# Phonetic-based Ranking for Improved Pseudo-Labeling in Low-Resource ASR

Marco Matassoni, Roberto Gretter, Daniele Falavigna, Mohamed Nabih Nawar, Alessio Brutti,
Mauro Cettolo, Marco Gaido, Matteo Negri, Sara Papi, Luisa Bentivogli
Fondazione Bruno Kessler, Trento, Italy
{gretter,matasso}@fbk.eu

**Abstract**

The rise of large language models has boosted speech and language technologies; however, where transcripts of audio data are limited, the performance of current technology is not yet satisfactory. One common strategy to tackle data scarcity is leveraging pseudo-labels, for example automatically transcribing data with a pre-trained ASR. One critical issue of this approach is assessing the quality of the automatic transcriptions, that may be rather bad for low-resourced languages. While several filtering approaches exist in literature, they typically work with decent pre-trained ASR models but may fail otherwise. In this work we propose a phonetic-based ranking, enabling an effective selection with controllable computational resources; the resulting subset of pseudo-labels serves as additional material for fine-tuning the source ASR models. Experiments on common benchmarks in three low-resource languages demonstrate the effectiveness of the proposed approach, yielding up to a 3-point reduction in WER.

**Keywords:** low-resource language, speech recognition, weak supervision, phonetic transcription, pseudo-labeling, data filtering

## 1. Introduction

Recent years have seen remarkable progress in language and speech technologies, fueled by breakthroughs in deep learning and the growing availability of large-scale datasets (Prabhavalkar et al., 2023). In particular, the emergence of self-supervised representation learning methods has substantially reduced the dependence on annotated data (Ericsson et al., 2022). Despite these advances, low-resource scenarios (i.e. data availability and manual annotations are limited) continue to pose significant challenges for systems deployed in production environments (Lam-Yee-Mui et al., 2023), which traditionally rely on large-scale speech corpora for effective training. Additionally, the process of collecting and manually transcribing spoken language data remains both resource-intensive and costly (Yu et al., 2024; Gaido et al., 2024). Given the vast linguistic diversity across the globe, the development of speech recognition technologies tailored for underrepresented languages and specialized domains has become increasingly crucial in broadening access to these modern computational tools for a wider range of users.

A widely adopted strategy to mitigate data scarcity in ASR is transfer learning (Lam-Yee-Mui et al., 2023; Zhou et al., 2024), which harnesses the knowledge of large-scale corpora to improve model performance in low-resource tasks and languages by exploiting multilingual and cross-lingual similarities (Farooq et al., 2023; Yu et al., 2023; Piñeiro-Martín et al., 2024). Data augmentation (Bartelds et al., 2023; Lu and Li, 2024) offers a compelling strategy to enhance model generalization to unseen data by artificially expanding the training data set through various techniques. Generally, unlabeled speech data are available in relatively large quantities, even for underrepresented languages, making its use increasingly viable for augmenting speech datasets with pseudo-labels automatically generated using a pre-trained ASR. Typically popular models exhibit transcription performance below 20% WERs and this can guaranty on average reliable quality for weakly supervised training; however, for low resource languages, the resulting WERs may be higher than 50% - e.g. (Radford et al., 2023) presents tables highlighting specific datasets or languages that exhibit particularly WERs - and in this case the strategy may fail.

In this study, we propose an approach for assessing the quality of pseudo-labels by leveraging multilingual phonetic transcribers and analyzing phonetic units rather than graphemes; we use the terms *pseudo-labeling* and *weak supervision* interchangeably. Previous research has demonstrated the effectiveness of multilingual alphabets in the development of multilingual ASR systems for low-resource languages using models trained from scratch (Liu et al., 2020; Schultz and Schlippe, 2014; Yusuyin et al., 2024; Lee et al., 2025). Additionally, selecting the most informative subsets of data for fine-tuning foundation models based on error metrics has been shown to be effective in active learning-based ASR approaches (Riccardi and Hakkani-Tur, 2005; Long et al., 2013; Fu and Ru, 2019) as well as in vision applications (Mariya et al., 2019; Paul et al., 2021; Sorscher et al., 2022). Therefore, we investigate the application of this approach to select subsets of utterances from an un-



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10130–10139
11–16 May 2026. 

labeled dataset that meet the quality requirements for effective model training.

Additionally, we assess the usefulness of these weakly-labeled, filtered speech datasets in the context of European languages (Gaido et al., 2024; Rao Koluguri et al., 2025), where notable disparities in recognition performance have been observed across official languages. We pursue this direction by moving into the phonetic domain to better explore cross-lingual similarities. Specifically, we rank automatic transcripts based on estimated accuracy using representations derived from the International Phonetic Alphabet (IPA). These rankings are then used to select the most promising utterances, aiming to balance the trade-off between the quantity of training data and the quality of the labels.

In summary, the contributions are as follows. *a)* Starting from automatic transcripts obtained with a state-of-the-art ASR, we present a filtering scheme that **ranks the automatic transcripts** (i.e. pseudo labels) based on an estimate of their accuracies using representations based on IPA. These rankings are then used to select the most promising utterances, balancing the trade-off between the amount of training data and the quality of the labels. The method may be applied to **any language, particularly in case of very low ASR performance**, and it does not require high computational resources. *b)* on three standard ASR benchmarkswe show that our selection methods allow to **increase the ASR performance** in three low-resource languages: Lithuanian (lt), Maltese (mt) and Slovenian (sl) for which ASR performance is worse than other EU languages, or the labeled audio resources are scarce, gaining approximately 2 to 3 WER points.

