PSEUDO-LABELING FOR MASSIVELY MULTILINGUAL SPEECH RECOGNITION

Loren Lugosch[1*], *Tatiana Likhomanenko*[2†], *Gabriel Synnaeve*[2], *Ronan Collobert*[2†]

[1]McGill University / Mila, [2]Facebook AI Research

ABSTRACT

Semi-supervised learning through pseudo-labeling has become a staple of state-of-the-art monolingual speech recognition systems. In this work, we extend pseudo-labeling to massively multilingual speech recognition with 60 languages. We propose a simple pseudo-labeling recipe that works well even with low-resource languages: train a supervised multilingual model, fine-tune it with semi-supervised learning on a target language, generate pseudo-labels for that language, and train a final model using pseudo-labels for all languages, either from scratch or by fine-tuning. Experiments on the labeled Common Voice and unlabeled VoxPopuli datasets show that our recipe can yield a model with better performance for many languages that also transfers well to LibriSpeech.

***Index Terms*—** speech recognition, massively multilingual models, semi-supervised learning, pseudo-labeling

1. INTRODUCTION

One of the long-term goals of automatic speech recognition (ASR) research is a single system that can transcribe speech in any language [1, 2]. Such a multilingual system would be simpler to maintain than a collection of monolingual models, enable users to comfortably speak any language without needing to tell the system which language to expect in advance, and share knowledge between all languages for improved performance.

A key ingredient of modern state-of-the-art monolingual ASR missing from current multilingual models is *pseudo-labeling* [3], a technique for harnessing unlabeled datasets that has recently begun consistently yielding performance gains even for ASR tasks with large labeled datasets like LibriSpeech [4, 5, 6]. In pseudo-labeling, a model trained on a labeled dataset is used to generate labels for an unlabeled dataset, and those pseudo-labels (PLs) are then used to train a model. Many variants of pseudo-labeling exist: for instance, the same model used to generate PLs can also be trained on those PLs [7, 8, 9], or PLs generated by a teacher model can be used to train a new student model [4, 6, 10, 11].

In this work, we go beyond the monolingual setting and demonstrate the use of pseudo-labeling to improve a massively multilingual speech recognizer trained on all 60 languages of the Common Voice dataset [12] simultaneously. First, we show that self-training on all unlabeled data in the multilingual VoxPopuli dataset [13] at once tends to produce poor PLs for low-resource languages, and instead propose a simple recipe (Fig. 1) in which the model is first fine-tuned for a particular language before pseudo-labeling. Next, we compare a number of methods for training with the generated PLs, and find that training a larger model from scratch on all labeled and pseudo-labeled data works best. Finally, we show that the use of pseudo-labeled data improves out-of-domain generalization through experiments on LibriSpeech [14]. Unlike much previous work on this topic, our experiments use only open-source data, and we release our code and models.[1]

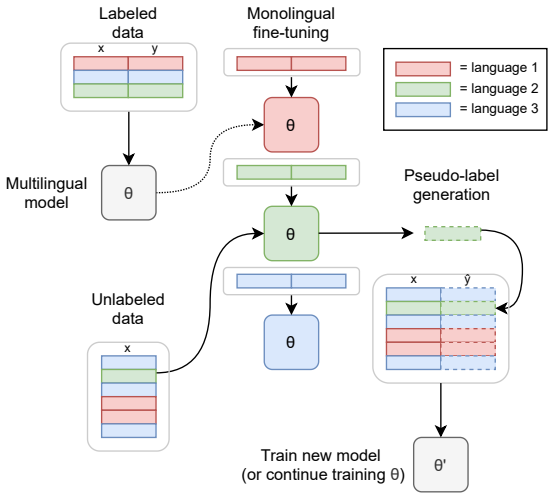

Fig. 1. Illustration of our method: to produce better pseudo-labels for a given language, we first fine-tune the multilingual model on that language.

2. MODEL

The model used in our experiments is identical to the neural network used for LibriSpeech in [9], except for the output layer(s). The output of the encoder is fed to a CTC [15] head and a language identification (LID) head. The CTC head

*Work done during an internship at FAIR. †Currently at Apple.

[1]https://github.com/flashlight/wav2letter/blob/49087d575ddf77aa5a99a01fee980fc00e92c802/recipes/mling_pl/README.md

is a linear layer with 8065 outputs: one for each character (most of which are Chinese characters), including punctuation, space, and the CTC <blank> symbol. The CTC head is shared across all languages: it is a "joint" multilingual model, using the terminology of [2]. The LID head is a linear layer with 60 outputs (one per language), followed by mean-pooling to aggregate the variable-length sequence of output vectors into a single vector of logits. The LID head outputs are only used during training: during inference, standard decoding algorithms can be applied to the CTC head outputs. The model is implemented and trained using Flashlight [16].

