The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

Improving Low-Resource Speech Recognition through Multilingual Fine-Tuning with Language Identifiers and Self-Training

Karol Nowakowski
Tohoku University of
Community Service and Science
karol@koeki-u.ac.jp

Michal Ptaszynski
Kitami Institute of Technology
michal@mail.kitami-it.ac.jp

Abstract

Previous work has demonstrated that multilingual fine-tuning of a pretrained multilingual speech representation model can lead to improved speech recognition accuracy when there is extremely little target language data available. In this paper we show that fine-tuning on labeled speech data from multiple languages sharing common phonological traits, preprocessed by attaching a language identifier to each speech sample, yields competitive results compared to monolingual fine-tuning, even if a moderate amount of target language data is available. In order to further improve the performance of our system, we apply self-training using unlabeled speech data. Our results indicate that fine-tuning a speech recognition model jointly on a combination of multilingual data and pseudo-labeled data yields superior performance compared to using any of the two augmentation techniques individually. We also find that models fine-tuned on multilingual data with language identifiers produce better results even if explicit information about language identity is not provided at inference time.

Keywords: Speech recognition, Under-resourced language, Ainu, Multilingual learning, Transfer learning, Cross-lingual transfer, Language identifiers, Self-training

1 Introduction

It is believed that speech processing technologies can be leveraged in language documentation projects to speed up labor-intensive tasks such as speech transcription. However, for many languages it is difficult to develop a speech recognition system useful in real-world applications, as the accuracy of current machine learning-based methods in a low-data scenario still lags behind, compared to languages with ample training data available. In order to push forward the development of low-resource speech recognition, previous studies have proposed various data augmentation techniques – such as self-training (Synnaeve et al., 2020; Xu et al., 2020) – transfer learning utilizing speech representations learned in unsupervised manner from raw speech data (Schneider et al., 2019; Baevski et al., 2020; Hsu et al., 2021), and cross-lingual transfer methods (Toshniwal et al., 2018; Conneau et al., 2021). It has been shown that pretraining speech representations jointly on unlabeled speech data in multiple languages results in models with better downstream performance for low-resource languages than training on each language individually (Conneau et al., 2021), especially if data from related languages is available in relatively large amounts. Recently, Nowakowski et al. (2023) found that the benefits of cross-lingual transfer to an under-resourced language from similar speech varieties also extend to supervised fine-tuning, if there is very little (less than 1 hour) labeled data in the target language available.

If a speech recognition model is trained on data in multiple languages simultaneously and only provided with the acoustic features of speech samples as input, it must implicitly learn to distinguish between different languages appearing in the training data in order to be able to produce a correct output, which can be particularly challenging in low-data scenarios. This requirement can be relaxed by introducing explicit information about the identity of the input language (Toshniwal et al., 2018). In this paper we investigate the possibility of improving the performance of a `wav2vec 2.0` model (Baevski et al., 2020) pretrained on multiple languages, in automatic transcription of an under-resourced language (namely, Sakhalin Ainu) by performing multilingual supervised fine-tuning with a language identifier attached to each speech sample. We find that (i) the proposed method results in lower error

The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

rates than in the case of models fine-tuned without this additional information, (ii) after this modification, using additional labeled data from a single language with similar phonological characteristics as the target language yields models that perform as good as or better than a model fine-tuned on monolingual data only, even if a moderate amount (nearly 10 hours) of labeled target language data is available, and (iii) models fine-tuned on multilingual data with language identifiers produce better results than those fine-tuned without explicit information about language identity, even if this information is absent at inference time. Additionally, we combine multilingual fine-tuning with self-training and find that it results in further improvements.

The remainder of this paper is organized as follows. In Section 2, we provide a short overview of related studies. In Section 3, we introduce our data and describe the details of our system and the training procedure. In Section 4, we analyze the results of our experiments. Finally, Section 5 contains conclusions and ideas for future improvements.

2 Related Work

Previous studies on various NLP problems, including neural machine translation (Johnson et al., 2017; Tang et al., 2020; Eronen et al., 2023) and speech recognition (Toshniwal et al., 2018; Conneau et al., 2021; Nowakowski et al., 2023), found that the information shared among languages in multilingual learning can facilitate the modeling of individual languages (or language pairs, in the case of machine translation), leading to better performance on downstream tasks. This is particularly true for under-resourced languages, especially when additional training data from related language(s) is available (Tang et al., 2020; Conneau et al., 2021; Nowakowski et al., 2023).