### 1.1. Related works

The use of weak supervision, or pseudo-labels, obtained by automatically transcribing unlabeled audio data with a pre-trained ASR has been explored in several works to target data scarcity (Attia et al., 2025; Shao et al., 2025; Damianos et al., 2025; Rangappa et al., 2025a; Cheng et al., 2024). **Iterative pseudo-labeling** (IPL) (Xu et al., 2020; Likhomanenko et al., 2023) consists of iteratively refining the ASR model and, therefore, the quality of the transcript. Another line of work investigates **training strategies** specifically designed for pseudo-labels as in (Higuchi et al., 2022), which also operate in an iterative fashion, and (Ling et al., 2022; Gao et al., 2023) or transfer learning from similar high-resource languages (San et al., 2024).

Another popular direction involves **filtering the weak-labels** in order to retain only the most accurate ones. This can be achieved by predicting errors in the transcriptions by comparing the output of different systems (Ali and Renals, 2018, 2020; Chowdhury and Ali, 2023; Bhogale et al., 2024), implementing some heuristics and using confidence measures as in (Tang et al., 2022; Bhogale et al., 2024) or relying on word-ratio and perplexity distribution metrics (Rangappa et al., 2025b). These approaches typically rely on decent accuracy of the pseudo-labels whose WER may be around 20-25%.

The method proposed in this paper falls into the last category, namely, it is a filtering approach that can be applied in combination with iterative training or other methods specifically designed for pseudo-labels. Unlike the approaches mentioned above, our method has several advantages. First, does not need preliminary **training**, meaning that no additional training data is required to rank the transcripts. Second, it **does not rely on information from the ASR system** (e.g., confidence scores), allowing the ASR to be treated as a black-box off-the-shelf component. Third, the method is **language-independent**. Fourth, it uses **computationally lightweight IPA recognizers**, which are more efficient than approaches that require two or more ASR engines. Finally, our filtering strategy is **particularly effective for low-resource languages**, where initial models often perform poorly because the IPA model can leverage cross-lingual information.

## 2. Methodology

Our proposed approach is depicted in Figure 1. We use fine-tuned versions of `whisper-large-v3` to automatically transcribe the unlabeled data, generating the pseudo labels. Given the input audio sequence $s$ and the Whisper (word-based) ASR engine $W$ we obtain the automatic transcripts $l_W$:

$$l_W = W(s). \tag{1}$$

In order to estimate the quality of the transcripts, we leverage the availability of (multi-lingual) pre-trained models that exploit cross-lingual representations and are capable of generating phonetic-based hypotheses (i.e. sequences of IPA symbols) (Xu et al., 2022; Taguchi et al., 2023). In this work, we use the model `wav2vec-lv-60-espeak-cv-ft`[1] described in (Xu et al., 2022). The model is based on `Wav2Vec2`, which is fine-tuned on multilingual data to produce IPA symbols as output (Taguchi et al., 2023). From $s$ we can therefore obtain the sequence of IPA symbols $p_E$:

$$p_E = E(s), \tag{2}$$

where $E$ is a phonetic transcriber (Bernard and Titeux, 2021). Since this model learns phonetic

[1] https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft

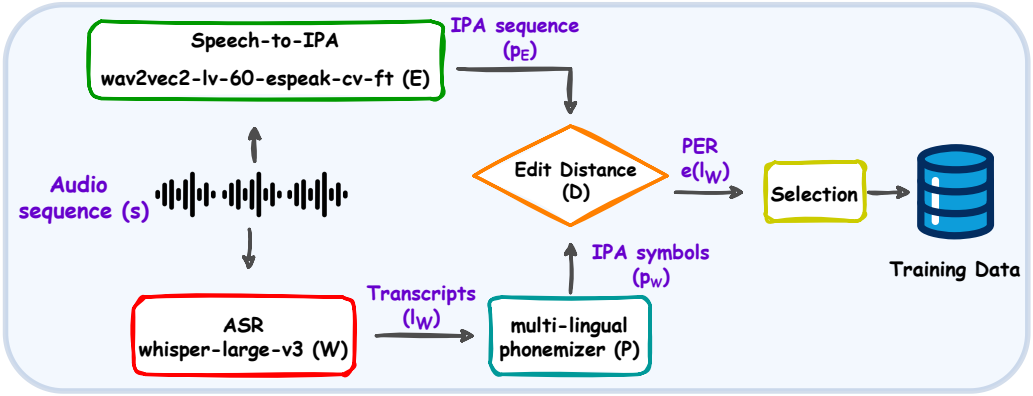


Figure 1: Pipeline for training data selection. Each audio sample is transcribed with two pipelines; the edit distance between the two resulting phonetic sequences (PER) is used to rank the data.

representations, it operates more at the acoustic level. Therefore, we can assume that its output is less dependent on both the specific language and the specific domain. Note that, in principle, more than one ASR system could be used to produce $l_W$ or $p_E$, thus generating several alternate hypotheses to combine and, consequently, improving the automatic selection process and the final performance as in (Ali and Renals, 2018, 2020; Chowdhury and Ali, 2023; Bhogale et al., 2024). However, decoding large amounts of unlabeled data with current ASR systems is extremely costly and time consuming. Moreover, if their quality is very low, as in low-resourced settings, their comparison is often not useful.

