EXPLORING THE IMPACT OF DATA QUANTITY ON ASR IN EXTREMELY LOW-RESOURCE LANGUAGES

Yao-Fei Cheng[1], *Li-Wei Chen*[2,3], *Hung-Shin Lee*[3], *Hsin-Min Wang*[4]

[1]University of Washington, [2]National Tsing Hua University
[3]United Link Co., Ltd., [4]Academia Sinica
nlp5566@uw.edu

ABSTRACT

This study investigates the efficacy of data augmentation techniques for low-resource automatic speech recognition (ASR), focusing on two endangered Austronesian languages, Amis and Seediq. Recognizing the potential of self-supervised learning (SSL) in low-resource settings, we explore the impact of data volume on the continued pre-training of SSL models. We propose a novel data-selection scheme leveraging a multilingual corpus to augment the limited target language data. This scheme utilizes a language classifier to extract utterance embeddings and employs one-class classifiers to identify utterances phonetically and phonologically proximate to the target languages. Utterances are ranked and selected based on their decision scores, ensuring the inclusion of highly relevant data in the SSL-ASR pipeline. Our experimental results demonstrate the effectiveness of this approach, yielding substantial improvements in ASR performance for both Amis and Seediq. These findings underscore the feasibility and promise of data augmentation through cross-lingual transfer learning for low-resource language ASR.

Index Terms— self-supervised learning, low-resource language, automatic speech recognition

1. INTRODUCTION

The development of contemporary end-to-end automatic speech recognition (ASR) systems has achieved remarkable success in high-resource languages such as English, French, and Chinese, largely due to the availability of extensive speech-text paired data [1, 2]. However, this abundance does not extend to endangered languages like Amis and Seediq, where acquiring high-quality transcriptions is particularly challenging and resource-intensive. The scarcity of annotated data, coupled with factors such as limited speaker populations and non-standardized writing systems, poses significant obstacles to building effective ASR systems. In response, self-supervised learning (SSL) has emerged as a promising approach to mitigate data sparsity in low-resource ASR settings [3–5]. SSL models, pre-trained on large amounts of unlabeled speech, can be fine-tuned with limited labeled data. However, the pre-training stage typically requires far more data than supervised methods—for instance, the `wav2vec 2.0 base` model demands 960 hours of speech for effective pre-training [4]. This makes training SSL models from scratch impractical for endangered languages, where large-scale data collection is often infeasible.

To address the broader issue of resource scarcity in natural language processing, researchers have explored various strategies, including language representation tools like `lang2vec` [6] and data augmentation techniques [7, 8]. However, these approaches face limitations when applied to low-resourced languages like Amis and Seediq. `lang2vec` currently lacks support for these languages, while data augmentation often depends on textual data, which is scarce or unavailable for many spoken and low-resource languages.

Consequently, recent research has shifted towards leveraging multilingual SSL models for low-resource ASR [9, 10]. This approach capitalizes on phonetic similarities across languages, enabling knowledge transfer and improved transcription accuracy during fine-tuning. Further emphasizing the potential of this approach, Nowakowski et al. demonstrated the effectiveness of continuing pre-training with target language data, albeit with a significant data requirement exceeding 234 hours [11]. Their work also explored leveraging phonetically similar languages for data augmentation, achieving promising results. Similarly, San et al. proposed utilizing donor speech data from related languages for fine-tuning, highlighting the potential of cross-lingual transfer learning in low-resource settings [12].

This research diverges from previous studies by exploring the feasibility of achieving robust ASR performance with minimal paired and unpaired data for the target language. Focusing on the severely low-resource Austronesian languages, Amis and Seediq, we aim to develop effective ASR systems using a modest multilingual corpus and limited paired data (less than one hour per language). Our approach centers on a novel data-selection strategy designed to identify and incorporate utterances from the multilingual corpus that exhibit close phonetic and phonological similarity to the target languages, thus refining the SSL-ASR pipeline. We achieve this

