Multilingual DistilWhisper: Efficient Distillation of Multi-Task Speech Models via Language-Specific Experts

Thomas Palmeira Ferraz[◇♣] **Marcely Zanon Boito**[◇] **Caroline Brun**[◇] **Vassilina Nikoulina**[◇]

[◇] NAVER LABS Europe, France
[♣] Télécom Paris, Institut Polytechnique de Paris, France
github.com/naver/multilingual-distilwhisper

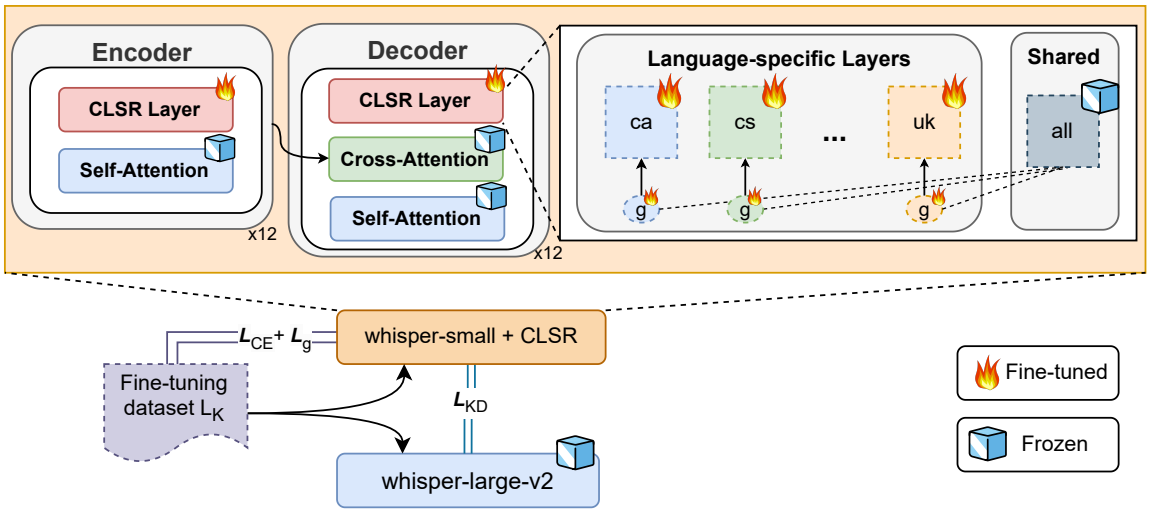

Figure 1: The Multilingual DistilWhisper architecture (top) and optimization framework (bottom). **Architecture:** We extend whisper-small by replacing its feed-forward network (FFN) modules with Conditional Language-Specific Routing (CLSR) modules in both encoder and decoder. Each CLSR module contains language-specific gates (g) that route tokens through either frozen multilingual representations initialized with previous FFN block (*shared*) or learnable language-specific modules (*LS*). **Training:** Our dual optimization combines ASR supervised fine-tuning, gate budget and knowledge distillation from frozen robust teacher (whisper-large-v2).

Abstract

Whisper is a multitask and multilingual speech model covering 99 languages. It yields commendable automatic speech recognition (ASR) results in a subset of its covered languages, but the model still underperforms on a non-negligible number of under-represented languages, a problem exacerbated in smaller model versions. In this work, we propose DistilWhisper, an approach able to bridge the performance gap in ASR for these languages while retaining the advantages of multitask and multilingual capabilities. Our approach involves two key strategies: lightweight modular ASR fine-tuning of whisper-small using language-specific experts, and knowledge distillation from whisper-large-v2. This dual approach allows us to effectively boost ASR performance while keeping the robustness inherited from the multitask and multilingual pre-training. Results demonstrate that our approach is more effective than standard fine-tuning or LoRA adapters, boosting performance in the targeted languages for both in- and out-of-domain test sets, while introducing only a negligible parameter overhead at inference.

Keywords: knowledge distillation, multitask speech processing, automatic speech recognition, multilingual speech processing, language experts

Corresponding author(s): [thomas.palmeira,marcely.zanon-boito]@naverlabs.com

1. Introduction

Whisper (Radford et al., 2023) is a popular multilingual and multitask speech model that is known for its robustness (i.e. invariant performance over different out-of-domain data) for automatic speech recognition (ASR) (Gandhi et al., 2022). This model covers 99 languages, and jointly trains on ASR, speech translation (many-to-English), language identification, and voice activity detection tasks. The original paper attributes this multitask training as a reason for the observed robustness of the model to out-of-domain data: compared to the English wav2vec 2.0 model (Baevski et al., 2020), Whisper performance seems to generalize better to unseen domains. Available in many sizes (from tiny to large-v2), Whisper exhibits an important gap in ASR performance between whisper-large-v2 (largest model) and whisper-small (second smallest model) on a large set of languages, including low-resource languages, but also many high- and mid-resource ones. This phenomenon in NLP is often referred as *curse of multilinguality*, where the performance drop due to the growing amount of covered languages can only be recovered via extensive model scaling (Arivazhagan et al., 2019, Conneau et al., 2020, Goyal et al., 2021). Such scaling comes with an important inference cost increase: for instance, whisper-large-v2 is 2-3 times slower than whisper-small.

