Towards Robust Speech Representation Learning for Thousands of Languages

William Chen[1], **Wangyou Zhang**[1,2], **Yifan Peng**[1] **Xinjian Li**[1]
Jinchuan Tian[1], **Jiatong Shi**[1], **Xuankai Chang**[1], **Soumi Maiti**[1]
Karen Livescu[1,3], **Shinji Watanabe**[1]
[1] Carnegie Mellon University, [2] Shanghai Jiaotong University
[3] Toyota Technological Institute at Chicago
williamchen@cmu.edu

Abstract

Self-supervised learning (SSL) has helped extend speech technologies to more languages by reducing the need for labeled data. However, models are still far from supporting the world's 7000+ languages. We propose XEUS, a Cross-lingual Encoder for Universal Speech, trained on over 1 million hours of data across 4057 languages, extending the language coverage of SSL models 4-fold. We combine 1 million hours of speech from existing publicly accessible corpora with a newly created corpus of 7400+ hours from 4057 languages, which will be publicly released. To handle the diverse conditions of multilingual speech data, we augment the typical SSL masked prediction approach with a novel dereverberation objective, increasing robustness. We evaluate XEUS on several benchmarks, and show that it consistently outperforms or achieves comparable results to state-of-the-art (SOTA) SSL models across a variety of tasks. XEUS sets a new SOTA on the ML-SUPERB benchmark: it outperforms MMS 1B and w2v-BERT 2.0 v2 by 0.8% and 4.4% respectively, despite having less parameters or pre-training data. Checkpoints, code, and data are found in https://www.wavlab.org/activities/2024/xeus/.

1 Introduction

Our planet is home to over 7000 languages (Austin and Sallabank, 2011), yet most speech processing models are only capable of serving at most 100-150 of them (Barrault et al., 2023b; Radford et al., 2023). The biggest constraint in supporting more of languages is the lack of transcribed speech: only around half of the world's languages have a formal writing system (Ethnologue, 2017), and even fewer of them have the resources to support the scale of annotated data required for training neural models. A common approach to address this limitation is self-supervised learning (SSL) on large amounts of unlabeled multilingual speech (Zhang et al., 2023; Black, 2019; Li et al., 2022), which allows for strong downstream performance even when access to annotated data is limited.

While SSL models have more relaxed data requirements relative to supervised models, few works have fully exploited this aspect to scale models to more languages. In fact, most multilingual SSL models remain in the 50-150 language range of coverage (Babu et al., 2022; Chen et al., 2023b; Chiu et al., 2022), reducing the benefits of this advantage. The MMS project (Pratap et al., 2023) sought to address this issue by directly crawling data for more languages, scaling SSL pre-training to 1,000+ languages. The authors collected speech across 1,406 languages and used it to train an SSL model, showing state-of-the-art (SOTA) results after fine-tuning on multilingual automatic speech recognition (ASR) and 3,900-way language identification (LID). While the MMS models were publicly released, the crawled data was not, preventing it from being used in future work and thus the models from being reproduced (Table 1).

An additional issue that is relatively unexplored in SSL research is robustness to noisy data. This aspect is important for multilingual models, since the available recordings of low-resource languages tend to be particularly noisy (Ardila et al., 2020). The issue is exacerbated by the fact that existing multilingual SSL corpora lack diversity not only in languages but also in speaking style and recording conditions. WavLM (Chen et al., 2022) and WavLabLM (Chen et al., 2023b) tackled this issue by simulating noisy conditions during training, overlapping utterances or adding acoustic noise to simulate multi-speaker and noisy environments respectively. While effective, we believe that this technique can be improved to cover an even wider variety of recording conditions.

Our goal is to thus build a universal speech encoder that can handle both linguistically and acoustically diverse speech. To achieve this, we

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10205–10224
November 12-16, 2024 ©2024 Association for Computational Linguistics

propose XEUS (pronounced Zeus) — a Cross-lingual Encoder for Universal Speech. XEUS is an E-Branchformer (Kim et al., 2023) encoder pre-trained on over 1 million hours of publicly available data across a wide variety of recording conditions. We first curate the data from 37 existing corpora to ensure a diverse selection of speech and recording conditions not often found in standard ASR datasets, including but not limited to spontaneous speech, accented speech, code-switching, indigenous languages, and singing voices. We expand the language coverage of XEUS by introducing a new SSL corpus that uses data sources previously unseen in speech literature. This corpus, which will be publicly released, contains 7,413 hours of unlabeled audio across 4,057 languages, the widest coverage of any speech processing dataset.