While we do not perform explicit empirical comparisons with other multilingual models in the literature (as the focus of this work is on pseudo-labeling), it is worth noting that our model is significantly simpler than existing multilingual models, forgoing the use of language- or language-family-specific parameters, decoders, and tokenizers. We are not the first to use an encoder-only CTC architecture for multilingual ASR [17, 18, 19], but we believe we are the first to demonstrate this for *massively* multilingual end-to-end ASR. Previous work on this topic [20, 21, 22, 23, 24] has instead used more sophisticated sequence transduction models with autoregressive decoder networks [25, 26, 27, 28], citing the flaw of CTC's conditional independence assumption. In practice, CTC models implemented using modern neural network architectures are able to learn strong implicit language models [4, 9] and achieve state-of-the-art results for the low-resource setting [29, 9]. For those reasons, we focus on CTC models in this paper.

3. DATA

The model is trained using the December 2020 release (6.1) of Common Voice (CV) [12], which has 3.6k hours of training data. CV is a continuously growing multilingual speech dataset recorded online by volunteer speakers. The 60 constituent languages vary greatly in the amount of available data: 7 languages have more than 100h of data, and 10 languages have less than 1h of data. We do not remove punctuation and capitalization from the CV transcripts, as this makes it easier to replicate our setup[2] and learning speed was not noticeably impacted. We downsample all audio to 16kHz.

In addition to CV, we use VoxPopuli (VP) [13], a very large scale (384k hours) unlabeled multilingual dataset of European languages. The dataset is split into 23 languages. 19 of the 23 VP languages are in CV (Czech, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovenian, and Swedish): we use only those 19 languages for semi-supervised learning.

[2]While there have been attempts to standardize the formatting of transcripts for Common Voice for English [30], most reported results use an ad-hoc normalization scheme, and so cannot readily be compared.

4. SUPERVISED TRAINING

We train supervised models on CV for ∼500k updates. The hyperparameters and training procedure are identical to those used in [9], except we use 2 SpecAugment [31] time masks instead of 10 (using 10 masks was found to cover too much of the shorter CV audio), and the learning rate is halved just once, at 250k updates. We do not use the language balancing technique of [21, 2] to sample languages evenly (which we found easily overfit to the low-resource languages), or curriculum learning as in [2]. In addition to the base model (275M params), we also train larger models (1.06B params) by doubling the feedforward and self-attention dimensions.

Following [32], we add an LID loss, so that the loss ℓ used for training is $\ell = \ell_{\text{CTC}} + \gamma \cdot \ell_{\text{LID}}$, where ℓ_{CTC} represents the CTC loss, ℓ_{LID} represents the LID loss (the cross-entropy between the LID head outputs and the one-hot language label for a given utterance), and γ is a hyperparameter. We trained models on CV with $\gamma \in \{0, 0.1, 1, 10\}$: $\gamma = 1$ yielded the best results, with 2.6% absolute improvement in average validation character error rate (CER) over the baseline with $\gamma = 0$ (no LID), using greedy decoding.

5. SEMI-SUPERVISED TRAINING

To train on the unlabeled data in VP, we use slimIPL [9], an iterative approach in which a model is trained for a number of updates on labeled data, followed by continuous training using labeled data and pseudo-labeled data stored in a dynamic cache which is periodically updated with pseudo-labels (PLs) re-generated by the current model state using greedy decoding without an external language model (LM). We use a cache size of 1000, replacement probability 0.1, and $\lambda = 10$ (ratio of unlabeled batches to labeled batches).

5.1. Fine-tuning before pseudo-labeling

The simplest way to perform semi-supervised learning would be to pool the unlabeled data for all languages, as we do for the labeled data, and run slimIPL. We found that doing so led to poor PLs for low-resource languages, such as Greek, which has only 2.75h of training data (see top of Fig. 2 — the transcript has a mix of Greek and Latin characters).

basic:	αναfeoρisi του κιακτίμανου sengen, μεταγιακρατί, critερία ποsμο πisama καίγατisυποpψήfiεs χορες, ενοτοsenginsiζitiτe διακί βervitika, τα proclimaταθas neχήsounaifista, γατοθα πτέπι παδiadramatisμε, μεγαliteroorol δίτε τησίζiti si γατi Rumania, κhe tη vulgaria to, efricaicoκi propecocino volie caneoti, boruse eni.
finetuned:	αναθεώρηση του κι κτημένου σέγγαν μεταϊδιακκρητήρια που χρισμοπίσαμε και για τις υποψήφιες χώρες ενώ το σέγγαν συζητήτε διακυβερνητικά τα προλήματαθα συνεχήσου να ηφήστάντα, γι αυτό θα πρέπει να διατραματήσουμε μεγαλύτερορόλο, δίτε τη σηζήτηση γιατη ρουμανία και τη βουλγαρίατο ευρωκαικόκικρωπακό κινοβούλιο έκανε ότηπορούσεεγή

Fig. 2. Pseudo-labels for an utterance from the Greek subset of VP with basic slimIPL (top) or with slimIPL after monolingual fine-tuning (bottom). Red letters are Latin characters.

