
UDW-16-16									
– proxy	48.30	39.79	20.21	53.06	45.92	25.69	29.15	37.13	38.01
– sonar	43.94	36.24	23.60	55.10	50.14	28.77	31.05	34.65	38.63
UDW-32-16									
– proxy	40.72	29.89	16.23	**47.03**	**41.45**	23.42	23.72	27.26	<u>31.80</u>
– sonar	**38.34**	<u>28.61</u>	<u>16.02</u>	50.02	44.94	<u>19.87</u>	<u>23.13</u>	<u>27.17</u>	**31.79**

Table 12: CER results on the Casablanca dataset. The best results are shown in bold. The second-best results are underlined. CER (↓) scores are reported after normalization and removing diacritics. We report the score on the test split of each dataset.

Parameter	Value
`warmup_steps`	50
`learning_rate`	0.0001
`lr_scheduler_type`	`constant_with_warmup`
`batch_size`	128
`max_label_length`	225
`gradient_accumulation_steps`	1
`dtype`	`bfloat16`

Table 13: Training parameters. All the training parameters are the default ones provided in Huggingface Seq2SeqTrainingArguments unless specified otherwise in this table.

Model	Size	CV15.0	MGB2	MGB3	MGB5	Fleurs	In-house Data					SADA
							ALG	JOR	PAL	UAE	YEM	
Baslines												
Amazon	-/-	-/-	-/-	-/-	-/-	-/-	88.0/71.6	59.2/29.1	63.4/32.2	71.1/44.3	77.4/47.7	-
XLS-R	0.96	92.7/46.7	97.7/54.5	99.1/64.5	99.6/70.1	95.1/45.4	99.7/68.0	99.3/62.9	99.2/62.8	99.5/66.4	99.7/66.4	99.6/69.5
HuBERT	0.31	76.5/31.0	59.4/20.3	**43.3/16.5**	95.0/48.7	48.9/14.4	96.2/45.6	70.6/25.4	81.5/31.4	87.9/39.9	91.3/40.8	81.3/37.1
W-FT	1.5	70.0/33.8	29.4/10.9	60.1/32.2	105.0/64.3	28.7/7.3	114.5/70.3	75.1/39.0	81.3/38.7	113.7/70.9	110.1/65.6	101.4/67.6
MMS-all	1.0	106.0/82.5	40.3/14.0	77.7/38.1	90.4/48.5	28.8/7.8	100.2/77.8	91.5/56.2	100.0/75.8	100.1/78.4	100.1/76.8	79.8/39.1
SM4T-M	1.2	42.3/18.2	28.1/11.2	50.2/26.8	88.2/50.8	19.5/6.0	84.5/42.8	55.2/18.7	63.0/23.0	68.0/28.1	79.4/34.5	73.2/42.8
SM4T-L-v1	2.3	44.2/19.1	25.9/11.7	52.5/27.6	92.8/55.9	22.6/7.6	89.7/50.3	59.1/21.7	64.7/25.8	69.0/30.3	81.5/37.0	72.4/40.8
SM4T-L-v2	2.3	**37.7/15.8**	22.4/9.9	46.7/23.9	92.1/58.4	19.8/6.5	94.8/55.2	51.3/17.6	**58.5/20.1**	65.6/26.9	80.6/35.5	72.2/44.4
W-S	0.24	68.9/31.8	49.5/25.7	84.8/55.4	228.6/164.5	33.4/10.3	129.15/87.85	75.25/36.55	79.73/39.3	103.83/63	112.69/70.69	144.5/106.6
W-M	0.77	55.1/24.2	37.6/19.6	71.5/43.7	129.7/89.4	24.0/7.1	103.9/71.4	59.0/23.9	66.8/27.6	90.7/55.7	95.2/56.2	106.0/76.3
W-L-v2	1.5	46.9/19.6	33.7/16.9	60.6/37.7	101.1/71.1	19.7/5.6	106.9/74.6	51.2/19.6	60.2/25.2	73.2/41.2	86.9/50.1	78.0/53.5
W-L-v3	1.5	43.2/16.9	**20.4/8.6**	44.6/22.5	82.0/47.7	**16.4/4.8**	103.8/68.9	52.7/18.9	64.3/26.4	72.3/35.9	86.0/43.3	74.6/47.9
DW-16-16	0.80	48.0/18.9	33.2/12.5	57.1/29.6	84.1/46.2	26.2/8.5	83.8/40.2	57.8/20.5	68.2/26.2	72.0/31.0	80.0/35.6	72.0/40.9
DW-32-16	1.12	45.6/17.7	27.7/10.3	51.2/26.1	80.9/**43.4**	22.0/6.6	**80.5/35.1**	52.6/17.1	62.9/22.4	66.7/26.3	77.3/32.6	66.9/47.3
DW-16-16++	0.80	44.1/17.1	28.5/10.5	54.5/28.5	83.2/45.6	22.4/6.9	82.3/38.7	55.4/18.9	65.2/24.9	69.3/28.2	76.8/33.0	76.0/42.7
DW-32-16++	1.12	44.7/17.3	25.2/10.0	48.8/25.2	**79.0**/43.7	20.2/5.0	76.4/35.4	**50.0/15.9**	60.1/21.8	**63.2/24.7**	**73.5/31.5**	67.2/38.1
No-filter												