Using an open-source *phonemizer*[2] $P$, we convert the transcript into IPA symbols:

$$p_W = P(l_W) \tag{3}$$

and evaluate the similarity between the two phonetic sequences through the edit distance. This results in a Phone Error Rate (PER) associated with each weak transcript:

$$e(l_W) = \text{PER}(D(p_W, p_E)) \tag{4}$$

where $D$ is the edit distance. We speculate that the PER correlates with the quality of the automatic transcriptions and can be used to rank them.

As a preliminary analysis of the efficacy of our proposed method, Table 1 reports the Pearson Correlation Coefficients (PCC) that measures the correlation between the PERs introduced above (i.e. measuring the edit distance between the hypotheses produced by the two models) and the actual CERs measured on labeled datasets (the transcribed Vox Populi test sets for Lithuanian and Slovenian and the MASRI test set for Maltese, see Section 3.1 for details on these corpora). Note the rather high correlation between our estimator and the actual CER of the transcript, which confirms that the proposed phonetic ranking is effective. As a comparison, we also report the PCCs with an estimator based on two diverse ASR models (i.e. `whisper-large-v3` and the popular Meta's MMS, `mms-1b-fl102`, trained in 102 languages [3]), which is a reasonable alternative. On average, the correlations obtained with the phonetic model are slightly better than those obtained by using two ASR engines. However, even if correlations are similar, using an auxiliary ASR model is extremely time-consuming, whereas the phonetic model is very fast and computationally efficient. Therefore, in this work, the filtering mechanism is based on this phonetically-based comparison.

Table 1: Agreement, measured as Pearson Correlation coefficients (PCC), between real CERs obtained with whisper-large-v3 and our PERs (the proposed quality estimator), obtained on the labelled test set of Vox Populi for Lithuanian and Slovenian, and on the MASRI test set for Maltese (first row). The second row reports the PCCs beetween the actual CERs and the CERs estimated using two ASR transcripts (i.e. with two different ASR models: Whisper and MMS).

| | lt | mt | sl |
|---|---|---|---|
| whisper-large-v3 vs. wav2vec2-lv-60-espeak-cv-ft | 0.97 | 0.90 | 0.86 |
| whisper-large-v3 vs. mms-1b-fl102 | 0.98 | 0.86 | 0.89 |

Finally, PERs associated with the different sentences can be used to select the weak labels following two approaches: *a)* sentences are selected according to a given threshold, leveraging a trade-

[2] https://github.com/bootphon/phonemizer

[3] https://huggingface.co/facebook/mms-1b-fl102

off between quality and quantity, or *b)* ranking the transcribed utterances to select the top-K, in case a given amount of training data is necessary. In the next section, we experimentally demonstrate the effectiveness of this ranking method.

## 3. Experimental Setup

### 3.1. Datasets

We evaluate the proposed approach considering three public datasets that cover many European languages: Common Voice 19 (CV) (Ardila et al., 2020), FLEURS (FL) (Conneau et al., 2023), and VoxPopuli (VP) (Wang et al., 2021). Specifically, three European languages are selected for which limited amounts of public labeled training data are available and for which the performance of current publicly available multilingual ASR systems is rather low (if compared against other more represented languages): Lithuanian (lt), Maltese (mt) and Slovenian (sl). Statistics related to the three datasets for the languages under analysis are reported in Table 2.

The reported volumes confirm that unlabeled material is often available in quantity. Even if VP does not include a transcribed section for Maltese, we are particularly interested in the language due to the very low performance of current multilingual models. Therefore, we also consider the MASRI (test) corpus (Hernandez Mena et al., 2020). It consists of YouTube videos belonging to the University of Malta channel (1 hour) and is gender balanced; the audio transcription was performed by the MASRI Team at the University of Malta.

### 3.2. Implementation

As ASR baseline, we consider the `whisper-large-v3` model (Radford et al., 2023): a popular state-of-the-art multilingual speech recognizer trained on 99 languages. Greedy search is used as the decoding strategy in the experiments.

Figure 2 shows our approach: Whisper is first fine-tuned on CV and FL training sets, for each language. Then, the resulting fine-tuned model is used to transcribe the unlabeled VP data, which also undergo phoneme recognition. The most promising unlabeled VP data, using the PER selection procedure, are then used to augment the CV and FL training sets and fine-tune again the original Whisper model.

As a preliminary analysis, Table 3 reports the performance (as Word Error Rate, WER%) of the baseline model (first row) on both CV and FL evaluation sets. It is evident that the WERs for the three languages we are focusing on are not satisfactory, particularly compared to the performance reported in high-resource languages, where WERs are typically much lower than 10% % for these datasets.