Table 1. (Semi-)supervised learning results with slimIPL for the CV Greek data given different training sets.

Labeled	Unlabeled	Valid CER	Test CER
CV All	–	53.2	47.8
CV Greek	–	30.6	33.6
CV Greek	VP Greek	23.9	25.1
CV Greek	VP English[3]	24.3	28.4
CV All → CV Greek	–	9.9	9.6
CV All → CV Greek	VP Greek	8.7	8.5

Instead, to produce PLs for a VP language, we first fine-tune the trained multilingual model by training only on CV data for that language for 10k updates, and then run slimIPL using the corresponding VP data (bottom of Fig. 2). The same effect could also be achieved by generating PLs using a monolingual model, but our proposed approach yields better results by taking advantage of multi-task learning (Table 1).

After training slimIPL models for all 19 languages in CV ∩ VP, we generate a final set of PLs for all unlabeled VP utterances using the appropriate slimIPL models. We filter out all utterances for which the PL length is 0 or >630 (maximum label length supported by the CTC loss implementation). The PLs for all languages can then be pooled and used either by continuing training the non-fine-tuned multilingual model checkpoint with all available CV and VP data, or by training a new model on that data from scratch.

5.2. Avoiding collapse: cropping warmup period

Another difficulty arose from the fact that the utterances of VP (average duration of 30s) are much longer than those of CV (average duration of 5.3s). The model trained only on CV generates mostly empty transcripts for VP, a commonly observed failure mode for out-of-domain audio or utterances longer than those observed during training [8, 33, 34]. Semi-supervised learning failed as a result, usually collapsing to generating all blanks even for the labeled data. To acclimate the model to the longer VP utterances, we use a warmup period of 10k updates during which we crop unlabeled audio into 10s segments before running the acoustic model, then stitch the resulting logit sequences back together and decode to obtain PLs. The model is then trained on the original uncropped utterance using those PLs.

6. RESULTS

Table 2 lists the performance of the multilingual model averaged over all CV languages in various settings. Table 3 reports the same information for CV languages that are in VP. All results for CV are reported using greedy decoding in terms of character error rate (CER), as suggested in [12]. In addition to the base model (trained only on CV), we report performance when the VP audio with the final PLs are added back into the training set, either by fine-tuning the model already trained on CV ("+ all PLs (fine-tune)") or by training a model from scratch on CV+VP ("+ all PLs (from scratch)"). We only report results for the large model when training it from scratch on CV+VP, as the large model overfit to CV after a few epochs. Test CER is measured by selecting the checkpoint with the best average validation CER across all languages. While performance is degraded on average (Fig. 3), it is greatly improved for the VP languages (Fig. 4), with the best results achieved training a larger model from scratch.

Table 2. CER averaged over all CV languages.

Model	Valid CER	Test CER
Base model	26.8	28.8
+ all PLs (fine-tune)	27.6	29.7
+ all PLs (from scratch, base)	38.0	39.9
↪ fine-tune on CV only	26.6	28.2
+ all PLs (from scratch, large)	35.4	37.1
Monolingual baseline	33.8	35.5
Supervised fine-tuning	10.6	11.4

Table 3. CER averaged over languages in (CV languages ∩ VP languages).

Model	Valid CER	Test CER
Base model	24.4	24.8
+ all PLs (fine-tune)	17.5	17.9
+ all PLs (from scratch, base)	15.0	15.6
↪ fine-tune on CV only	13.8	14.0
+ all PLs (from scratch, large)	11.7	12.2
Monolingual baseline	25.1	26.8
Supervised fine-tuning	7.7	8.3
slimIPL fine-tuning	6.9	7.5

We also train a monolingual model for each language separately using the same hyperparameters as the multilingual model, and report the performance of those models along with the performance of the multilingual model when fine-tuned using only labeled data for that language ("supervised fine-tuning") or, when unlabeled data is available (Table 3), using both labeled and unlabeled data for that language ("slimIPL fine-tuning"). For monolingual models, or multilingual models with monolingual fine-tuning, the test CER is measured using the checkpoint with the best validation CER. There is still a large gap between the base model and fine-tuned models (see e.g. Greek in Table 1), but the gap is reduced for the VP languages when training on the pseudo-labeled data.