The benefits of multilingual training are observed both for systems learned in a supervised manner (Johnson et al., 2017; Toshniwal et al., 2018) and for self-supervised language representation models (Tang et al., 2020; Conneau et al., 2021). Conneau et al. (2021) pretrained a single `wav2vec 2.0` model on unlabeled speech data in 53 languages and tested it in speech recognition, obtaining better performance than with monolingual models, particularly for low-resource languages. They also found that pretraining with additional data from a related language has a stronger positive effect on the model's performance on a low-resource language than using data from a distant language. A study by Nowakowski et al. (2023) also used a multilingual pretrained speech representation model and found that in a scenario where labeled data in the target language is extremely scarce, performing multilingual supervised fine-tuning of such a model using additional transcribed data from a closely related language or an unrelated language with similar phonological characteristics, can lead to further improvements in speech recognition accuracy.

It has been also demonstrated that multilingual neural models perform better when provided with explicit information about language identity of the input. For example, Toshniwal et al. (2018) built a single end-to-end ASR model for 9 different Indian languages and found that feeding a language identifier as an additional input feature resulted in improved performance. Similar results were reported by Abe et al. (2020) who trained a machine translation model jointly on multiple dialects spoken in Japan. They carried out experiments with and without a special token specifying the dialect, attached to the beginning of the input sequence, and observed better performance with the former variant. In this research, we extend the work of Nowakowski et al. (2023) by performing multilingual fine-tuning with language identifiers.

Another technique for improving the effectiveness of low-resource speech recognition which we investigate in this research, is self-training (Synnaeve et al., 2020; Xu et al., 2020, 2021; Khurana et al., 2022; Bartelds et al., 2023). In this approach, the available human-annotated data is first used to train an initial model (often referred to as the 'teacher model'), which is then utilized to generate predictions for a relatively large amount of unlabeled data. Finally, those pseudo-labels are used as an additional training data for the final model (the 'student model'), which – due to having access to more samples from the target distribution – typically exhibits better performance than the teacher model. Recently, it has been shown that self-training is beneficial with models pretrained in a self-supervised manner, as well (Xu et al., 2021; Bartelds et al., 2023).

3 Experiment Setup

3.1 Data

In this research, we are working with actual fieldwork data from a language documentation project.

Table 1: Statistics of human-labeled speech data used in our fine-tuning experiments. We use less than 1h of labeled speech from our target domain (i.e., the Tokoro tapes), less than 10h from our target language (Sakhalin Ainu), and relatively large amounts of data from 3 other speech varieties. For validation and testing we use the remaining two stories from Murasaki and Fujiyama (2010) (namely, `Fu13-700326` and `Fu11-690328`, respectively).

Data	(Main) language/dialect	Total duration (h)
"Wenenekaype" (`Fu12-690401`) (Murasaki and Fujiyama, 2010)	Sakhalin Ainu	0.8
Tuytah (Murasaki and Asai, 2001)	Sakhalin Ainu	8.9
Ainu Language Archive (An=ukokor Aynu ikor oma kenru (National Ainu Museum), 2017–2022)	Hokkaido Ainu	62.2
A Topical Dictionary of Conversational Ainu (National Institute for Japanese Language and Linguistics, 2015)	Hokkaido Ainu	2.3
Common Voice (Japanese) (Ardila et al., 2020)	Japanese	40.6
JSUT (Sonobe et al., 2017)	Japanese	10.3
LibriSpeech (Panayotov et al., 2015)	English	100.6