To enhance the model's robustness, XEUS is also pre-trained with a novel SSL objective of acoustic dereverberation, which requires the model to predict clean discrete phonetic pseudo-labels from simulated reverberant audio. By combining this objective with HuBERT-style masked prediction (Hsu et al., 2021) and WavLM-style denoising (Chen et al., 2022), XEUS is designed to be the next step towards a truly universal speech encoder for any language or recording condition.

In our downstream evaluations, we find that XEUS consistently improves over SOTA SSL models across a wide variety of tasks. XEUS sets a new SOTA on the ML-SUPERB multilingual ASR benchmark, outperforming SSL models such as MMS (Pratap et al., 2023) and w2v-BERT 2.0 v2 (Barrault et al., 2023b) while having fewer parameters or less training data. Our speech translation (ST) results show the effectiveness of XEUS' wide language coverage, even for languages with less than 10 hours of data in the pre-training corpus. We also explore XEUS' potential in generative tasks and show its superiority on speech resynthesis when compared to other SSL encoders. Finally, we evaluate XEUS' representations on a variety of tasks through the English-only SUPERB benchmark, where it sets a new SOTA on 4 tasks despite XEUS' focus on multilingual performance.

To conduct SSL pre-training at such scale, we had to make significant optimizations to existing speech processing toolkits. To encourage further SSL research and reproducibility, we will publicly release this code, along with the training configurations and checkpoints for XEUS. We also release all 200+ intermediate checkpoints and training logs obtained throughout the pre-training for further research in the training dynamics of large-scale multilingual SSL models.

To summarize, our main contributions are as follows:

1. We publicly release a new corpus that contains 7,413 hours of unlabeled speech across 4,057 languages, 25+ times wider coverage than current public datasets (Shi et al., 2023a).

2. We introduce a new self-supervised task that improves model robustness by implicitly learning to clean reverberant audio.

3. We publicly release XEUS, a SSL speech encoder trained on over 1 million hours of data across 4,057 languages.

4. We evaluate XEUS on numerous downstream tasks, and show that it outperforms SOTA SSL models such as MMS (Pratap et al., 2023), w2v-BERT 2.0 v2 (Barrault et al., 2023b), and WavLM on tasks such as ASR, ST, and speech resynthesis.

2 Motivation and Related Work

2.1 Speech Representation Learning

SSL has seen tremendous success in speech processing by having neural networks learn rich feature representations from large-scale unlabeled data (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022), which can then be fine-tuned on various downstream tasks (Chen et al., 2023c; Zhang et al., 2023). Multilingual SSL is a natural extension of this technique (Conneau et al., 2021; Babu et al., 2022) and facilitates cross-lingual transfer learning at scale. However, almost no studies leverage this capability to scale multilingual SSL models to truly massively multilingual settings, with the exception of Meta's MMS capable of covering ∼1,000+ languages (Pratap et al., 2023). However, MMS relies upon the older wav2vec 2.0 SSL objective, which has now been consistently outperformed by newer SSL objective (wen Yang et al., 2021; Hsu et al., 2021; Chiu et al., 2022). In our work, we scale to 4 times the language coverage of MMS while further boosting performance with more powerful model architectures (Kim et al., 2023) and training objectives (Chen et al., 2022).

Table 1: Comparison of multilingual SSL models by number of languages, training data size, and transparency. We define transparency in terms of the public availability of the data, model checkpoints (weights), training code and/or configurations, and the release of any other training artifacts (logs, intermediate checkpoints).

Model	Langs.	Hours	Transparency			
			Data	Weights	Training Code	Other Artifacts
XLS-R 128 (Babu et al., 2022)	128	436K	✓	✓	✗	✗
w2v-BERT 51 (Conneau et al., 2022)	51	429K	✓	✗	✗	✗
MR-HuBERT (Shi et al., 2024)	23	100K	✓	✓	✓	✗
WavLabLM (Chen et al., 2023b)	136	40K	✓	✓	✓	✗
MMS (Pratap et al., 2023)	1,406	491K	✗	✓	✗	✗
USM (Zhang et al., 2023)	300	12.5M	✗	✗	✗	✗
w2v-BERT 2.0 v1 (Barrault et al., 2023a)	143	1M	✗	✓	✗	✗
w2v-BERT 2.0 v2 (Barrault et al., 2023b)	143	4.5M	✗	✓	✗	✗
XEUS (ours)	4,057	1M	✓	✓	✓	✓