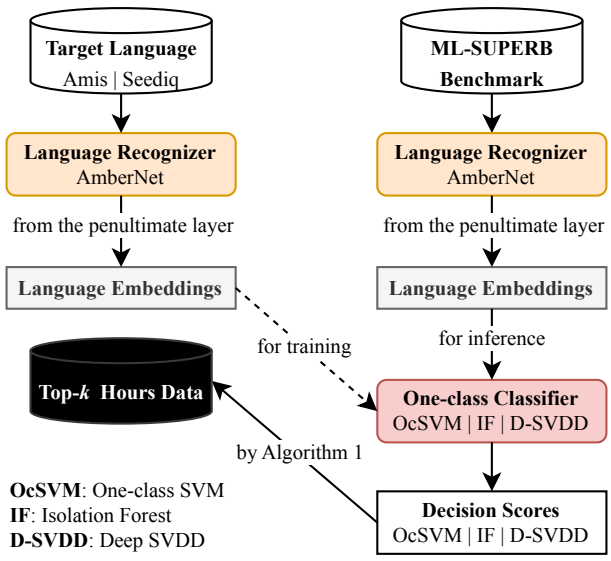


Fig. 1. Diagram for picking the top-k hours data.

by first employing a language classifier to extract language-specific embeddings for each utterance [13, 14]. Then, leveraging a suite of one-class classifiers – One-class SVM [15], Isolation Forest [16], and Deep SVDD [17] – we rank and select utterances based on their decision scores, prioritizing those most similar to the target languages. Our results demonstrate the effectiveness of this approach for both Amis and Seediq, yielding promising ASR performance despite the significant data constraints.

2. DATA DESCRIPTION

Taiwan is home to 16 indigenous ethnic groups, encompassing 42 dialects. The Taiwan Council of Indigenous Peoples introduced Klokah[1], an e-learning platform with abundant audio materials to foster native language use among the younger generation. For ASR system development, we curated a multi-speaker dataset from two subjects: reading-and-writing and contextual-language, chosen based on their suitable audio lengths. We selected Amis and Seediq for experimentation, allocating 1 hour or 10 minutes of speech for training and 10 minutes each for validation and testing.

2.1. Amis

Amis has approximately $218,500$ speakers in Taiwan. It is one of the endangered languages identified by the United Nations Educational, Scientific, and Cultural Organization (UNESCO). Its writing system consists of 21 Latin letters and one digraph (ng). Another 5 Latin letters (b, g, j, q, and v) are used only for loanwords. The epiglottal stop /ʔ/ and glottal stop /ʔ/ are marked as ’ and ˆ, respectively. The vowel length /ː/ is marked as a colon symbol : , meaning semantic emphasis.

Amis has five dialect groups: ’Amisay a Pangcah, Siwkolan ’Amis, Pasawalian Pangcah, Farangaw Amis, and Palidaw ’Amis. Its written system can be vague, i.e., the same pronunciation has different written symbols. For instance, only ’Amisay a Pangcah uses the symbol *o* to represent /o/, and the other four groups use the symbol *u* to represent /o/.

2.2. Seediq

Seediq is also an endangered language identified by UNESCO. Its language population in Taiwan is about $10,970$. The written system of Seediq consists of 23 Latin letters and seven diphthongs (aw, ay, ey, ow, uy, ug, and ng). The Seediq text is written in syllabic units of semantic words, with an apostrophe ’ to mark where the syllables in the same word have to be separated because of the sound level or where there is still a problem with the attribution of the pronunciation, or where the vowel sound /ʔ/ is indicated.

3. METHOD

Inspired by Nowakowski et al.’s work [11], which demonstrated that continued pre-training with a related language (Japanese) significantly improved ASR performance for Ainu, a low-resource language[2], we propose a novel data augmentation strategy. Since `lang2vec` [6]—the tool they used to identify related languages—does not support our target languages, we present an alternative approach: leveraging multilingual fine-tuning of an SSL model with data from languages that share phonetic and phonological similarities with our target languages. This strategy aims to enhance speech recognition performance by enriching the training data with acoustically similar utterances from related languages.

Our methodology, illustrated in Fig. 1, employs a two-phase approach. In the first phase, a neural spoken language identification model analyzes utterances to extract language-specific embeddings [13, 14] (orange box in Fig. 1). These embeddings, derived from the penultimate layer of the classifier, encapsulate rich phonetic and phonological information. The second phase leverages these embeddings to train a suite of one-class classifiers (red box in Fig. 1). These classifiers are then employed to identify and select utterances from a larger multilingual dataset that exhibit high similarity to the target language, thereby enriching the training data for subsequent SSL model fine-tuning.

First, One-class Support Vector Machine (OcSVM) aims to construct a hyperplane in a high-dimensional feature space that encloses the majority of data points while maximizing the distance between this hyperplane and the origin [15]. This

[1] https://web.klokah.tw

[2] Interestingly, the Ainu speech data utilized for continued pre-training of the SSL model exceeded 200 hours.