A common approach to efficient inference is distilling knowledge from a large multilingual teacher model into a smaller model (Mohammadshahi et al., 2022, Sanh et al., 2020). However, applying knowledge distillation (KD) to whisper-large-v2, the best and largest Whisper model, presents a challenge. We would ideally need access to training data across all tasks and languages to preserve robustness, but such data is unavailable, making it challenging to maintain the model's out-of-domain generalization capabilities.

In another direction, Pfeiffer et al. (2022) and Pratap et al. (2023) have demonstrated that the *curse of multilinguality* can also be solved by equipping a moderately sized model with language-specific (LS) modules. Such architectures allow extending model parameters via extra modules when more languages are added into the model, thus maintaining consistent performance across languages, with no (or very low) extra computations at inference.

Inspired by these findings, we propose DistilWhisper, which extends whisper-small with language-specific feed-forward modules, that are used in parallel with the original feed-forward layers of the model. In order to preserve the robustness of the original model, DistilWhisper introduces the following extensions of previous works:

1. Following Zhang et al. (2021), we extend conditional language-specific routing (CLSR) modules with the gating mechanism that can route input representation either through the original feed-forward module or through newly learned LS feed-forward module;
2. When learning language-specific modules, we use whisper-large-v2 as a teacher model with the hypothesis that the KD loss should help reproduce the robustness of the larger Whisper model.

Through extensive experiments on a diverse set of languages we demonstrate the effectiveness of DistilWhisper compared to standard fine-tuning or LoRA adapters (Hu et al., 2021). Our lightweight ASR fine-tuning approach based on CLSR modules generalizes better than LoRA, and the introduction of KD further boosts results in both in- and out-of-domain test sets. We perform additional ablation studies showing our approach can cope with different amounts of training data. Finally, we demonstrate that the flexibility introduced by the gating mechanism equips DistilWhisper with an efficient adaptation approach, leveraging the language-specific modules only when those are relevant. We make available the models' weights[1] and code[2] developed in this work.

2. Background

2.1. State of the art for ASR

Current approaches for ASR mainly rely on the adaptation of pre-trained Transformer stacks learned through self-supervision (i.e. SSL models) on unlabeled audio data. Such pre-trained models vary on the usage of pretext tasks (Baevski et al., 2020, Chen et al., 2022, Hsu et al., 2021) and language coverage (Babu et al., 2022, Conneau et al., 2021, Pratap et al., 2023, Zhang et al., 2023). In contrast to this branch of research, the Whisper

[1] Weights available at: https://huggingface.co/collections/naver/multilingual-distilwhisper-6576ecae8d209fc6a767d9e7.
[2] Code available at: https://github.com/naver/multilingual-distilwhisper.

model relies on weak supervision, which means that the architecture is trained on weakly labeled data only (no self-supervision). Nonetheless, Radford et al. (2023) show that with sufficient amounts of data, the Weakly Supervised Whisper model reaches competitive results compared to monolingual and multilingual SSL models (Gandhi et al., 2022, Pratap et al., 2023).

2.2. Knowledge distillation

Knowledge distillation (KD) has been initially proposed by Hinton et al. (2015) to distill knowledge from ensemble of models into a single model for ASR. It has further been used to distill knowledge from a large teacher model into smaller student models (Mohammadshahi et al., 2022, Sanh et al., 2020, Shen et al., 2023). While original KD methods relied on minimization of KL-divergence between a teacher model and a student model, Wen et al. (2023) and Go et al. (2023) have recently shown that symmetric divergences, such as Jensen-Shannon (JS) divergence, suffer less from borderline behaviors and lead to better results on sequence level distillation.

2.3. Parameter-efficient Fine-tuning

Adapters are small lightweight modules which are commonly used in NLP to adapt pre-trained models to new tasks or domains. In speech-related tasks, adapter-based fine-tuning has been utilized for speech translation (Antonios et al., 2022, Gow-Smith et al., 2023, Le et al., 2021), and domain adaptation (Thomas et al., 2022, Tomanek et al., 2021), for which they exhibit a similar performance to standard fine-tuning, but with only a fraction of trainable parameters. We also find work on task-adaptation of Whisper (Feng and Narayanan, 2023, Radhakrishnan et al., 2023, Wang et al., 2023) using LoRA adapters. In contrast to adapters, in this work we introduce gated LS modules into Whisper, and propose a parameter-efficient KD approach that allows us to increase robustness to out-of-domain data.