2.2 Robust Speech Representations

As most SSL speech encoders are trained solely on clean speech (Baevski et al., 2020; Hsu et al., 2021; Chung et al., 2021), they perform noticeably worse on noisy audio (Chang et al., 2021). A common approach to alleviate this issue is to perform continued pre-training of the SSL model in noisy conditions (Chang and Glass, 2023; Ng et al., 2023; Huang et al., 2023; Zhu et al., 2023). While computationally efficient, the performance of these methods is ultimately limited by the underlying SSL model. WavLM (Chen et al., 2022) solves this issue by introducing an implicit denoising task during SSL pre-training, where the model must predict clean phonetic pseudo-labels when given an utterance corrupted with acoustic noise. WavLabLM (Chen et al., 2023b) extends this approach to the multilingual setting. Unlike XEUS, however, neither model considers the impact of reverberation, and both are trained on much smaller corpora (40K-86K hours vs. 1+ million hours).

2.3 Open Foundation Models

State-of-the-art speech foundation models vary significantly in their degree of openness (Table 1). The best performing models like Whisper (Radford et al., 2023), Google USM (Zhang et al., 2023), w2v-BERT 2.0 v1 (Barrault et al., 2023a), and w2v-BERT 2.0 v2 (Barrault et al., 2023b) are all trained on fully closed data. Whisper and w2v-BERT 2.0 v1/v2 only report pre-training data quantity and the languages covered. The USM report includes much more information about their data sources, but the model checkpoints remain unreleased.

Smaller scale multilingual speech models follow more open release practices. XLSR 53 (Conneau et al., 2021) and XLS-R 128 (Babu et al., 2022) came with checkpoints and only use publicly accessible datasets but did not release training code. Similarly, MMS (Pratap et al., 2023) released checkpoints but did not release their training code and crawled data. WavLabLM (Chen et al., 2023b) and MR-HuBERT (Shi et al., 2024) released code and checkpoints but operated on a smaller scale.

Software infrastructure remains a critical barrier to democratizing speech SSL research. While there is plenty of infrastructure for large-scale training of text-based models (Workshop et al., 2022; Andonian et al., 2023; Liu et al., 2023; Groeneveld et al., 2024), no similar work has achieved speech pre-training at our scale. AR-HuBERT (Chen et al., 2023a) and the OWSM project (Peng et al., 2023b, 2024b,a) sought to reproduce SOTA speech models in an open manner but use more than 80% less data than our work. With XEUS, we release the entire pre-training framework. We only use publicly accessible datasets and release all of the additional pre-training data that we crawled. To facilitate research in the training dynamics of large-scale SSL models, we also release all model logs and are the first to also release all intermediate checkpoints. As we are the first to create an open SSL speech model at such data and model scale, we also release all of our heavily optimized training code.

3 Data

3.1 Existing Datasets

We begin by combining a large variety of pre-training data from 37 publicly accessible[1] speech processing datasets across 150+ languages, which

[1]We follow Peng et al. (2023b)and include licensed data such as BABEL (IARPA) as part of this definition, as exact copies can be obtained, unlike that of closed/unreleased data.

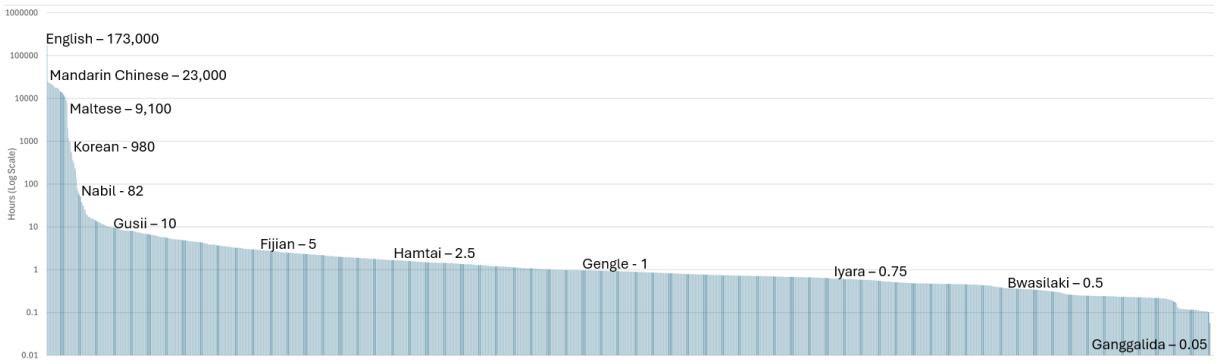

Figure 1: Distribution of XEUS pre-training data by language (log scale). We exclude data from YODAS (Li et al., 2023) due to the noisiness of the language labels.