To see how well the multilingual models perform on out-of-domain audio, we evaluate them on LibriSpeech in Table 4.

[3]To our surprise, we found that using the *wrong language* (English) for unlabeled data could also improve test performance over the supervised monolingual baseline. We leave exploring this phenomenon to future work.

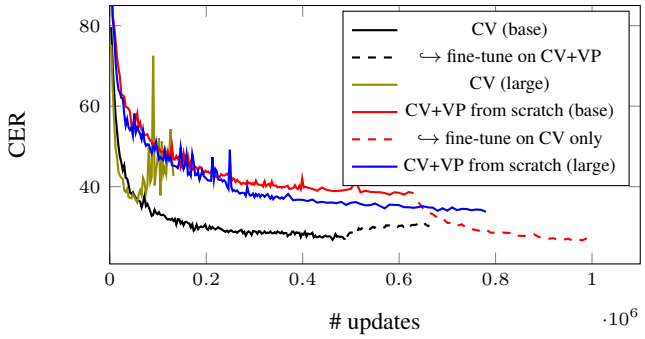

Fig. 3. Average dev CER for all languages when training from scratch or fine-tuning. (Fine-tuning begins at ∼500k updates.)

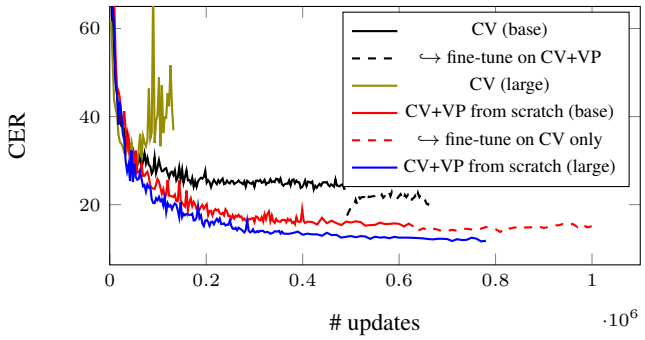

Fig. 4. Average dev CER for languages in CV ∩ VP when training from scratch or fine-tuning.

Word error rate (WER) is reported both using greedy decoding and using a beam search for

$$\underset{\mathbf{y}}{\operatorname{argmax}} \ \log p_\theta(\mathbf{y}|\mathbf{x}) + \alpha \log p^{\mathrm{LM}}(\mathbf{y}) + \beta|\mathbf{y}|, \qquad (1)$$

where p^{LM} is the probability according to an external 4-gram LM, and α, β are set using a small grid search on the dev sets. We find that the multilingual model fine-tuned with all VP PLs performs much better on LibriSpeech across all settings. It can be seen from Table 5, in which test-other is split by the duration of utterances, that the improvement is due mostly to the model's ability to process longer sequences acquired from training on the longer VP utterances (see Sec. 5.2).

We also demonstrate the base model's transfer capability by fine-tuning it either on the 100h or 960h subset of LibriSpeech (Table 4, "CV → LS-{100,960}"). With fine-tuning on LibriSpeech, performance is significantly improved for the 100h setup over the 100h-only training, while with 960h performance is similar or slightly worse. We have not yet made these comparisons for the CV+VP models, but our other results suggest that similar benefits may be observed.

[4]CV models output punctuation/capitalization (accounts for some errors).

Table 4. LibriSpeech WER for different training sets.[4]

Data	LM	Dev WER		Test WER	
		clean	other	clean	other
CV	-	59.7	60.1	62.0	62.8
	4-gram	33.7	34.3	37.6	37.7
CV → CV+VP	-	34.1	41.7	33.5	42.5
	4-gram	8.8	15.9	9.0	16.8
CV → LS-100	-	4.8	13.7	5.1	13.6
	4-gram	3.3	9.7	3.8	9.9
CV → LS-960	-	3.0	7.5	3.1	7.4
	4-gram	2.1	5.3	2.6	5.8
LS-100	-	6.2	16.8	6.2	16.8
	4-gram	4.1	12.4	4.5	12.7
LS-960	-	2.7	6.8	2.8	6.9
	4-gram	2.0	5.1	2.6	5.7

Table 5. WERs for test-other split over audio duration.