3. DistilWhisper

With the goal of increasing performance for different languages in models of limited capacity, we propose the DistilWhisper approach: we plug conditional language-specific routing (CLSR) modules (Zhang et al., 2021) into a small Whisper (whisper-small), and optimize these modules jointly on ASR fine-tuning and KD from a larger Whisper (whisper-large-v2). Figure 1 presents our architecture, below we detail its key components.

3.1. Conditional Language-Specific Routing (CLSR) module

We extend CLSR modules for the first time to the speech domain. This module learns a hard binary gate $g(\cdot)$ for each input token by using its hidden embedding z^l. These decisions enable a layer to selectively guide information through either a Language-Specific path denoted as h^{lang} or a shared path referred to as h^{shared}, as in Eq 1:

$$\text{CLSR}(z^l) = g(z^l) \cdot h^{lang}(z^l) + (1 - g(z^l)) \cdot h^{shared}(z^l). \quad (1)$$

In contrast to the original CLSR, in this work we use language-specific gates as shown in Figure 1, instead of sharing them across languages. This allows us to train LS modules individually (i.e. in parallel), and then only load the relevant modules at inference. Moreover, our approach also differs from the original CLSR by the positioning: supported by Zhang et al. (2021) and Pfeiffer et al. (2022) works, we limit CLSR to the feed-forward, which we entirely replace with the CLSR module, reducing further the number of parameters.

Following Zhang et al. (2021), the gating mechanism is implemented as follows: each gate $g(.)$ consists of a two-layer bottleneck network, to which we add an increasing zero-mean Gaussian noise during training to facilitate discretization and enable gradient flow through the binary decisions. At inference time, we adopt hard gating, where the gate outputs are deterministically set to either 0 or 1 based on the learned routing decisions.

3.2. DistilWhisper approach

DistilWhisper approach is detailed in Figure 1. Our student is enriched with CLSR modules at each feed-forward for each language. These CLSR layers are initialized from the frozen weights of the corresponding feed-forward layer. At training time, for each language the model updates only the corresponding language-specific modules and gates. At inference time, the model loads the shared modules (multilingual) and the LS modules and gates for the languages of interest, resulting in a limited parameter overhead. We highlight that

the use of CLSR modules brings more flexibility to our architecture when compared to adapters, as it allows for routing at the token-level. This makes this approach more capable of leveraging pre-existing knowledge (shared frozen module) via LS gating activation.

3.3. DistilWhisper optimization

Following Zhang et al. (2021), when learning CLSR module parameters, in addition to standard cross-entropy loss $\mathcal{L}_{\text{CE}}$, we employ a gate budget loss $\mathcal{L}_g$ (Eq 3) to balance models' usage of LS and language-shared modules.

The gate budget loss relies on the gate $g(.)$ activation values for a pair (audio, text) (X, Y) in a batch $\mathcal{B}$, which is expressed by:

$$\mathcal{G}_{(X,Y)} = \sum_{x \in X} \sum_{m \in \mathcal{M}_{\text{enc}}} g_m(x) + \sum_{y \in Y} \sum_{m \in \mathcal{M}_{\text{dec}}} g_m(y) \tag{2}$$

where $\mathcal{M}_{\text{enc}}$ and $\mathcal{M}_{\text{dec}}$ are respectively the encoders and decoders layers, and $g_m(.) = 1$ when LS module is selected, or 0 otherwise. The average of this gate usage is constrained to a budget b (Eq 3):

$$\mathcal{L}_g = \left| \frac{\sum_{(X,Y) \in \mathcal{B}} \mathcal{G}_{(X,Y)}}{\sum_{(X,Y) \in \mathcal{B}} (|X||\mathcal{M}_{\text{enc}}| + |Y||\mathcal{M}_{\text{dec}}|)} - b \right|, \tag{3}$$

For KD, following Wen et al. (2023) and Go et al. (2023), we use JS divergence, whose loss is detailed in Eq 4:

$$\mathcal{L}_{\text{KD}} = \frac{1}{2} \mathbb{E}_{\boldsymbol{Y} \sim p} \left[\log \frac{p(\boldsymbol{Y})}{m(\boldsymbol{Y})} \right] + \frac{1}{2} \mathbb{E}_{\boldsymbol{Y}' \sim q_\theta} \left[\log \frac{q_\theta(\boldsymbol{Y}')}{m(\boldsymbol{Y}')} \right] \tag{4}$$

where p is the teacher distribution, q_θ is the student distribution, $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ are sampled from the teacher's and student's distributions and compared with their average $m(\cdot) = \frac{1}{2} p(\cdot) + \frac{1}{2} q_\theta(\cdot)$.