Table 2: Statistics for Common Voice, FLEURS and VoxPopuli in hours and minutes (hh:mm). Compared to high-resource languages, the labeled audio material is very limited while the availability of unlabeled data is vastly greater (Wang et al., 2021). For Common Voice we consider the official partitions.

| | **CommonVoice (CV)** | | |
|---|---|---|---|
| | Train | Dev | Test |
| lt | 10:11 | 06:33 | 07:08 |
| mt | 02:23 | 02:05 | 02:19 |
| sl | 01:20 | 01:22 | 01:25 |
| | **FLEURS (FL)** | | |
| | Train | Dev | Test |
| lt | 09:45 | 01:10 | 02:58 |
| mt | 09:54 | 01:29 | 03:32 |
| sl | 07:45 | 00:53 | 02:16 |
| | **VoxPopuli (VP)** | | |
| | Transcribed | Unlabelled | |
| lt | 01:28 | 14404:00 | |
| mt | - | 9083:00 | |
| sl | 09:39 | 11331:00 | |

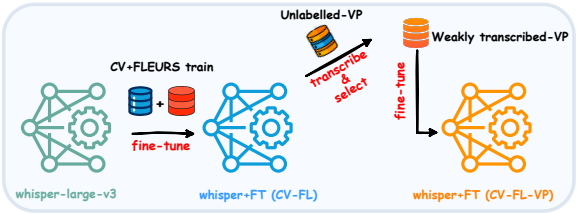


Figure 2: Whisper is fine-tuned on CV and FL training sets. The new model is used to transcribe the unlabelled VP data. The selected transcripts augment CV and FL training sets to fine-tune again the Whisper model.

As mentioned in Sec. 1, one common approach is to specialize the multi-lingual model to a given language with labeled data. This is in fact beneficial, as shown in Table 3. Starting from the original Whisper model, full fine-tuning (FT) or adapting (using default LoRA transformations on the query and value projection layers of the self-attention mechanism) on the CV training set improves performance even using a very small amount of training data (less than 1.5 hours for Slovenian, for example).

However, the procedure shows model brittleness, since improvements on FL are negligible or even negative, and LoRA is not always effective. In general, fine tuning the model may improve the performance, but WERs are largerly far from that of high-resourced languages and there is room for

Table 3: WER (%) on CV and FL with Whisper models: whisper-large-v3, whisper-large-v3 fine-tuned on CV, whisper-large-v3 adapted with LoRA on CV and whisper-large-v3 fine-tuned on CV and FT.

| | | CV | FL |
|---|---|---|---|
| **lt** | whisper-large-v3 | 31.1 | 25.3 |
| | + FT on CV | 17.2 | 24.0 |
| | + LoRA on CV | 19.8 | 28.1 |
| | **+ FT on CV-FL** | 15.5 | 22.6 |
| **mt** | whisper-large-v3 | 82.9 | 74.3 |
| | + FT on CV | 28.3 | 32.6 |
| | + LoRA on CV | 76.6 | 68.9 |
| | **+ FT on CV-FL** | 14.7 | 16.6 |
| **sl** | whisper-large-v3 | 18.8 | 20.6 |
| | + FT on CV | 16.1 | 26.2 |
| | + LoRA on CV | 11.8 | 22.2 |
| | **+ FT on CV-FL** | 15.9 | 23.8 |

further improvement

These experimental results highlight the necessity of exploring strategies to enhance the effectiveness of multilingual models in low-resource language settings, particularly by leveraging the availability of unlabeled audio data. Henceforth, we define the model fine-tuned on both CV and FL as the baseline. Given the contrastive findings observed with LoRA in the preceding experiments, the remainder of this paper focuses exclusively on full fine-tuning.

### 3.3. Selection of weakly supervised data

Using the quality predictor described in Section 2, we select different volumes of VP data to be used to supplement the initial training set. Table 4 illustrates the data volumes selected according to different PER thresholds: with higher values, the amount of weakly supervised material increases at the expense of the expected quality of the associated transcripts.

Table 4: Data selection on the VP unlabeled dataset, based on automatic PER-based ranking: data volumes selected according to different PER thresholds for the three languages.

| PER % | lt | mt | sl |
|---|---|---|---|
| 25 | 43h | 7m | 87h |
| 30 | 640h | 200m | 1191h |
| 35 | 3168h | 64h | 3906h |
| 40 | 7059h | 576h | 6400h |
| 45 | 9932h | 2153h | 7845h |
| 50 | 11243h | 4006h | 8531h |

As expected, very different subsets across languages are created for a given threshold. The very different quality of the transcriptions combined with the capabilities of both the phonetic model and the phonemizer induce very specific phonetic distances on the three languages. Note that the Maltese subsets are considerably smaller compared to the other two. While after fine tuning on CV and FL Whisper largely improves for Maltese on the related test sets (14.7% and 16.6% WER as reported in Table 3), performance is probably still low for VP which leads to selecting smaller amount of data for a given PER in comparison with the other languages. Note that the proposed quality estimation method correlates well with the actual WER on the transcribed part of VP as reported in Table 1, so we can reasonably assume that the fine-tuned model does not work well on VP. Given these augmented training sets, the Whisper model is fine-tuned again using CV, FL and the weak labels from VP.

### 3.4. Fine-tuning setup

We experiment with different values of the PER threshold to estimate the best trade-off between quantity and quality. The popular *transformers* library (Wolf et al., 2020) is used for the experiments, which are carried out on a NVIDIA H100 GPU. The learning rate and batch size are set to $0.00005$ and $32$, respectively; Adam with weight decay is used as optimizer; the best checkpoint is selected on the FL development set, using the Character Error Rate (CER) as metric.