– DW-16-16	0.0	48.3/19.1	34.2/13.0	60.2/33.9	96.8/61.3	24.9/7.9	93.8/53.1	58.9/22.4	72.3/30.2	75.5/37.2	84.8/42.3	83.9/51.6
– DW-32-16	0.0	47.2/18.8	29.0/11.8	58.3/35.0	92.5/60.8	23.5/7.0	88.0/64.2	55.3/22.4	65.5/27.9	72.5/38.8	80.4/53.8	76.7/55.5
Ours												
UDW-16-16												
– nll	0.0	49.12/19.55	32.18/12.64	60.76/32.4	94.05/57.88	25.75/8.09	88.88/44.93	70.14/28.53	59.84/22.8	78.73/38.95	93.3/50.73	89.2/56.9
– pesq	0.0	49.46/19.73	34.27/13.55	60.58/31.99	95.63/57.22	26.8/8.64	87.91/43.83	71.92/30.19	60.01/22.47	76.7/35.72	87.51/45.81	84.7/52.1
– entropy	0.0	49.03/19.5	33.4/13.57	62.74/35.19	96.45/58.9	25.81/8.29	90.31/49.75	61.72/25.08	71.29/30.42	80.69/40.86	101.99/55.25	96.8/63.7
– conf	0.0	48.73/19.3	34.01/13.44	59.2/32.17	89.77/53.36	24.68/7.86	89.61/45.69	60.05/22.57	71.27/28.52	76.71/35.08	85.44/42.53	84.0/51.3
– proxy	0.0	47.9/18.76	30.96/12.7	56.03/30.8	86.44/52.66	24.69/9.25	81.7/45.74	68.43/26.62	58.08/22.07	72.96/34.2	84.6/43.64	74.1/48.2
– sonar	0.0	48.82/19.4	33.88/13.19	61/33.03	87.08/52.03	27.75/8.91	87.6/45.09	62.72/24.81	71.92/30.21	76.9/35.99	82.49/39.72	78.9/47.3
UDW-32-16												
– nnll	0.0	45.55/17.88	29.02/11.52	54.66/29.79	96.01/57.91	20.19/6.13	84.03/39.91	63.45/23.8	52.63/18.43	71.98/33.71	84.57/43.01	79.6/48.0
– pesq	0.0	47.5/19.0	32.0/12.4	55.6/30.8	88.9/53.8	23.5/7.1	89.3/44.9	55.2/19.6	66.4/25.6	75.0/34.2	83.3/40.4	79.0/50.0
– entropy	0.0	45.68/18.18	29.86/12.07	53.88/32.97	88.92/68.55	21.55/6.32	91.78/84.74	54.32/23.38	66.47/33.82	73.62/47.04	83.91/56.95	74.7/54.1
– conf	0.0	47.07/18.29	31.12/11.62	54.04/28.14	85.53/51.41	22.78/6.68	81.01/39.84	65.69/25	69.83/29.36	55.32/20.11	86.59/44.52	72.9/43.6
– proxy	0.0	45.38/17.71	28.22/11.32	51.47/27.06	84.54/48.86	21.41/7.93	81.25/37.94	53.42/18.05	63.9/23.76	70.19/29.16	78.38/34.03	71.0/44.1
– sonar	0.0	44.74/17.38	27.35/10.35	52.72/28.08	82.33/49.95	21.05/6.29	80.58/38.85	52.29/17.53	63.17/22.67	67.71/27.87	75.42/32.57	65.0/37.7
UDW-16-16++												
– proxy	0.0	52.0/22.5	28.4/11.6	53.8/28.8	86.5/50.6	21.9/6.7	83.6/43.2	54.2/18.8	64.9/24.1	69.3/30.9	78.9/37.2	66.2/37.3
– sonar	0.0	45.5/17.7	29.9/11.4	53.9/28.8	83.7/50.1	22.8/7.0	82.6/40.6	55.4/19.6	65.0/24.4	70.2/29.6	77.5/33.9	65.7/36.0
UDW-32-16-++												
– proxy	0.0	44.9/17.5	26.6/10.5	50.4/25.8	84.4/47.1	20.7/6.1	81.6/39.0	51.5/17.0	62.6/23.0	66.5/26.6	77.6/33.8	62.1/33.6
– sonar	0.0	44.2/17.1	29.0/11.5	51.2/27.0	81.7/47.2	20.6/5.7	79.5/36.5	62.3/22.2	52.5/17.5	67.1/26.8	76.0/31.8	54.7/31.3

Table 14: WER/CER (↓) scores on orthographic transcription. Average is the mean score across all the evaluation sets. All distilled models are trained with a filtering threshold of 80. We report the score on the test split of each dataset. Best results are shown in **bold**. Second best results are underlined. We report the score on the test split of each dataset. Abbreviations. **W** - Whisper, **FT** - Finetuned, **M** - Medium, **L** - Large, **S** - Small, **DW** - Distill Whisper, **UDW** - Unsupervised Distill Whisper, **nll** - negative log likelihood, **conf** - confidence score.