Thus, CLSR modules parameters are learned to minimize final loss expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{g}} + \alpha \mathcal{L}_{\text{KD}}. \tag{5}$$

4. Experimental Setup

4.1. Datasets

We downsample the train and validation sets of the CommonVoice 13.0 (CV-13) dataset (Ardila et al., 2020), using equal amounts of training data for each selected language: 10k utterances for training (approx. 14 h), 1k for validation. Data selection depends on the amount of up-votes utterances received by annotators. We do not downsample the test set. The FLEURS (Conneau et al., 2023) dataset is used for out-of-domain evaluation, as it provides both a good language overlap with CV-13, and an effective out-of-domain setting for ASR evaluation. For instance, average number of tokens per sample for CV-13 is 36, and 97 for FLEURS.

4.2. Language Selection

We consider all Whisper languages with a WER gap of more than 11 between large and small models on CV-13. We then narrow this list considering: 1) minimum amount of utterances (10k); 2) overlap with the FLEURS dataset. The final list of languages is: Catalan (ca), Czech (cs), Galician (gl), Hungarian (hu), Polish (pl), Thai (th), Tamil (ta) and Ukrainian (uk).[3] These languages encompass 5 language sub-families and vary widely in terms of coverage in the available Whisper training data, spanning from 4,300 h (pl) to just 9 h (gl).

4.3. Models

We compare our approach to both whisper-small (pre-trained student) and whisper-large-v2 (teacher) models, as well as two approaches of fine-tuning (FT) for the student: standard fine-tuning (all weights are updated), and LoRA adaptation on top of the feed-forward layer. Finally, we also investigate the impact of the CLSR layer without the use of KD (CLSR-FT), decoupling the effect of KD from the flexibility offered by the routing mechanism on the consequent robustness of the model.

[3] Although Arabic would also qualify considering our criteria, we find that the dialect from FLEURS differs from the ones present on CV-13.

FLEURS (out-of-domain)										
Model	**# params**	**avg**	**ca**	**cs**	**gl**	**hu**	**pl**	**ta**	**th**	**uk**
whisper large-v2	1.5 B	12.5	5.6	14.3	16.6	17.9	5.9	19.3	12.2	8.1
whisper-small	244 M	28.3	**14.6**	40.4	32.7	43.0	**16.7**	36.0	22.8	**20.5**
whisper-small + FT	244 M	23.3±0.06	15.5	31.0	16.9	**36.7**	22.0	22.7	15.6	25.9
whisper-small + LoRA-FT	379 M	24.9±0.07	17.6	36.9	18.2	41.6	25.9	**15.2**	**11.7**	31.8
whisper-small + CLSR-FT	369 M	23.4±0.19	15.7	30.5	17.2	36.9	22.8	22.7	15.6	25.8
DistilWhisper	369 M	**22.8**±0.21	**15.3**	**30.2**	**16.7**	36.9	**21.4**	21.8	15.1	**24.9**

Common Voice 13.0 (in-domain for FT only)										
Model	**# params**	**avg**	**ca**	**cs**	**gl**	**hu**	**pl**	**ta**	**th**	**uk**
whisper large-v2	1.5 B	14.9	16.9	14.4	18.9	18.7	8.0	17.3	9.2	15.5
whisper-small	244 M	31.4	30.1	38.4	35.5	45.6	18.6	30.0	20.3	32.3
whisper-small + FT	244 M	16.3±0.09	**13.7**	20.5	**11.3**	24.1	16.3	13.6	7.4	23.4
whisper-small + LoRA-FT	379 M	18.2±0.02	14.0	23.7	12.7	28.0	21.2	**12.0**	7.9	26.4
whisper-small + CLSR-FT	369 M	16.3±0.08	14.1	20.3	11.6	24.3	16.1	13.3	7.4	23.4
DistilWhisper	369 M	**16.0**±0.04	**13.8**	**20.0**	11.8	**24.0**	**15.9**	12.6	**7.2**	**23.1**

Table 1: WER (↓) for both in-domain and out-of-domain evaluation settings. **Top panel: FLEURS** (out-of-domain). **Bottom panel: Common Voice 13.0 (CV-13)** (in-domain only for fine-tuned (FT) models) - Results of pre-trained models on CV-13 are shown in gray, as they are not directly comparable to FT models. Each panel reports the number of active parameters, dataset average WER (mean ±std) and per-language WER. Rows are grouped as *baselines* (top), *adaptation approaches* (middle), and *our method* (bottom), with FT-only (CLSR-FT) or with distillation (DistilWhisper). Best results for fine-tuned whisper-small as well as cases where pre-FT models performed better are shown in **bold**.