Specifically, our goal is to develop a system for automatic transcription of unpublished materials from several dialects of the Ainu language formerly spoken in Sakhalin (hereinafter referred to as the "Tokoro tapes", owing to the name of the town in Hokkaido, Japan, where they were recorded), collected in the 1960s and 1970s by professor Kyoko Murasaki in cooperation with Haru Fujiyama and several other speakers of those dialects. The total duration of the recordings is more than 20 hours (or more than 30 hours, if duplicate recordings are counted) which makes them one of the largest existing corpora of Sakhalin Ainu and an invaluable source of knowledge for linguistic and anthropological studies. A subset of the materials has been transcribed, translated to Japanese and published, e.g. in Murasaki and Fujiyama (2010), which includes three different versions of a single folktale, "Wenenekaype", with a total duration of 1.9h. We use the data from Murasaki and Fujiyama (2010) in our experiments as labeled data for model fine-tuning. All human-labeled data used for fine-tuning of our models is listed in Table 1. For monolingual fine-tuning, we use a total of 9.7h of Sakhalin Ainu data obtained from two sources: one story from Murasaki and Fujiyama (2010) (namely, `Fu12-690401`, running for 0.8h) and 8.9h of data from a different collection of Sakhalin Ainu speech recordings, published in Murasaki and Asai (2001). In experiments with multilingual fine-tuning, we add data from three other speech varieties: 64.5h from Hokkaido Ainu, 50.9h from Japanese and 100h of English data. We choose those languages in order to analyze the correlation between language similarity and the effectiveness of our method. Hokkaido Ainu belongs to the same phylogenetic group as our target language. Japanese is not genetically related to Ainu but they share some phonological features, such as the lack of consonant clusters, and quantitative analysis of typological features reveals that both languages are indeed relatively similar (Nowakowski et al., 2023). For comparison, we also use data from English which is both unrelated to Ainu and dissimilar in terms of the phonological system. For validation and testing we use the remaining two stories from Murasaki and Fujiyama (2010) (namely, `Fu13-700326` and `Fu11-690328`, respectively). We preprocess the fine-tuning data in the same way as Nowakowski et al. (2023).

3.2 System Architecture

Fine-tuning with Language Identifiers: Our speech transcription models are built by fine-tuning a multilingual pretrained `wav2vec 2.0` checkpoint on labeled data. Specifically, we use a publicly available model pretrained by Conneau et al. (2021) on 53 languages and further pretrained by Nowakowski et al. (2023) on Ainu language data[1]. We follow the fine-tuning procedure described by Baevski et al. (2020) and Conneau et al. (2021), namely, we add a linear output layer representing

[1] huggingface.co/karolnowakowski/wav2vec2-large-xlsr-53-pretrain-ain

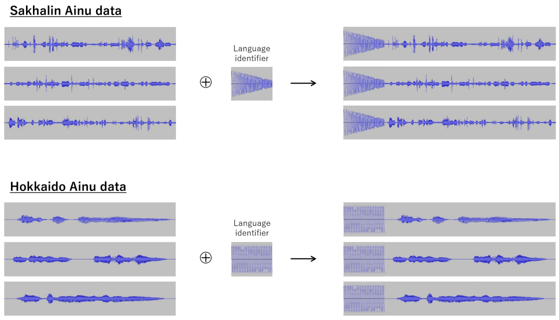

Figure 1: Visualization of our approach to including explicit information about language identity in multilingual fine-tuning data.

the letter vocabulary on top of the pretrained model and train it using Connectionist Temporal Classification (Graves et al., 2006). The only modification that we introduce is the addition of language identifiers. The information about language identity can be either conveyed by a separate language embedding vector concatenated to the model's input at each time step (Östling and Tiedemann, 2017; Toshniwal et al., 2018) or included directly in the data, in the form of an artificial token specifying the language (Tang et al., 2020; Abe et al., 2020). We take the latter approach as it is simpler and requires no changes to the model architecture. Since we are dealing with spoken audio data rather than written text, instead of an artificial textual token we use a fixed length audio clip with artificially generated sound wave (e.g. a sine wave) unique to each language, attached to the beginning of each speech segment in the dataset. The length of each clip is 400 samples (25ms)[2] which is equal to the receptive field of the feature encoder (Baevski et al., 2020). Unless stated otherwise, the language identifiers are used both in training and inference. Figure 1 illustrates our approach to data modification.

Self-training: Apart from multilingual fine-tuning, we carry out experiments with self-training. We use the model fine-tuned on Sakhalin Ainu data[3], released by Nowakowski et al. (2023), to pseudo-label all the speech data from the Tokoro tapes (nearly 32 hours in total, including duplicates) and use the output in addition to human-annotated data to fine-tune the model. Previous studies have shown that the performance gains from self-training can be increased by applying an iterative approach with multiple rounds of pseudo-labeling (Xu et al., 2020; Khurana et al., 2022) and pseudo-label filtering (Park et al., 2020; Khurana et al., 2022). However, in this research we only experiment with a simple approach and leave those methods for future investigation.

[2]In preliminary experiments we tested longer language identifiers (2000 samples), but it resulted in worse performance.