totals to 1.074 million hours of data. Details about these datasets can be found in Section A.1 in the Appendix, while an overview is presented in Table 2. Notable data used in this work that was unseen in prior SSL models includes accented speech (Ahamad et al., 2020; Sanabria et al., 2023), code-switching (Lyu et al., 2010), and singing voices (Smule, 2019). To prevent data corruption, we use only the official training splits of each dataset. To increase language coverage beyond these 150 languages, in the next subsections we describe our collection of three additional data sources.

3.2 MMS-unlab v2

We first reproduce the MMS-unlab dataset (Pratap et al., 2023), which was not publicly released, and scale it to 200 more languages. Like the original, we crawl religious audiobooks from the Global Recordings Network.[2] Our data, however, is processed in a different manner. Since we use it for SSL instead of language identification, we do not filter out languages with low amounts of data. We also perform VAD with an energy-based detector[3] instead of a neural model, the latter of which is more computationally expensive and likely less robust to unseen languages. This leads to a total of 6,700 hours of data across 4,023 ISO3 languages, which is wider in coverage than the original with 3,809 languages. We obtain explicit written permission for the use and redistribution of this data, as the Global Recordings Network website did not include clear licensing information.

3.3 WikiTongues

We also create new unlabeled speech corpora by crawling data from 2 data sources previously unseen in the speech processing literature. The first is WikiTongues,[4] based on a grassroots project that collected recordings of many languages and dialects spoken around the world with the goal of language preservation. Each 2-20 minute recording is released under the CC-BY-NC license, and contains 1-2 speakers casually speaking a particular language/dialect. Importantly, many of these languages are low-resource, if not endangered or extinct. In total, the dataset we obtain contains around 700 languages and/or dialects. We obtained 821 recordings, yielding about 70 hours of speech data. We use the same VAD settings as in Section 3.2 to segment each recording.

3.4 Jesus Dramas

The second corpus we collect is *Jesus Dramas*. The source of this data is Inspirational Films[5], which released the “Story of Jesus" audio drama in 430 languages under a non-commercial license. Each multi-speaker audio drama is 90 minutes long, totalling 645 hours. We use the same VAD settings as in Section 3.2 to segment these dramas into utterance-level clips.

3.5 Final Pre-Training Corpus

The new datasets we collect from Sections 3.2, 3.3 and 3.4 total 7,413 hours of data across 4,057 ISO3 languages. After aggregating it with the data from Section 3.1, we obtain a total of 1.081 million hours of pre-training data. We filter out all utterances longer than 40 seconds due to memory constraints. An overview of all of our pre-training corpora with their licensing information is presented in Table 11 in the Appendix. Figure 1 shows an overview of the language distribution in our data on a log-scale. Overall, our pre-training dataset spans 189

[2]https://globalrecordings.net/en/us
[3]https://github.com/wiseman/py-webrtcvad
[4]https://wikitongues.org
[5]https://www.inspirationalfilms.com

Table 2: Overview of datasets used for pre-training. The language column indicates the language used in monolingual datasets and the number of languages in multilingual datasets. **Bolded** dataset names indicate new corpora we will release.