Data	LM	Duration			
		<10s	10-15s	15-20s	>20s
CV	-	46.6	83.6	99.1	99.9
	4-gram	15.6	54.7	93.3	98.7
CV → CV+VP	-	43.5	38.6	41.3	47.7
	4-gram	17.0	15.2	17.2	20.8

7. CONCLUSION

We have demonstrated the use of pseudo-labeling to improve an end-to-end joint model for massively multilingual ASR with Common Voice, by fine-tuning a multilingual model with semi-supervised learning on each language of VoxPopuli separately. Training on all VoxPopuli pseudo-labels combined i) significantly improves the performance of the multilingual model for those 19 languages, ii) helps the model generalize to a new domain (LibriSpeech), and iii) enables training a larger model than was possible with Common Voice alone without overfitting. Future work could look into reducing the gap between the performance of the multilingual model on its own and after fine-tuning on a particular language, improving performance for languages without unlabeled data, integrating language models into the PL generation process, and running iterative pseudo-labeling instead of a single round. Our method also requires knowledge of which language is spoken in the unlabeled audio: overcoming this requirement, so that even more data in the wild can be used, would also be worth exploring.

8. REFERENCES

[1] Tanja Schultz et al., "Multilingual and crosslingual speech recognition," in *DARPA BNTU Workshop*, 1998.

[2] Vineel Pratap et al., "Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters," *Interspeech*, 2020.

[3] Dong-Hyun Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshop on Rep. Learning*, 2013.

[4] Gabriel Synnaeve et al., "End-to-end ASR: from supervised to semi-supervised learning with modern architectures," *ICML SAS Workshop*, 2020.

[5] Yu Zhang et al., "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv*, 2020.

[6] Qiantong Xu et al., "Self-training and pre-training are complementary for speech recognition," in *ICASSP*, 2021.

[7] Qiantong Xu et al., "Iterative pseudo-labeling for speech recognition," *Interspeech*, 2020.

[8] Yosuke Higuchi et al., "Momentum pseudo-labeling for semi-supervised speech recognition," *Interspeech*, 2021.

[9] Tatiana Likhomanenko et al., "slimIPL: Language-Model-Free Iterative Pseudo-Labeling," *Interspeech*, 2021.

[10] Jacob Kahn et al., "Self-training for end-to-end speech recognition," in *ICASSP*, 2020.

[11] Daniel S Park et al., "Improved noisy student training for automatic speech recognition," *Interspeech*, 2020.

[12] Rosana Ardila et al., "Common Voice: A Massively-Multilingual Speech Corpus," *LREC*, 2020.

[13] Changhan Wang et al., "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *ACL*, 2021.

[14] V. Panayotov et al., "LibriSpeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[15] Alex Graves et al., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[16] Jacob Kahn et al., "Flashlight: Enabling innovation in tools for machine learning," *arXiv preprint arXiv:2201.12465*, 2022.

[17] Markus Müller et al., "Phonemic and graphemic multilingual CTC based speech recognition," *arXiv*, 2017.

[18] Sibo Tong et al., "Cross-lingual adaptation of a CTC-based multilingual acoustic model," *Speech Communication*, vol. 104, pp. 39–46, 2018.

[19] Alexis Conneau et al., "Unsupervised cross-lingual representation learning for speech recognition," *arXiv*, 2020.

[20] Oliver Adams et al., "Massively multilingual adversarial speech recognition," 2019.

[21] Anjuli Kannan et al., "Large-scale multilingual speech recognition with a streaming end-to-end model," *Interspeech*, 2019.

[22] Jaejin Cho et al., "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," 2018.

[23] Bo Li et al., "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," 2018.

[24] Wenxin Hou et al., "Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning," in *Interspeech*, 2020.

[25] Jan Chorowski et al., "Attention-based models for speech recognition," *NeurIPS*, 2015.

[26] William Chan et al., "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016.

[27] Alex Graves, "Sequence transduction with recurrent neural networks," *ICML Rep. Learn. Workshop*, 2012.

[28] Yanzhang He et al., "Streaming end-to-end speech recognition for mobile devices," in *ICASSP*, 2019.

[29] Alexei Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.

[30] Tatiana Likhomanenko et al., "Rethinking evaluation in ASR: Are our models robust enough?," *Interspeech*, 2021.

[31] Daniel S Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," *Interspeech*, 2019.

[32] Shubham Toshniwal et al., "Multilingual speech recognition with a single end-to-end model," in *ICASSP*, 2018.

[33] Tatiana Likhomanenko et al., "CAPE: Encoding relative positions with continuous augmented positional embeddings," *NeurIPS*, 2021.

[34] Chung-Cheng Chiu et al., "RNN-T models fail to generalize to out-of-domain audio: Causes and solutions," in *SLT Workshop*, 2021.