Algorithm 1 Multi-list Utterance Selection

Given three lists of utterances U_1, U_2, and U_3 sorted by decision scores of Deep SVDD, One-class SVM, and Isolation Forest
Initial ranking limit L_0, final list R, and specified hours k.
$L \leftarrow L_0$
while the total length in $R < k$ hours **do**
    $\widehat{U_1} \leftarrow U_1[:L]$
    $\widehat{U_2} \leftarrow U_2[:L]$
    $\widehat{U_3} \leftarrow U_3[:L]$
    for $u \in \widehat{U_1}$ **do**
        if $u \notin R$ **and** $u \in \widehat{U_2}$ **and** $u \in \widehat{U_3}$ **then**
            R.append(u)
        end if
    end for
    $L \leftarrow L + L_0$
end while

approach effectively separates the data from regions with a sparse data density. Data points falling outside the hyperplane's boundary are then classified as anomalies.

Second, Isolation Forest (IF) leverages an ensemble of isolation trees to detect anomalies [18]. Each tree is constructed by recursively partitioning the data based on randomly selected features and split points. The algorithm aggregates these path lengths to provide an anomaly score for each data point.

Finally, Deep Support Vector Data Description (D-SVDD) utilizes a deep neural network to learn a hypersphere in the feature space that encapsulates the training data while minimizing the volume of this hypersphere [17].

Given the potential for variation in the rankings produced by each classifier, we introduce an ensemble algorithm to harmonize their predictions and enhance the reliability of the utterance selection process (see Algorithm 1). This algorithm prioritizes consistently flagged utterances similar to those of all three classifiers, ensuring a more robust and confident data selection for augmenting the SSL model's training set.

Algorithm 1 aims to create a list of utterances (R), given three pre-sorted lists (U_1, U_2, U_3) based on the scores of Deep SVDD, One-class SVM, and Isolation Forest. The algorithm prioritizes utterances appearing high in all three lists.

The process begins by initializing L with an initial ranking limit (L_0). It iteratively expands L until the total duration of utterances in R meets a predefined time limit (k hours). Each iteration involves selecting top L utterances from each list (U_1, U_2, U_3), denoted as $\widehat{U_1}$, $\widehat{U_2}$, and $\widehat{U_3}$. Subsequently, each utterance (u) in $\widehat{U_1}$ is evaluated. If u is not already in R and is present in both $\widehat{U_2}$ and $\widehat{U_3}$, it signifies high ranking across all three lists and is appended to R. After processing all utterances in $\widehat{U_1}$, L is incremented by L_0 for the next iteration. This cycle continues until the time constraint of R is met.

4. EXPERIMENTS

Our experimental procedure unfolds as follows: In addition to leveraging the existing, limited volume of target language data, we employ the top-k hours of sampled data from the ML-SUPERB multilingual corpus. This selection is facilitated by a one-class classifier or Algorithm 1, intended for the continued pre-training of the SSL model. The refined model is then utilized to generate SSL features, enabling fine-tuning of the downstream ASR model using target language speech data accompanied by transcriptions. For our framework, Fairseq is used for pre-training, while both S3PRL and ESPnet are employed for downstream ASR tasks [19–21].

4.1. Continued Pre-training

Following the methodology outlined by Nowakowski et al., we continued to pre-train the entire SSL model with the curated data. The training utterances were curated from the ML-SUPERB benchmark using one-class classifiers. This process involved $100,000$ updates, including $10,000$ warmup steps, with a batch size of 256 and a learning rate set at 10^{-4}. All settings were aligned with those of the `wav2vec 2.0 large` model. The optimal configuration was identified using validation data from the ML-SUPERB benchmark.

4.2. One-class Classification

Our process commenced using NVIDIA's pre-trained TitaNet-LID to derive language embeddings for each utterance [14]. This model attains a 7.0% error rate on the 103 languages test set. The obtained embeddings are situated in a 512-dimensional space, originating from the penultimate layer.

For the classifiers, we utilized scikit-learn to train the One-class SVM and Isolation Forest, largely sticking to the default parameters. In the case of Deep SVDD, we implemented a modification by substituting the original 2D convolution for a 1D variant. The autoencoder was pre-trained for $2,500$ epochs with a learning rate of 10^{-2}, and the encoder was further trained for $1,000$ epochs at a reduced learning rate of 10^{-3}. Negative samples were selected from the ML-SUPERB training data and combined with the original test utterances from Amis and Seediq as positive samples. Our experiments demonstrated the superior performance of Deep SVDD in this setup.