Dataset	Language(s)	Domain	Hours
YODAS (Li et al., 2023)	140	Variety	422K
VoxPopuli (Wang et al., 2021)	23	Legal	400K
LibriLight (Kahn et al., 2020)	English	Audiobook	60K
MLS (Pratap et al.)	8	Audiobook	44K
People's Speech (Galvez et al., 2021)	English	Variety	30K
WeNetSpeech (Zhang et al., 2022)	Mandarin	Variety	22K
Russian Open STT (Slizhikova et al., 2020)	Russian	Variety	20K
NPTEL2020 (AI4Bharat, 2020)	English	Talk	15K
Reazonspeech (Yin et al., 2023)	Japanese	Television	15K
Common Voice 13 (Ardila et al., 2020)	92	Read	13K
GigaSpeech (Chen et al., 2021)	English	Variety	10K
VoxLingua (Valk and Alumäe, 2021)	107	Variety	6800
MMS-unlab v2	4023	Religious	6700
SPGI (O'Neill et al., 2021)	English	Finance	5000
Fisher (Post et al., 2013)	English	Conversation	2000
Googlei18n (Chen et al., 2023b)	34	Variety	1328
BABEL (IARPA)	17	Conversation	1000
FLEURS (Conneau et al., 2022)	102	News	1000
KSponSpeech (Bang et al., 2020)	Korean	Conversation	970
LibriSpeech (Panayotov et al., 2015)	English	Audiobook	960
MagicData (Yang et al., 2022)	Mandarin	Conversation	755
mTEDx (Salesky et al., 2021)	8	Talk	753
Jesus Dramas	430	Religious	643
Althingi (Helgadóttir et al., 2019)	Icelandic	Legal	542
TEDLIUM3 (Hernandez et al., 2018)	English	Talk	500
VoxForge (VoxForge)	8	Read	235
AISHELL (Bu et al., 2017)	Mandarin	Read	200
SEAME (Lyu et al., 2010)	Codeswitch	Conversation	192
DAMP-MVP (Smule, 2019)	English	Singing	150
NorwegianParl. (Solberg and Ortiz, 2022)	Norwegian	Legal	140
AIDATATANG (aid)	Mandarin	Read	140
AMI (Carletta, 2007)	English	Meetings	100
Nahuatl (Shi et al., 2021a)	Nahuatl	Conversation	82
WSJ (Paul and Baker, 1992)	English	Read	81
Mixtec (Shi et al., 2021b)	Mixtec	Conversation	70
WikiTongues	700	Conversation	70
Siminchik (Cardenas et al., 2018)	Quechua	Radio	50
Edinburgh Accent (Sanabria et al., 2023)	English	Conversation	40
VCTK (Yamagishi et al., 2019)	English	Read	25
AccentDB (Ahamad et al., 2020)	English	Read	20
Totonac (Berrebbi et al., 2022)	Totonac	Monologue	17

language families (Appendix Figure 3). We find that the languages follow a long tail distribution, with the top 50 languages accounting for 99.5% of the data. However, a very encouraging finding is that around 2,000 languages have 1 or more hours of data each. Data from YODAS (Appendix Section A.1) is excluded from the above analyses, as we found the language labels to be very noisy.

4 Self-Supervised Pre-Training

XEUS' training, sketched in Figure 2, combines ideas from HuBERT's masked prediction, WavLM's denoising objective, and a new dereverberation objective. These components are described in the following sub-sections.

4.1 Masked Prediction and Denoising

To obtain the target phonetic pseudo-labels for HuBERT masked prediction, we first extract encoded representations from a pre-trained WavLabLM MS model (Chen et al., 2023b). The representations are then clustered using k-means, with $k = 2,048$.

Algorithm 1 Simulation of Reverberant Speech

Require: a batch of utterances B, a set of RIRs D, and reverberation probability p_r.
for $u \in B$ **do**
  Sample v from cont. dist. $\mathcal{U}(0, 1)$
  if $v < p_r$ **then**
    Sample a random RIR u_n from D
    $dt = \min(\text{argmax}(u_n))$
    $r = u \circledast u_n$
    Realign r to u using dt
    Rescale r to have the same energy as u
  end if
end for
return B

The data used for the feature extraction and clustering is a subset of our training data. Specifically, we sample 6,000 hours from a combination of Common Voice, MLS, and Googlei18n, 6,000 hours from YODAS, and 6,000 hours from MMS-unlab v2, and combine it with the entirety of FLEURS and BABEL's training data. This leads to a total of 20K hours used for k-means.

We also integrate the acoustic denoising task proposed by WavLM into XEUS pre-training. During training, an input utterance has some probability p to be augmented with either random noise from the Deep Noise Suppression Challenge (Reddy et al., 2021) or another utterance in the batch as interference. As the target labels are obtained solely from uncorrupted speech, the model learns to implicitly clean the input audio.

4.2 Dereverberation

We extend the concept of acoustic denoising introduced by WavLM by proposing another speech enhancement task for SSL pre-training: dereverberation. Similarly to the WavLM dynamic mixing augmentation, we simulate reverberant conditions in the input audio during training while the target pseudo-labels are again left untouched. The model must thus implicitly learn to remove the reverberation from the audio to predict the clean pseudo-labels. We note that it is possible for *both* the noise and reverberation augmentation to be applied for a single utterance. For simplicity, the noise augmentation is always applied first.