## 4. Results

We initially focus on Maltese, as it is the language with less data and poorer baseline performance. First, we further experimentally validate our preliminary observations in Table 1 on the correlation between the PER and the transcription usefulness completing the entire pipeline. We select 100 hours of automatic VP transcripts for fine-tuning, either randomly or using our PER-based quality estimator.

The results reported in Table 5 confirm that selecting training samples with the lowest PER for fine-tuning is appropriate, as it consistently outperforms random selection. This is particularly evident on the MASRI test set for which no training data is used at all. On both CV and FL test sets, for which labeled training data are available and have been used to fine-tune the model, the improvement is smaller, but consistent.

Having assessed the usefulness and potential of the proposed approach, we investigate the trade-off between quantity and quality of the weak labels. Table 6 reports the results obtained by selecting different amounts of training data on Maltese. The results confirm the validity of our approach on all

Table 5: WER (%) achieved on the Maltese evaluation sets selecting 100h of VP automatic transcriptions randomly or using our proposed approach.

| selection method | CV test | FL test | MASRI test |
|---|---|---|---|
| random | 15.5 | 15.6 | 51.6 |
| PER-based (proposed) | **13.3** | **14.5** | **33.5** |

test sets. The selection procedure adds training material that consistently improves the resulting fine-tuned model. Although a certain degree of adaptation to the dataset is observable, WERs, in particular on MASRI, confirm the positive impact of the augmentation strategy, which improves the WER from 45.2% to 33.0% using a value of PER threshold equal to 35%, which corresponds to further 64 hours of training data (see Table 4).

As mentioned above, the improvement is more evident in MASRI because both FL and CV data have already been utilized in the first fine-tuning step. Nevertheless, significant WER improvements are observed also on both CV and FL tests sets: from 14.7% to 13.6% and from 16.6% to 15.1%, respectively with PER threshold equal to $35\%$. Note also that there seems to exist an optimal quality-quantity trade-off: selecting data of lower quality (i.e. PER higher than 40%) deteriorates the performance.

Table 6: WER (%) results on Maltese: the amount of weakly supervised VP portion is selected based on PER. Whisper+FT-CV-FL represents the model fine tuned on CV and FL (i.e., no weak label augmentation).

| PER (%) th | CV test | FL test | MASRI test |
|---|---|---|---|
| **Maltese** | | | |
| whisper-large-v3 + FT on CV-FL | 14.7 | 16.6 | 45.2 |
| 30 | 14.5 | **14.3** | 34.0 |
| 35 | **13.6** | 15.1 | **33.0** |
| 40 | 14.2 | 16.3 | 38.4 |
| 45 | 14.3 | 16.6 | 39.9 |
| 50 | 14.8 | 16.6 | 43.5 |

To further investigate the impact of transcription quality, we generated new transcripts using a better model fine-tuned on proprietary data by the MASRI Project (Hernandez Mena, 2023) .

Given the quantity and type of data in the MASRI dataset, this model is expected to generate better weak labels than our model (only) fine-tuned on CV and FL. Figure 3 illustrates that augmenting the training material with this transcript results in improved models (the yellow curve is consistently below the red curve). Since the transcripts are of higher quality, the same PER threshold selects more training material (for instance, PER 35% corresponds to 295h instead of 64h), leading to better WERs.

However, also this curve exhibits a local minimum, indicating that using all transcripts is not optimal for fine-tuning. This suggests a trade-off between audio quantity and transcript quality.

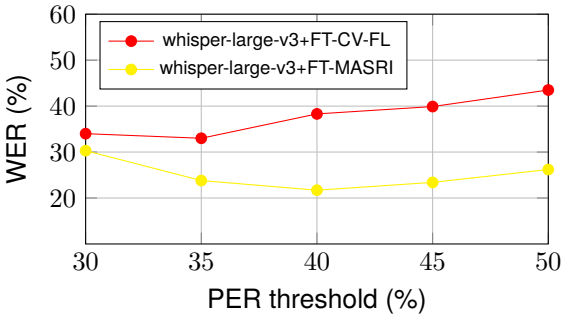


Figure 3: WER results on MASRI test set with `whisper-large-v3` fine-tuned on a training set which includes CV, FL, and different portion of VP.

To complete the experimental analysis, we examine the remaining two languages under investigation. The results, presented in Table 7, demonstrate a consistent pattern, indicating that an optimal data quantity — corresponding to a 25% PER — yields superior performance across multiple datasets and languages. In the case of Lithuanian, the initial WERs are recorded at 15.3%, 22.6%, and 35.2%. Incorporating selected VP data into the training material improves performance in all three test sets, resulting in WERs of 15.5%, 19.5%, and 33.1%.

Similarly, for Slovenian initial WERs are 15.9%, 23.8%, and 34.6% whilst after fine-tuning with additional 87 hours (i.e., threshold 25) we obtain 13.4%, 20.4%, and 27.7%, respectively. Note that this trade-off corresponds to 43 and 87 hours for Lithuanian and Slovenian, respectively, which aligns with the 64 hours trade-off observed for Maltese.