4.3. Fine-tuning

The downstream ASR task followed ML-SUPERB's design. The frozen SSL representations were weighted summed with

[3]https://huggingface.co/TencentGameMate/chinese-wav2vec2-base
[4]https://huggingface.co/TencentGameMate/chinese-wav2vec2-large
[5]https://huggingface.co/facebook/wav2vec2-xls-r-300m
[6]https://huggingface.co/TencentGameMate/chinese-hubert-base
[7]https://huggingface.co/TencentGameMate/chinese-hubert-large

SSL/Acoustic Features	# Params. (M)	# Languages	Amis-10min	Amis-1h	Seediq-10min	Seediq-1h
Mel-filterbank	N/A	N/A	40.5	28.5	49.4	36.0
wav2vec2-base [4]	95	1	14.6	9.8	19.8	11.9
wav2vec2-large [4]	317	1	13.4	9.5	18.5	11.4
robust-wav2vec2-large [22]	317	1	13.5	9.3	18.6	11.5
chinese-wav2vec2-base[3][23]	95	1	14.4	9.4	18.7	11.2
chinese-wav2vec2-large[4][23]	317	1	89.4	10.0	22.2	12.5
chinese-wav2vec2-large* [23]	317	1	15.3	10.0	21.6	12.3
wav2vec2-base-23 [10]	95	23	14.5	9.2	19.6	12.1
wav2vec2-large-23 [10]	317	23	103.4	9.6	19.2	11.3
wav2vec2-large-23* [10]	317	23	13.5	9.2	18.4	11.4
XLSR-53 [24]	317	53	81.4	10.4	18.6	14.0
XLSR-300M[5][25]	317	128	90.8	7.7	14.0	8.4
XLSR-300M* [25]	317	128	10.9	7.9	13.7	8.6
HuBERT-base [5]	95	1	12.6	8.8	18.2	11.0
HuBERT-large [5]	317	1	80.6	7.4	84.1	8.9
HuBERT-large* [5]	317	1	10.3	7.5	14.0	9.1
chinese-hubert-base[6][23]	317	1	13.2	9.0	18.4	10.9
chinese-hubert-large[7][23]	317	1	11.0	7.3	14.7	8.1
mHuBERT-base [26]	95	3	12.9	8.9	17.3	11.5

Table 1. CER (%) evaluated on Amis and Seediq using SSL models. * indicates a smaller downstream ASR model. 10min and 1h mean the amount of training data used in downstream ASR models.

learnable weights. The ASR model consisted of a CNN layer that downsamples the SSL feature sequence to half the original length and two Transformer encoders with an attention dimension of 256, a feedforward layer with a dimension of $1,024$, and 8 heads. The model is trained with connectionist temporal classification (CTC) and Adam optimizer with a learning rate of 10^{-4} and 10^{-6} weight decay [27, 28]. We set dropout to 0.1, batch size to 8, and gradient accumulation to 4. SpecAugment was applied to SSL features [29].

5. RESULTS

5.1. ASR Results on Amis and Seediq

Table 1 presents an overview of ASR results for both Amis and Seediq, where the downstream ASR models are directly fine-tuned from SSL models in various settings (see Section 4.3 for details). We have compared models that were pre-trained with utterances in multilingual, English-only, and Mandarin Chinese-only. Additionally, we evaluated the effectiveness of the model architecture and parameter size.

`chinese-hubert-large` (Hubert-large fine-tuned with Mandarin Chinese data) consistently outperformed other models across different training data volumes, except for the Seediq 10-minute task. This superior performance was observed despite using identical training data for `wav2vec2` (see results of chinese-hubert-large vs. chinese-wav2vec2-large), suggesting potential advantages of the `HuBERT` architecture for these low-resource scenarios. Furthermore, `XLSR-300M` demonstrated comparable effectiveness, achieving the lowest CER of 13.7% on the Seediq 10-minute task. These results highlight several noteworthy trends:

- `HuBERT` generally outperform `wav2vec2` when trained on the same data.
- Larger SSL models yield modest but consistent gains, especially in the 1-hour training setup.
- Multilingual pre-training contributes positively to ASR performance in low-resource settings.