[3]`huggingface.co/karolnowakowski/wav2vec2-large-xlsr-53-ain-sakh`

3.3 Training Settings

Following Nowakowski et al. (2023), we oversample the ”Wenenekaype” data so that it constitutes roughly half of the training set. In the experiments using relatively large amounts of data from speech varieties other than Sakhalin Ainu, we also oversample the Tuytah data by a factor ranging from 6 to 11. Furthermore, in self-training experiments using additional data in Hokkaido Ainu or Japanese, we oversample the pseudo-labeled data by a factor of 2.

We fine-tune our models with a learning rate of 3e-5 and a total batch size of 25.6M samples, for up to 80k updates (for monolingual models and bilingual models fine-tuned on human-transcribed data only) or 120k updates (for models fine-tuned on data from 3 languages and bilingual models fine-tuned with the addition of pseudo-labeled data). We apply early stopping after 20k updates without improvement on the validation set. Concerning other hyperparameters, we follow the configuration for the LARGE model reported by Baevski et al. (2020). We perform all experiments using the `fairseq` library (Ott et al., 2019).

3.4 Inference

We decode the output of the fine-tuned models without a text-based language model, as previous studies did not observe positive effects on speech recognition performance in a low-resource setting, with limited amount of textual data available for language model training (Nowakowski et al., 2023; San et al., 2023). Before evaluation, we preprocess the transcriptions generated by the models by converting all alphabetic characters to lower case.

4 Results and Discussion

Results obtained by models fine-tuned with and without language identifiers are presented in Table 2. We see that using the language identifiers in multilingual fine-tuning generally results in better performance, with the exception of the bilin-

Table 2: Comparison of models fine-tuned with and without language identifiers in speech transcription on the test set. We report Character Error Rates and Word Error Rates. Best results are displayed in bold font. With the exception of the model fine-tuned on Sakhalin Ainu + Japanese, using language identifiers in multilingual fine-tuning leads to significant improvements. Fine-tuning with language identifiers and additional labeled data from a single language with similar phonological characteristics as the target language (namely, Hokkaido Ainu or Japanese) yields models that perform as good as or better than a model fine-tuned on monolingual Sakhalin Ainu data.

Lang. IDs:	NO		YES	
Fine-tuning data	CER	WER	CER	WER
Sakhalin Ainu (“Wenenekaype” + Tuytah)	**9.6**	29.3	N/A	N/A
Sakhalin Ainu + Hokkaido Ainu	10.2	31.4	**9.6** (-0.6)	29.2 (-2.2)
Sakhalin Ainu + Japanese	**9.6**	29.2	9.7 (+0.1)	**29.1** (-0.1)
Sakhalin Ainu + English	14.1	44.2	12.9 (-1.2)	42.1 (-2.1)
Sakhalin Ainu + Hokk. Ainu + Jap.	10.0	31.0	9.8 (-0.2)	29.7 (-1.3)

Table 3: Error rates calculated separately for test samples including Japanese script characters (either in the reference transcriptions or in the model's predictions) and other test samples.

		Lang. IDs: NO			Lang. IDs: YES		
	Fine-tuning data	CER	WER	# samples	CER	WER	# samples
Test samples without Japanese characters	Sakh. Ainu	8.9	28.1	270	N/A	N/A	N/A
	Sakh. Ainu + Hokk. Ainu	9.3	30.0	260	8.9	27.8	265
	Sakh. Ainu + Japanese	8.9	28.3	266	**8.8**	**27.4**	258
	Sakh. Ainu + English	13.1	42.9	281	12.0	40.7	281
	Sakh. Ainu + Hokk. Ainu + Jap.	9.2	29.3	271	9.1	28.4	265
Test samples including Japanese characters	Sakh. Ainu	14.0	37.3	35	N/A	N/A	N/A
	Sakh. Ainu + Hokk. Ainu	14.1	38.1	45	13.7	36.8	40
	Sakh. Ainu + Japanese	**13.6**	**34.9**	39	14.4	37.2	47
	Sakh. Ainu + English	23.5	56.7	24	21.5	56.2	24
	Sakh. Ainu + Hokk. Ainu + Jap.	15.4	42.1	34	14.4	37.0	40