4.4. Implementation details

We train all models using the Transformers library (Wolf et al., 2020), using the pre-trained weights for whisper-small and whisper-large-v2 available on HuggingFace[4]. All models are trained for 10 epochs using a learning rate of 10^{-4} with linear decay, one epoch of warm-up, batch size of 16, and label smoothing factor of 0.1. For LoRA, we use the hyperparameters proposed by Wang et al. (2023). For CLSR training, we set the gate budget $b = 0.5$ and skip-gate probability $s = 0.2$. For knowledge distillation (KD), we employ the JS divergence with temperature $\tau = 1$. The full learning objective for our experiments is given by:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{g}} + 2\mathcal{L}_{\text{KD}} \tag{6}$$

We report normalized WER using the Whisper normalization procedure, with a slight modification to avoid splitting numbers and latin-scripted text into individual characters for languages that do not use space delimitation (e.g., Thai). In all cases, the best model is chosen based on WER on the down-sampled CV-13 validation set.

5. Results

We conduct training for each setting using three distinct seeds and present the average scores. Table 1 presents our results. The top portion presents whisper-large-v2 (upper bound) and whisper-small (lower bound) pre-trained scores. The middle portion presents standard fine-tuning (FT) and LoRA adaptation at the feed-forward layers (LoRA-FT). Our results are presented in the bottom: CLSR-FT corresponds to the setting without $\mathcal{L}_{\text{KD}}$, while DistilWhisper is the complete setting in which both CLSR and KD losses are leveraged.

[4]Models weights are available at: https://huggingface.co/openai/whisper-small and https://huggingface.co/openai/whisper-large-v2

	Train size	FLEURS avg	CV-13 avg	FLEURS ca	FLEURS ta	FLEURS th	CV-13 ca	CV-13 ta	CV-13 th
whisper-small+CLSR-FT	3k	20.5±0.17	15.0±0.07	17.9	25.6	18.0	19.0	16.4	9.8
DistilWhisper	3k	**20.2**±0.13	**14.6**±0.08	**17.4**	25.5	**17.7**	**18.7**	**15.7**	**9.6**
whisper-small+CLSR-FT	10k	18.0±0.25	11.6±0.01	15.7	22.7	15.6	14.1	13.3	7.4
DistilWhisper	10k	**17.4**±0.13	**11.2**±0.08	**15.3**	**21.8**	**15.1**	**13.8**	**12.6**	**7.2**
whisper-small+CLSR-FT	28k	15.7±0.15	9.5±0.13	13.5	19.8	13.9	11.3	11.3	6.0
DistilWhisper	28k	**15.5**±0.03	**9.3**±0.06	**13.3**	**19.3**	**13.7**	11.3	**11.0**	**5.7**

Table 2: Average WER (↓) for models trained with different training data sizes (3k, 10k, and 28k utterances). Results are reported on in-domain (**CV-13**) and out-of-domain (**FLEURS**) test sets, including per-language scores for the languages with at least 28k utterances at training split (ca, ta, and th). Best results for each training size are shown in **bold**.

5.1. DistilWhisper vs. other adaptation approaches

For whisper-small, we observe that both FT and LoRA-FT approaches (middle portion of Table 1) are able to improve performance on both in- and out-of-domain test sets. However, FT achieves this improvement at the cost of language specialization, reducing performance in other languages. In contrast to that, LoRA-FT is a light adaptation technique that does not modify the pre-trained representation. This method increases performance on both in-domain (avg -13.1) and out-of-domain (avg -3.5) test sets compared to whisper-small. DistilWhisper further improves performance over whisper-small (avg -15.3) and LoRA-FT (avg -2.2) for in-domain data. It also presents better out-of-domain adaptation capabilities compared to LoRA-FT (avg -2.1).

5.2. Impact of knowledge distillation

We observe that DistilWhisper on average outperforms all other adaptation approaches (FT, LoRA-FT) for in- and out-of-domain test sets (bottom portion of Table 1). Comparing our models (CLSR-FT and DistilWhisper), we observe that the version with KD (DistilWhisper) exhibits a slight increase in average in-domain performance (-0.3). In out-of-domain settings, this model consistently outperforms CLSR-FT across all languages (avg -0.6), which confirms our initial hypothesis that the KD loss leverages the robustness from the teacher into the final model. Overall, these results highlight the effectiveness of our proposed architecture: we reduce the out-of-domain performance gap between whisper-large-v2 and whisper-small by 35.2% (avg -5.5) with a parameter overhead at inference time of only 10% (25 M).

5.3. Effect of training data size

We now show the effectiveness of our approach on lower and higher data resource settings. For this, we select a subset of languages for which we find more training data available on CV-13 (ca, th, ta). Table 2 presents results for our approach in low (3k utterances; ~4 h), and higher-resource settings (28k utterances; ~40 h), compared to the 10k results from Table 1. We observe that, as expected, increasing the amount of trainable examples leads to superior ASR performance for both approaches, with the leveraging of KD (DistilWhisper) being consistently superior to CLSR-FT. For the 28k setup (ca, th, ta), we reduce the out-of-domain WER gap between whisper-large-v2 and whisper-small by 75% (from 12 to 3 WER).[5] For the 3k setup, we reduce the WER gap by 35.8% using only 4 h of training data. This implies that our approach has the potential to improve ASR performance across low-resource languages for which less training data is available.