As observed in Table 1, large SSL models, while generally outperforming their smaller counterparts, exhibit a tendency to overfit when trained on limited data. This overfitting was particularly evident for models like `wav2vec2-large-23`, `chinese-wav2vec2-large`, `XLSR-53`, `XLSR-300M`, and `HuBERT-large`. To mitigate this issue, we modified the downstream ASR model architecture by halving the number of attention heads to 4, reducing the feedforward dimension to 512, and increasing the dropout rate to 0.3. These adjustments, as demonstrated in Table 1 (models denoted by *), enabled the smaller ASR model to achieve comparable performance even when paired with larger SSL models, effectively addressing the overfitting problem.

5.2. Proposed Data Augmentation Method

This section explores the effectiveness of our proposed data augmentation method. Specifically, we attempt to improve the downstream ASR results that we discussed previously (see Section 5.1) by leveraging our proposed data augmentation method (see Section 3 for more details).

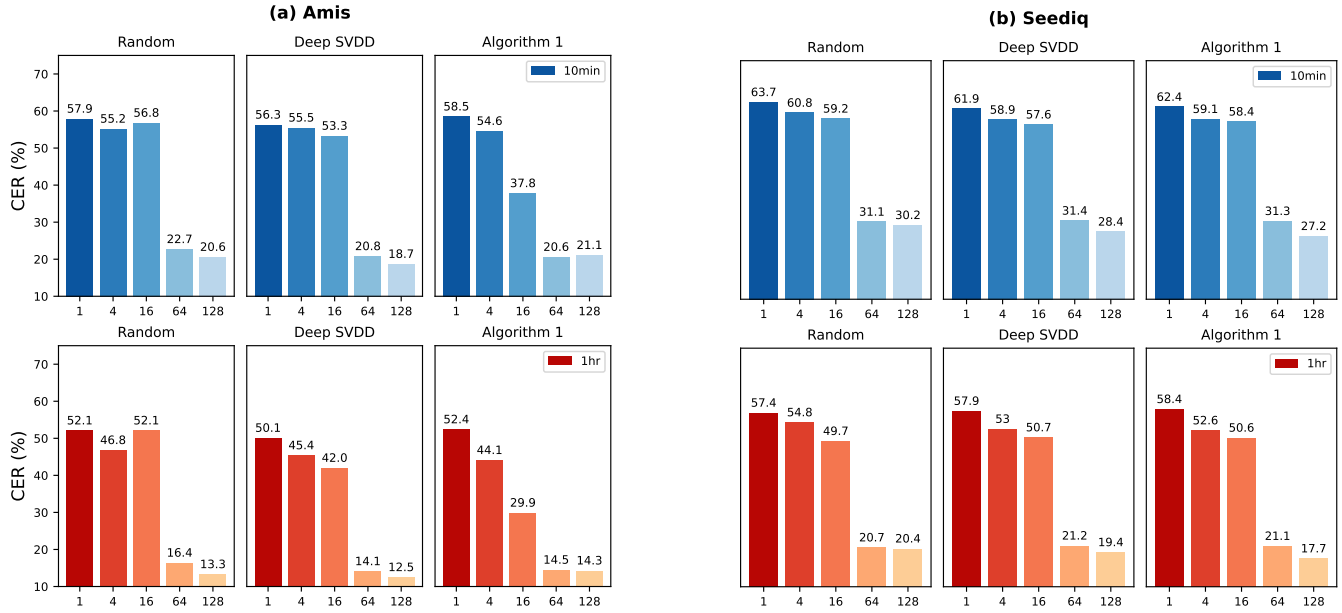


Fig. 2. The effect of data amount and selection on continued pre-training XLSR-128 for Amis and Seediq. The x-axis represents the amount of sampled data (in hours), and '1' means 1 hour of non-target language data is used together with 1 hour of target language data in continued pre-training. The shades in blue and red represent the amount of pre-training data.

While `XLSR-300M` did not consistently achieve the highest performance across all experimental settings, we selected it as our primary benchmark model. This decision stems from the fact that `XLSR-300M` is a multilingual model not specifically pre-trained on languages closely related to Amis and Seediq, making it a more generalizable choice for broader low-resource language research where such closely related pre-trained models might not be readily available.