gual model trained with the addition of Japanese data, which achieves relatively good results without language identifiers and no significant change is observed after adding them. We hypothesize that this behavior is related to the fact that the Ainu data, including the test set used in our experiments, contains many code-switched fragments in Japanese. Namely, a model fine-tuned not only on Ainu speech, but also on monolingual Japanese data, might be able to learn a better representation of the latter language and as a result, have easier time deciding whether a certain part of an utterance is in Ainu or in Japanese as well as transcribing such code-switched fragments. In order to verify if this is true, we calculate the error rates separately for test samples including Japanese script characters (either in the reference transcriptions or in the model's predictions) and samples without any code-switching. Analysis of the results (presented in Table 3) seems to partially confirm our hypothesis: while all other models fine-tuned on multilingual data without language identifiers perform worse on test samples with Japanese characters than a monolingual Sakhalin Ainu model, for the model fine-tuned with Japanese data we observe an improvement. On the other hand, it also yields the best results among multilingual models for samples without Japanese script, which indicates that its relatively good performance cannot be fully explained only by code-switching.

Models fine-tuned on Sakhalin Ainu + Japanese and Sakhalin Ainu + Hokkaido Ainu (in the latter case, only when training with language identifiers) perform competitively to the monolingual Sakhalin Ainu model, whereas fine-tuning with English data leads to significantly worse results. This outcome

Table 4: Results of the experiments using pseudo-labels generated through self-training. Best results are displayed in bold font. The best overall results are achieved by combining multilingual and pseudo-labeled data and fine-tuning with language identifiers.

Lang. IDs:	NO		YES	
Fine-tuning data	CER	WER	CER	WER
Sakhalin Ainu (incl. pseudo-labels)	9.4	29.0	N/A	N/A
Sakh. Ainu (incl. pseudo-labels) + Hokk. Ainu	9.6	29.0	**9.1**	**28.1**
Sakh. Ainu (incl. pseudo-labels) + Japanese	9.2	28.2	9.2	28.4

Table 5: Comparison of the results obtained by (i) not using language identifiers at all, (ii) training with language identifiers but testing on data without them, and (iii) using data with language identifiers both in training and inference. In most cases applying language identifiers at training time only gives better results than not using them at all.

Lang. IDs:	NO		YES (train.)		YES (train.+infer.)	
Fine-tuning data	CER	WER	CER	WER	CER	WER
Sakhalin Ainu + Hokkaido Ainu	10.2	31.4	9.7	29.6	9.6	29.2
Sakhalin Ainu + Japanese	9.6	29.2	9.7	29.3	9.7	29.1
Sakhalin Ainu + English	14.1	44.2	12.6	40.6	12.9	42.1
Sakhalin Ainu + Hokk. Ainu + Jap.	10.0	31.0	9.9	29.7	9.8	29.7
Sakh. Ainu (incl. pseudo-labels) + Hokk. Ainu	9.6	29.0	9.2	**28.0**	**9.1**	28.1
Sakh. Ainu (incl. pseudo-labels) + Japanese	9.2	28.2	9.2	28.2	9.2	28.4

confirms the correlation between language similarity and the effectiveness of cross-lingual transfer, also observed in previous studies. Fine-tuning with data from two additional languages (specifically, Hokkaido Ainu and Japanese) at the same time does not achieve the best results, indicating that the potential benefits from additional cross-lingual signal are outweighed by the reduction in the number of model parameters per language.

Results of the self-training experiment are shown in Table 4. Similarly to previous studies, we observe improved performance after training with pseudo-labeled data. Concerning the model fine-tuned on Sakhalin Ainu data only, self-training provides a 2% relative improvement of CER compared to the supervised-only counterpart. Combining self-training and multilingual data results in further improvements. The best overall results are achieved by fine-tuning on human-annotated Sakhalin Ainu and Hokkaido Ainu data as well as pseudo-labeled Sakhalin Ainu data and using language identifiers. This yields a 5% relative improvement of CER compared to the baseline model fine-tuned on monolingual Sakhalin Ainu data.