Finally, Figure 4 provides a comprehensive summary of the key findings presented so far, illustrating the progressive improvements in WER. The analysis considers the Whisper baseline, the fine-tuned models trained on the initial small dataset, and the additional enhancements introduced by the proposed data selection method for weakly supervised training. For this evaluation, we adopt the optimal PER threshold, determined from the MASRI and VP transcribed out-of-domain test datasets.

The results indicate that only a relatively small amount of additional training data is required to achieve significant model improvements: in our setup, the largest performance gains are observed with Slovenian (87 hours), Lithuanian (43 hours),

Table 7: WER results on Lithuanian and Slovenian test sets with fine-tuned whisper models, where the amount of weakly supervised VP portion is selected based on PER.

| PER % th | CV test | FL test | VP transcribed |
|---|---|---|---|
| **Lithuanian** | | | |
| whisper-large-v3 + FT on CV-FL | **15.3** | 22.6 | 35.2 |
| 25 | 15.5 | **19.5** | **33.1** |
| 30 | 16.1 | 20.8 | 33.4 |
| 35 | 17.4 | 22.8 | 35.0 |
| **Slovenian** | | | |
| whisper-large-v3 + FT on CV-FL | 15.9 | 23.8 | 34.6 |
| 25 | **13.4** | **20.4** | **27.7** |
| 30 | 14.9 | 23.4 | 30.3 |
| 35 | 17.9 | 24.4 | 30.8 |

and Maltese (64 hours). A more granular investigation is necessary to further evaluate the precise trade-off between data quantity and transcript quality.

## 5. Conclusion

In this work, we have investigated a data selection mechanism based on a ranking strategy based on phonetic comparison between a multilingual phone model and a Whisper model. The quality of the transcript is then estimated and in turn used to select a suitable subset as additional training material.

In an ASR scenario for low-resource European languages, we have shown the effectiveness of weakly supervised training on unlabeled data in three different languages (Maltese, Lithuanian, and Slovenian). We proved that this selection improves the resulting fine-tuned model, also in comparison with a model fine-tuned on a small labeled dataset; it is particularly useful when the quality of pseudo-labels is expected to be low, making it especially suitable for low-resource languages that typically exhibit suboptimal ASR performance.

As expected, the optimal data volume varies according to the target language and the source model used to generate the automatic transcriptions. More interestingly, our study shows that a small percentage of the initial unlabeled dataset allows us to obtain consistent improvements across datasets.

Promising future directions are the study of other phonetically driven techniques for data selection and the use of other cross-lingual phone-base models tailored for the languages under investigation (Kim et al., 2025), as well as a more sophisticated distance or clustering strategy to better accommodate phonetic units.

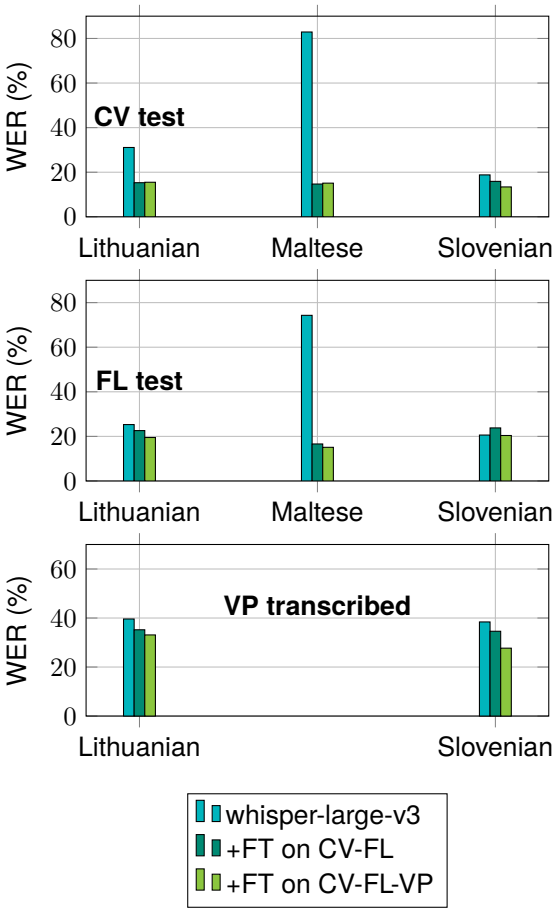


Figure 4: WERs on Common Voice test and FLEURS test using `whisper-large-v3` and the corresponding fine-tuned models using CV, FL and VP portions automatically transcribed and selected via the proposed method.

## 6. Ethics Statement and Limitations

All experiments in this work were conducted using publicly available speech corpora that are distributed for research purposes. No personally identifiable or private data was collected or used. We carefully reviewed the terms of use and data documentation for each dataset to ensure compliance with their intended scope and ethical guidelines.

Our study focuses on audio data selection and pseudo-labeling methods for ASR; transcripts were generated automatically from publicly released models and were used solely for research on weakly supervised learning techniques.

Specifically, the VoxPopuli corpus was employed as the source for pseudo labels; it consists of recordings of European Parliament proceedings, featuring speakers who are predominantly fluent

and articulate, and who operate within a formal and controlled acoustic environment.