Fig. 2 illustrates the impact of data volume and selection strategy on the performance of `XLSR-300M` after continued pre-training for both Amis and Seediq. We evaluated three data selection approaches: random sampling, Deep SVDD, and our proposed ensemble algorithm (Algorithm 1). As expected, increasing the amount of pre-training data generally leads to improved ASR performance, mitigating the risk of overfitting. Notably, using only 1 hour of data from ML-SUPERB resulted in suboptimal performance across all selection methods, underscoring the importance of sufficient training data volume. Expanding the pre-training data to 64 hours significantly reduced character error rate (CER) for both languages, highlighting the consistent benefit of larger datasets across languages.

Our proposed methods (SVDD and Algorithm 1) consistently outperform the random selection baseline across nearly all experiments. The only exception is noted in the Seediq 1-hour experiments, where our methods demonstrate their effectiveness after selecting 128 hours of non-target language data. Specifically, in both the Seediq 10-minute and 1-hour scenarios with 128 hours of data selection, Algorithm 1 consistently yields superior results compared to both random selection and SVDD. In contrast, SVDD performs the best in Amis 10-minute and 1-hour scenarios with 128 hours of data selection. However, increasing the data size to 128 hours yielded only marginal improvements, suggesting that the model might be approaching saturation given the limited representation of the target languages within the available multilingual speech data.

Our findings for Amis reveal that when pre-training data is limited (e.g., 16 hours), utilizing Algorithm 1) for data selection significantly outperforms both random sampling and Deep SVDD, particularly in the 10-minute and 1-hour training scenarios. Moreover, Fig. 2 demonstrates that the SSL model pre-trained with data selected using our algorithm exhibits greater robustness across almost all experimental conditions than models trained with randomly selected or Deep SVDD-sampled data.

6. CONCLUSIONS

This study introduces speech corpora for two low-resource Austronesian languages, Amis and Seediq, and presents a data augmentation strategy for enhancing low-resource ASR. We propose leveraging a novel ensemble algorithm to select acoustically similar utterances from a multilingual corpus for continued pre-training of the `XLSR-300M` model. Our results demonstrate that while the performance improvements are currently modest, the positive correlation between pre-training data volume and ASR accuracy highlights the potential of this approach.

7. References

[1] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020.

[2] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-Branchformer: Branchformer with Enhanced merging for speech recognition," in *Proc. SLT*, 2022.

[3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[5] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[6] P. Littell, D. R. Mortensen, K. Lin, K. Kairis, C. Turner, and L. Levin, "URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors," in *Proc. EACL*, 2017.

[7] M. Bartelds, D. Jurafsky, and M. Wieling, "Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation," in *Proc. ACL*, 2023.

[8] J. Safri, W. Sahrozi, B. Foley, B. McDonnell, and D. Jurafsky, "Leveraging supplementary text data to kick-start automatic speech recognition system development with limited transcriptions," in *Proc. the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 2023.

[9] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. Interspeech*, 2021.

[10] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. ACL*, 2021.

[11] K. Nowakowski, M. Ptaszynski, K. Murasaki, and J. Nieuważny, "Adapting multilingual speech representation model for a enw, under-resourced language through multilingual fine-tuning and continued pretraining," *Inf. Process. Manag.*, vol. 60, no. 2, p. 103 148, 2023.

[12] N. San, G. Paraskevopoulos, A. Arora, X. He, P. Kaur, O. Adams, and D. Jurafsky, "Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens," in *Proc. SIGTYP*, 2024.

[13] F. Jia, N. R. Koluguri, J. Balam, and B. Ginsburg, "AmberNet: A compact end-to-end model for spoken language identification," in *Arxiv preprint arXiv:2210.15781*, 2022.

[14] F. Jia, N. R. Koluguri, J. Balam, and B. Ginsburg, "A compact end-to-end model with local and global context for spoken language identification," in *Proc. Interspeech*, 2023.

[15] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural. Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.

[16] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. ICDM*, 2008.

[17] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *Proc. ICML*, 2018.

[18] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 1, pp. 1–39, 2012.

[19] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "Fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. NAACL*, 2019.

[20] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018.

[21] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021.

[22] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Proc. Interspeech*, 2021.

[23] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, D. Wu, and Z. Peng, "WenetSpeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *Proc. ICASSP*, 2022.

[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual depresentation learning at scale," in *Proc. ACL*, 2020.

[25] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022.

[26] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, Y. Adi, J. Pino, J. Gu, and W.-N. Hsu, "Textless speech-to-speech translation on real data," in *Proc. NAACL*, 2022.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[28] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.

[29] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019.