5.4. Gate Activation Analysis

To better understand how the model uses routing mechanism, we plot gate activation statistics for both CLSR-FT and DistilWhisper in Figure 2. We observe that the models tend to rely more on the new language-specific modules in out-of-domain settings (FLEURS vs CV-13), which could be attributed to the greater complexity and larger size of sentences in FLEURS. Also, as expected, increasing the training data size leads to more reliable LS

[5] whisper-large-v2 and whisper-small avg FLEURS scores for ca, th, ta are respectively 12.5 and 24.5.

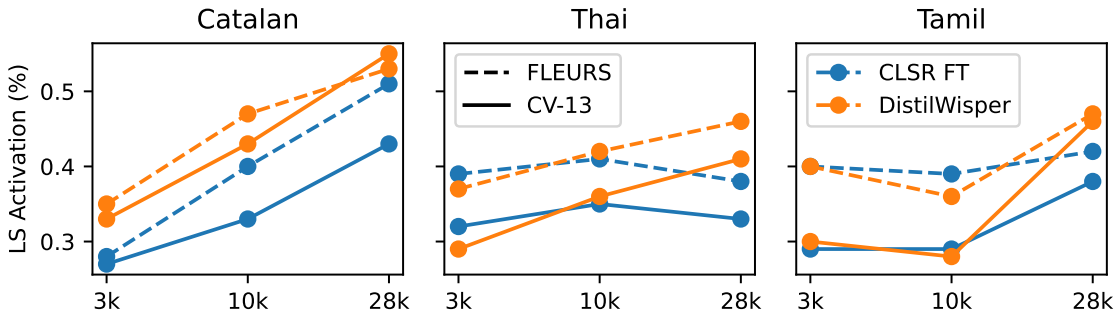

Figure 2: Ratio of language-specific (LS) expert activations selected by each model across different training data sizes (x-axis), evaluation domains — in-domain (**CV-13**, solid lines) and out-of-domain (**FLEURS**, dashed lines) — and languages. We compare the direct baseline (**CLSR-FT**) and with our knowledge-distilled CLSR-FT+KD (**DistilWhisper**).

modules, and therefore higher LS usage. The only exception for this is Thai at the 28k setup, and this might be due to dataset quality and requires further investigation. When comparing the 3 languages, we observe that Catalan exhibits a higher reliance on LS routes, which could also be related to the data quality for this language in CV-13. Finally, we observe that for languages with a weaker teacher (Thai, Tamil) the model may receive contradictory signals at lower-resource settings (3k, 10k), leading to less LS routing usage with KD. However, in the higher resource setting (28k), KD usage leads systematically to more reliable LS module and therefore higher LS routing.

6. Conclusion

We presented DistilWhisper, a parameter-efficient distillation approach that boosts performance of whisper-small by leveraging the robustness from the whisper-large-v2 into a smaller model, while preserving its multilingual capabilities. This is done by adding language-specific gated modules, and by jointly optimizing ASR fine-tuning and KD losses. Compared to LoRA adapters, and across eight languages, we are able to consistently improve performance in both in- and out-of-domain test sets, while adding only a negligible number of parameters at inference time. We believe this architecture makes Whisper models more accessible to researchers and practitioners, as it boosts the performance of a low-inference cost model by 35.2% using only 14 h of training data.

Acknowledgements

This research was partially supported by the French *Agence Nationale de la Recherche, ANR* as part of the project **Diké - Bias, Fairness and Ethics of Compressed Language Models**, under grant number ANR-21-CE23-0026-02.

Co-funded by the European Union

This work was also co-funded by the European Union's Horizon Europe Research and Innovation programme through the project **UTTER – Unified Transcription and Translation for Extended Reality**[a] under Grant Agreement No. 101070631.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

[a]For more information please visit: https://he-utter.eu/.

References

Anastasopoulos Antonios, Barrault Loc, Luisa Bentivogli, Marcely Zanon Boito, Bojar Ondřej, Roldano Cattoni, Currey Anna, Dinu Georgiana, Duh Kevin, Elbayad Maha, et al. Findings of the iwslt 2022 evaluation campaign. In *Proc. of the 19th Int. Conf. on Spoken Language Translation (IWSLT 2022)*. ACL, 2022. 3

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.520. 4

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively multilingual neural machine translation in the wild: Findings and challenges, 2019. 2

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. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282, 2022. doi: 10.21437/Interspeech.2022-143. 2

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 2020. 2

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022. 2

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020. 2

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430, 2021. doi: 10.21437/Interspeech.2021-329. 2

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE SLT*, pages 798–805. IEEE, 2023. 4

Tiantian Feng and Shrikanth Narayanan. Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models. *arXiv preprint arXiv:2306.05350*, 2023. 3

Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush. Esb: A benchmark for multi-domain end-to-end speech recognition. *arXiv preprint arXiv:2210.13352*, 2022. 2, 3

Dongyoung Go, Tomasz Korbak, Germàn Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. https://proceedings.mlr.press/v202/go23a.html. 3, 4, 10

Edward Gow-Smith, Alexandre Berard, Marcely Zanon Boito, and Ioan Calapodescu. NAVER LABS Europe's multilingual speech translation systems for the IWSLT 2023 low-resource track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 144–158, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.10. https://aclanthology.org/2023.iwslt-1.10. 3

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. Larger-scale transformers for multilingual masked language modeling, 2021. 2

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. http://arxiv.org/abs/1503.02531. 3

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 2

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 2

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. Lightweight adapter tuning for multilingual speech translation. In *Proc. of the 59th Annual Meeting of the ACL and the 11th Int. Joint Conf. on Natural Language Processing*, 2021. 3

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.18653/v1/2022.emnlp-main.571. https://aclanthology.org/2022.emnlp-main.571. 2, 3

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.255. https://aclanthology.org/2022.naacl-main.255. 2, 3

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*, 2023. 2, 3

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 2, 3

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Narsis A Kiani, David Gomez-Cabrero, and Jesper N Tegner. A parameter-efficient learning approach to arabic dialect identification with pre-trained general-purpose speech model. *arXiv preprint arXiv:2305.11244*, 2023. 3

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. 2, 3

Zhijie Shen, Wu Guo, and Bin Gu. Language-universal adapter learning with knowledge distillation for end-to-end multilingual speech recognition. *arXiv preprint arXiv:2303.01249*, 2023. 3

Bethan Thomas, Samuel Kessler, and Salah Karout. Efficient adapter transfer of self-supervised speech models for automatic speech recognition. In *ICASSP 2022-2022 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022. 3

Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vaillancourt, and Fadi Biadsy. Residual adapters for parameter-efficient asr adaptation to atypical and accented speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6751–6760, 2021. 3

Minghan Wang, Yinglu Li, Jiaxin Guo, Xiaosong Qiao, Zongyao Li, Hengchao Shang, Daimeng Wei, Shimin Tao, Min Zhang, and Hao Yang. WhiSLU: End-to-End Spoken Language Understanding with Whisper. In *Proc. INTERSPEECH 2023*, pages 770–774, 2023. doi: 10.21437/Interspeech.2023-1505. 3, 5

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In *ACL*, July 2023. doi: 10.18653/v1/2023.acl-long.605. https://aclanthology.org/2023.acl-long.605. 3, 4, 10

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. 5

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*, 2021. https://openreview.net/forum?id=Wj4ODo0uyCF. 2, 3, 4

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, 2023. 2

A. Effect of Temperature and Distillation Loss

We explore two temperature settings (1 and 3), and compare Jensen-Shannon (JS) loss with the traditional Kullback-Leibler (KL) in the 28k setup. Results are presented for validation (Table 3) and test sets (Table 4).

We observe overall better validation scores using JS, a trend that is confirmed by our test results. Regarding temperature, JS equipped with $\tau = 3$ presents inferior validation scores, but superior CV-13 (in-domain) test

	FLEURS	CV-13	FLEURS			CV-13		
	avg	**avg**	**ca**	**ta**	**th**	**ca**	**ta**	**th**
JS w/ $\tau = 1$	**14.8**	8.8	**12.7**	10.7	12.9	**5.9**	**18.7**	9.9
JS w/ $\tau = 3$	15.4	**8.5**	14.5	**10.0**	**12.7**	6.0	18.9	**9.4**
KL w/ $\tau = 1$	15.6	10.2	15.1	13.8	12.8	6.3	18.8	10.5
KL w/ $\tau = 3$	15.7	8.6	15.2	10.5	13.0	**5.9**	18.8	9.5

Table 3: WER (↓) for in- and out-of-domain **validation sets** for DistilWhisper equipped with JS and KL losses, with different temperatures (τ). Best results in **bold**.

	FLEURS	CV-13	FLEURS			CV-13		
	avg	**avg**	**ca**	**ta**	**th**	**ca**	**ta**	**th**
JS w/ $\tau = 1$	**15.4**	**9.3**	**13.1**	19.2	14.0	**11.3**	**10.9**	**5.7**
JS w/ $\tau = 3$	16.3	9.7	14.8	20.1	14.1	11.8	11.3	5.9
KL w/ $\tau = 1$	15.6	10.8	14.6	**18.7**	**13.3**	14.9	11.3	6.2
KL w/ $\tau = 3$	16.5	9.7	15.8	19.8	14.0	12.2	11.1	5.9

Table 4: WER (↓) for in- and out-of-domain **test sets** for DistilWhisper equipped with JS and KL losses, with different temperatures (τ). Best results in **bold**.

scores. In this work, we selected our models based on their performance on the validation set of CV-13, which is why we report results for JS with $\tau = 1$ only.