We acknowledge that this limited demographic and stylistic diversity introduces potential biases in the resulting models. The relatively high quality of speech, professional context, and restricted linguistic and cultural variability may not reflect real-world speech conditions such as spontaneous conversation, diverse accents, or variable recording quality. Consequently, the findings derived from this dataset may not be generalized to broader populations or less formal speech domains.

Future work should incorporate more heterogeneous and representative datasets to mitigate these limitations and better evaluate the robustness of weakly supervised ASR approaches.

This research does not involve new data collection or interventions in the datasets. All results are shared according to the principles of responsible AI research, emphasizing reproducibility, transparency, and respect for the rights to data usage.

## 7. Acknowledgements

The authors acknowledge CINECA for the availability of high-performance computing resources and support.

Mohamed Nabih Nawar and Marco Gaido received funding from the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

Luisa Bentivogli, Mauro Cettolo, Matteo Negri and Sara Papi received funding from the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETings BetWEEN People).

Marco Matassoni, Roberto Gretter and Alessio Brutti received funding from the European Union's Horizon 2020 project ELOQUENCE (grant No 101070558).

## 8. Bibliographical References

Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24.

Ahmed Ali and Steve Renals. 2020. Word error rate estimation without ASR output: e-WER2. In *Interspeech 2020*, pages 616–620.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Ahmed Attia, Dorottya Demszky, Jing Liu, and Carol Espy-Wilson. 2025. From Weak Labels to Strong Results: Utilizing 5,000 Hours of Noisy Classroom Transcripts with Minimal Accurate Data. In *Interspeech 2025*, pages 3678–3682.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn B. Wieling. 2023. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. In *Annual Meeting of the Association for Computational Linguistics*.

Mathieu Bernard and Hadrien Titeux. 2021. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958.

Kaushal Santosh Bhogale, Deovrat Mehendale, Niharika Parasa, Sathish Kumar Reddy G, Tahir Javed, Pratyush Kumar, and Mitesh M. Khapra. 2024. Empowering Low-Resource Language ASR via Large-Scale Pseudo Labeling. In *Interspeech*, pages 2519–2523.

Yao-Fei Cheng, Li-Wei Chen, Hung-Shin Lee, and Hsin-min Wang. 2024. Exploring the impact of data quantity on asr in extremely low-resource languages.

Shammur Absar Chowdhury and Ahmed Ali. 2023. Multilingual word error rate estimation: e-WER3. In *ICASSP*, pages 1–5. IEEE.

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.

Dimitrios Damianos, Georgios Paraskevopoulos, and Alexandros Potamianos. 2025. MSDA: Combining Pseudo-labeling and Self-Supervision for Unsupervised Domain Adaptation in ASR. In *Interspeech 2025*, pages 3863–3867.

Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. 2022. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62.

Muhammad Umar Farooq, Rehan Ahmad, and Thomas Hain. 2023. MUST: A Multilingual Student-Teacher Learning Approach for Low-Resource Speech Recognition. *2023 IEEE Automatic Speech Recognition and Understanding Workshop*, pages 1–6.

Jiayi Fu and Kuang Ru. 2019. A dropout-based single model committee approach for active learning in asr. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 16–22.

Marco Gaido, Sara Papi, Luisa Bentivogli, Alessio Brutti, Mauro Cettolo, Roberto Gretter, Marco Matassoni, Mohamed Nabih, and Matteo Negri. 2024. MOSEL: 950,000 Hours of Speech Data for Open-Source Speech Foundation Model Training on EU Languages. In *2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Dongji Gao, Hainan Xu, Desh Raj, Leibny Paola García-Perera, Daniel Povey, and Sanjeev Khudanpur. 2023. Learning From Flawed Data: Weakly Supervised Automatic Speech Recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 1–8.

Carlos Daniel Hernandez Mena. 2023. Acoustic Model in Maltese: whisper-large-maltese-8k-steps-64h.

Carlos Daniel Hernandez Mena, Ayrton-Didier Brincat, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, and Iván Vladimir Meza Ruiz. 2020. MASRI-TEST CORPUS: Audio and Transcriptions in Maltese extracted from the YouTube channel of the University of Malta.

Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori. 2022. Momentum pseudo-labeling: Semi-supervised asr with continuously improving pseudo-labels. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1424–1438.

Minu Kim, Kangwook Jang, and Hoirin Kim. 2025. Improving cross-lingual phonetic representation of low-resource languages through language similarity analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Léa-Marie Lam-Yee-Mui, Waad Ben Kheder, Viet Bac Le, Claude Barras, and Jean-Luc Gauvain. 2023. Multilingual Models with Language Embeddings for Low-resource Speech Recognition. *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL)*.

Jaeyoung Lee, Masato Mimura, and Tatsuya Kawahara. 2025. Leveraging ipa and articulatory features as effective inductive biases for multilingual asr training. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Tatiana Likhomanenko, Ronan Collobert, Navdeep Jaitly, and Samy Bengio. 2023. Continuous soft pseudo-labeling in asr. volume 187 of *Proceedings of Machine Learning Research*, pages 66–84. PMLR.

Shaoshi Ling, Chen Shen, Meng Cai, and Zejun Ma. 2022. Improving pseudo-label training for end-to-end speech recognition using gradient mask. In *ICASSP*, pages 8397–8401. IEEE.

Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf, and Geoffrey Zweig. 2020. Multilingual graphemic hybrid ASR with massive data augmentation. In *Proceedings of SLTU and CCURL*, pages 46–52, Marseille, France.

Y. Long, M. J. F. Gales, P. Lanchantin, X. Liu, M. S. Seigel, and P. C. Woodland. 2013. Improving lightly supervised training for broadcast transcription. In *Proc. of Interspeech*, pages 2187–2191.

Hongxuan Lu and Biao Li. 2024. Sample adaptive data augmentation with progressive scheduling.

Toneva Mariya, Sordoni Alessandro, Tachet des Combes Remi, Trischler Adam, Bengio Yoshua, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *Proc. of ICLR*.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. In *Proc. of Neurips*, volume 34.

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, Maria del Carmen Lopez-Perez, and Georg Rehm. 2024. Weighted Cross-entropy for Low-Resource Languages in Multilingual Speech Recognition. *Interspeech*.

Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schluter, and Shinji Watanabe. 2023. End-to-End Speech Recognition: A Survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *40th International Conference on Machine Learning*.

Pradeep Rangappa, Andrés Carofilis, Jeena Prakash, Shashi Kumar, Sergio Burdisso, Srikanth Madikeri, Esaú Villatoro-Tello, Bidisha Sharma, Petr Motlicek, Kadri Hacioglu, Shankar Venkatesan, Saurabh Vyas, and Andreas Stolcke. 2025a. Efficient Data Selection for Domain Adaptation of ASR Using Pseudo-Labels and

Multi-Stage Filtering. In *Interspeech 2025*, pages 4928–4932.

Pradeep Rangappa, Juan Zuluaga-Gomez, Srikanth Madikeri, Andres Carofilis, Jeena Prakash, Sergio Burdisso, Shashi Kumar, Esaú Villatoro-Tello, Iuliia Nigmatulina, Petr Motlicek, Karthik Pandia, and Aravind Ganapathiraju. 2025b. Speech data selection for efficient asr fine-tuning using domain classifier and pseudo-label filtering. In *Proc. of ICASSP 2025*.

Nithin Rao Koluguri, Monica Sekoyan, George Zelenfroynd, Sasha Meister, Shuoyang Ding, Sofia Kostandian, He Huang, Nikolay Karpov, Jagadeesh Balam, Vitaly Lavrukhin, Yifan Peng, Sara Papi, Marco Gaido, Alessio Brutti, and Boris Ginsburg. 2025. Granary: Speech Recognition and Translation Dataset in 25 European Languages. In *Interspeech 2025*, pages 3923–3927.

Giuseppe Riccardi and Dilek Hakkani-Tur. 2005. Active learning: Theory and applications to automatic speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 13:504 – 511.

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. 2024. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.

Tanja Schultz and Tim Schlippe. 2014. GlobalPhone: Pronunciation dictionaries in 20 languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 337–341.

Mingchen Shao, Xinfa Zhu, Chengyou Wang, Bingshen Mu, Hai Li, Ying Yan, Junhui Liu, Danming Xie, and Lei Xie. 2025. Weakly Supervised Data Refinement and Flexible Sequence Compression for Efficient Thai LLM-based ASR. In *Interspeech 2025*, pages 748–752.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. In *Proc. of Neurips*, volume 35, pages 19523–19536.

Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. 2023. Universal Automatic Phonetic Transcription into the International Phonetic Alphabet. In *Interspeech 2023*, pages 2548–2552.

Raphael Tang, Karun Kumar, Gefei Yang, Akshat Pandey, Yajie Mao, Vladislav Belyaev, Madhuri Emmadi, Craig Murray, Ferhan Ture, and Jimmy Lin. 2022. SpeechNet: Weakly supervised, end-to-end speech recognition at industrial scale. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 285–293.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. pages 993–1003. Association for Computational Linguistics.

Thomas Wolf et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In *2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. In *Interspeech 2022*, pages 2113–2117.

Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert. 2020. Iterative pseudo-labeling for speech recognition. In *Interspeech 2020*, pages 1006–1010.

Jianwei Yu, Hangting Chen, Yanyao Bian, Xiang Li, Yimin Luo, Jinchuan Tian, Mengyang Liu, Jiayi Jiang, and Shuai Wang. 2024. AutoPrep: An Automatic Preprocessing Framework for In-The-Wild Speech Data. *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1136–1140.

Zhongzhi Yu, Yang Zhang, Kaizhi Qian, Cheng Wan, Yonggan Fu, Yongan Zhang, and Yingyan (Celine) Lin. 2023. Master-asr: achieving multilingual scalability and low-resource adaptation in asr with modular learning. In *Proceedings of the 40th International Conference on Machine Learning*.

Saierdaer Yusuyin, Te Ma, Hao Huang, Wenbo Zhao, and Zhijian Ou. 2024. Whistle: Data-efficient multilingual and crosslingual speech recognition via weakly phonetic supervision. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1440–1453.

Rui Zhou, Takaki Koshikawa, Akinori Ito, Takashi Nose, and Chia-Ping Chen. 2024. Multilingual Meta-Transfer Learning for Low-Resource Speech Recognition. *IEEE Access*, 12:158493–158504.