While in this research we are mainly focusing on a single language and only leveraging data in other speech varieties to improve the speech recognition performance on that language, there are also many studies aiming to develop systems that can be applied to multiple languages (Toshniwal et al., 2018; Radford et al., 2022; Pratap et al., 2023). One potential limitation of the proposed method using language identifiers is that the information about language identity may not be always available beforehand in real-world use in a multilingual setting. However, in our experiments we find that the lack of this information at inference time does not necessarily invalidate our approach. In Table 5 we compare the results obtained by (i) not using language identifiers at all, (ii) training with language identifiers but testing on data without them, and (iii) using data with language identifiers both in training and inference. We observe that in most cases, applying a model fine-tuned on data including language identifiers still yields significantly better results, even if they are not available at inference time. The model producing the lowest error rates on our test set yields nearly identical results in inference with and without language identifiers, and in the case of the model fine-tuned with the addition of English data, predictions made for the data without language identifiers are more accurate than with them. These results indicate that the additional knowledge about the relationships and differences between the languages used in fine-tuning, learned by the agency of the language identifiers, can be to a large extent reused in inference regardless of their presence in the new data. This would mean that our approach could be used to improve multilingual speech recognition without sacrificing versatility, but additional experiments on a larger number of languages are needed to verify our observations.

5 Conclusions and Future Work

We have demonstrated how low-resource speech recognition accuracy can be improved by leveraging labeled data from additional languages as well as unlabeled target language data. Firstly, we improved the effectiveness of multilingual supervised fine-tuning of a pretrained speech representation model by augmenting the data with language identifiers. Our results showed that fine-tuning on data preprocessed this way and including additional samples from a single language with similar phonological characteristics as the target language, produces models performing on par with or better than a model fine-tuned using monolingual data only, even if a moderate amount of labeled target language data is available. Furthermore, we found that supplying the model with the information about language identity at training time is helpful even if it is not provided later during inference, meaning that our approach could be potentially useful also in multilingual settings where such information is not available beforehand. Finally, we used unlabeled speech data to perform self-training and found that fine-tuning a speech recognition model jointly on a combination of multilingual data and pseudo-labeled target language data yields superior performance compared to using any of the two augmentation techniques individually.

In the future we will explore alternative methods for supplying the information about language identity, namely, additional language embedding vectors attached to the input of the encoder and/or the decoder at each time step. We also plan to enhance our self-training procedure by applying iterative pseudo-labeling and pseudo-label filtering techniques.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP22K17952.

References

Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2020. Multi-dialect neural machine translation for 48 low-resource japanese dialects. *Journal of Natural Language Processing*, 27(4):781–800.

An=ukokor Aynu ikor oma kenru (National Ainu Museum). 2017–2022. Ainu-go Ākaibu [Ainu Language Archive].

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 *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *ArXiv*, abs/2006.11477.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-lingual Representation Learning for Speech Recognition. In *Interspeech*.

Juuso Eronen, Michal Ptaszynski, Karol Nowakowski, Zheng Lin Chia, and Fumito Masui. 2023. Improving polish to english neural machine translation with transfer learning: Effects of data volume and language similarity. In *Workshop on Multilingual, Multimodal and Multitask Language Generation*, Tampere, Finland.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Magic dust for cross-lingual adaptation of monolingual wav2vec-2.0. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6647–6651.

Kyōko Murasaki and Take Asai. 2001. *Karafuto Ainu no mukashi-banashi: Tuytah [Sakhalin Ainu folktales: Tuytah]*. Sōfūkan, Tokyo.

Kyōko Murasaki and Haru Fujiyama. 2010. *Sakhalin Ainu Folktales (ucaskuma): Wenenekaype*, volume 2 of *ILCAA Norheast Asian Studies*. Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies, Tokyo.

National Institute for Japanese Language and Linguistics. 2015. A Topical Dictionary of Conversational Ainu.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le. 2020. Improved Noisy Student Training for Automatic Speech Recognition. In *Proc. Interspeech 2020*, pages 2817–2821.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Nay San, Martijn Bartelds, Blaine Billings, Ella de Falco, Hendi Feriza, Johan Safri, Wawan Sahrozi, Ben Foley, Bradley McDonnell, and Dan Jurafsky. 2023. Leveraging supplementary text data to kick-start automatic speech recognition system development with limited transcriptions. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–6, Remote. Association for Computational Linguistics.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-training for Speech Recognition. In *INTERSPEECH*.

Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2017. JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis. *ArXiv*, abs/1711.00354.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2020. End-to-end asr: from supervised to semi-supervised learning with modern architectures.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *ArXiv*, abs/2008.00401.

Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. 2018. Multilingual speech recognition with a single end-to-end model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4904–4908.

Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034.

Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert. 2020. Iterative Pseudo-Labeling for Speech Recognition. In *Proc. Interspeech 2020*, pages 1006–1010.