B. Stability of the Methods

We investigate robustness by repeating each approach — fine-tuned only (CLSR-FT), DistilWhisper with JS ($\tau = 1$), and DistilWhisper with KL ($\tau = 1$)—over three random seeds in the 28k setup. Table 5 reports per-language results and dataset averages as mean±std WER (↓) across seeds.

With a 95% confidence intervals on the macro-averaged scores across seeds, DistilWhisper with JS is statistically superior to the fine-tuning-only baseline, while KL is comparable to JS in mean but exhibits higher variability on some splits, with is probably related to the borderline effects reported in recent research (Go et al., 2023,Wen et al., 2023).[6] In short, JS (τ=1) offers the best trade-off between accuracy and stability across languages and domains, confirming that distillation improves over CLSR-FT with lower seed sensitivity.

[6]95% confidence intervals over seed means: CLSR-FT [15.54, 15.87], JS [15.42, 15.49], KL [15.24, 15.57].

Domain	Split	**CLSR-FT**				**JS** ($\tau = 1$)				**KL** ($\tau = 1$)			
		s1	s2	s3	**avg**	s1	s2	s3	**avg**	s1	s2	s3	**avg**
Catalan (ca)													
FLEURS	Val	13.1	13.9	13.6	**13.5**±0.4	12.7	12.6	13.1	**12.8**±0.3	15.1	14.8	14.7	**14.8**±0.2
	Test	13.5	13.3	13.6	**13.5**±0.2	13.1	13.6	13.3	**13.3**±0.2	14.6	14.6	14.3	**14.5**±0.2
CV-13	Val	9.6	9.6	9.0	**9.4**±0.4	10.7	9.3	8.8	**9.6**±1.0	13.8	13.8	13.7	**13.8**±0.1
	Test	11.4	11.3	11.2	**11.3**±0.1	11.3	11.3	11.3	**11.3**±0.03	14.9	14.6	15.0	**14.8**±0.2
Tamil (ta)													
FLEURS	Val	19.2	18.9	19.0	**19.1**±0.2	18.7	19.1	18.5	**18.8**±0.3	18.1	17.7	17.7	**17.8**±0.2
	Test	19.6	19.8	19.9	**19.8**±0.1	19.2	19.2	19.6	**19.3**±0.2	18.5	18.3	17.9	**18.2**±0.3
CV-13	Val	9.3	9.3	9.0	**9.2**±0.2	9.9	9.3	8.9	**9.4**±0.5	10.5	10.4	11.4	**10.8**±0.5
	Test	11.3	11.4	11.2	**11.3**±0.1	10.9	10.9	11.2	**11.0**±0.2	11.4	11.2	9.8	**10.8**±0.9
Thai (th)													
FLEURS	Val	13.8	13.6	13.5	**13.6**±0.2	12.9	12.9	12.9	**12.9**±0.02	12.8	13.0	13.0	**12.9**±0.1
	Test	13.8	13.8	14.2	**13.9**±0.2	14.0	13.6	13.6	**13.7**±0.2	13.3	13.6	13.6	**13.5**±0.1
CV-13	Val	6.4	6.0	5.7	**6.0**±0.3	5.9	5.8	5.9	**5.9**±0.1	6.3	6.5	6.3	**6.4**±0.1
	Test	6.2	6.0	5.9	**6.0**±0.2	5.7	5.8	5.8	**5.7**±0.1	6.2	6.3	6.2	**6.2**±0.1
Averages													
FLEURS	Val	15.4	15.5	15.4	**15.4**±0.04	14.8	14.9	14.8	**14.8**±0.1	15.3	15.1	15.1	**15.2**±0.1
	Test	15.6	15.6	15.9	**15.7**±0.2	15.4	15.5	15.5	**15.5**±0.03	15.5	15.5	15.2	**15.4**±0.2
CV-13	Val	8.4	8.3	7.9	**8.2**±0.3	8.8	8.1	7.9	**8.3**±0.5	10.2	10.2	10.5	**10.3**±0.2
	Test	9.7	9.6	9.4	**9.5**±0.1	9.3	9.4	9.4	**9.3**±0.1	10.8	10.7	10.3	**10.6**±0.3

Table 5: Val/Test WER (↓) for in-domain (CV-13) and out-of-domain (FLEURS) across three methods: fine-tuning-only version (CLSR-FT), and distillation versions (DISTILWHISPER) with either Jensen-Shannon (JS) loss, or traditional Kullback-Leibler (KL) one, both with temperature $\tau = 1$. All experiments are repeated with three seeds (s1–s3) and we also report in **bold** the **mean**±std. Rows are organized by per-language scores and dataset averages.