Our technique (Algorithm 1) consists of the following steps. First, each utterance in a mini-batch has a probability p_r for the reverberation augmentation to be applied. If an utterance u is to be augmented, we then randomly sample a Room Impulse Response (RIR) u_n to be used from the audio of Ko

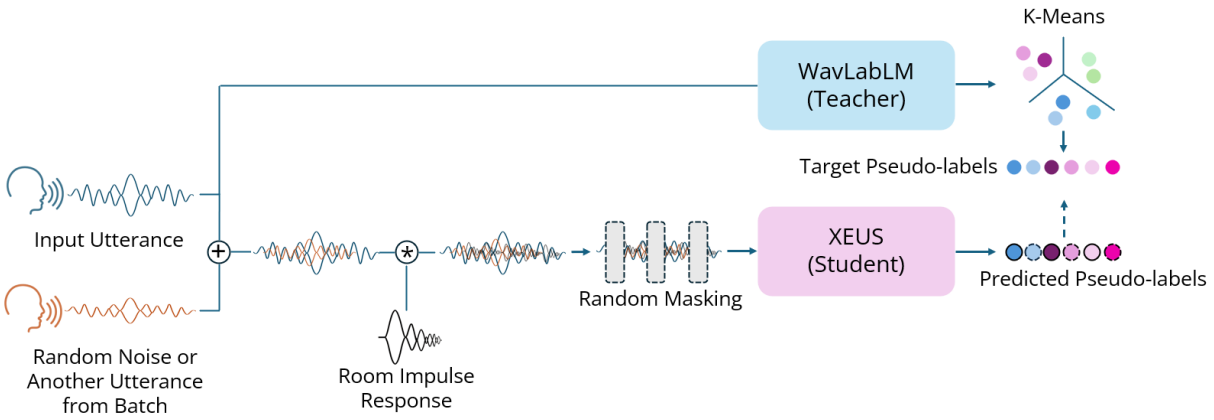

Figure 2: Overview of XEUS' pre-training. The teacher encoder generates phonetic pseudo-labels from clean speech, while the student must predict those pseudo-labels after masking, random noise and/or reverberation is applied to the input waveform.

et al. (2017). We first estimate dt, the sample shift imposed on u after adding the reverberation, according to the highest peak in u_n. The reverberant utterance r can then be simulated via convolution ($\circledast$) between u and u_n. Finally, we realign r with u using dt and normalize it to have the same energy as u. This final realignment step is crucial for the effectiveness of this technique; otherwise the audio would be shifted and misaligned with the target pseudo-labels. Quantitative analyses of our method can be found in Appendix Section A.3.

4.3 Model Architecture

XEUS is based on the HuBERT architecture (Hsu et al., 2021), with several modifications. We replace the Transformer (Vaswani et al., 2017) with an E-Branchformer (Peng et al., 2022; Kim et al., 2023), as convolution-augmented models achieve superior SSL performance (Chung et al., 2021). We choose the E-Branchformer over the Conformer (Gulati et al., 2020) due to the former's relative ease of training and superior downstream performance (Peng et al., 2023a). We also replace the original HuBERT loss with cross entropy, which is faster and leads to better downstream performance (Yang et al., 2023). XEUS consists of a convolutional feature extractor and 19 E-Branchformer layers. Each of the latter has 8 attention heads, a hidden dimension of 1,024, feed-forward size of 4,096, and kernel size of 31. The model size is 577M parameters. Ablations on these modifications can be found in Appendix Section A.2.

4.4 Pre-Training Settings

XEUS is pre-trained on 64 40GB NVIDIA A100 GPUs using the ESPnet toolkit (Watanabe et al., 2018). Each GPU has a maximum batch size of 100 seconds, leading to a total batch size of 106 minutes. We use a noise augmentation probability p of 0.2, where there is an equal probability of the corruption being random noise or another utterance (Section 4.1). We use a dereverberation augmentation probability p_r of 0.3 (Section 4.2). We perform a two passes through the training set, totalling 670K steps. More details and a breakdown of the training costs can be found in Appendix Section A.4.

The scale of XEUS' pre-training is unseen outside of a few other works (Radford et al., 2023; Zhang et al., 2023; Barrault et al., 2023b), with the amount of pre-training data being 5-25 times the size of those used in prior models trained with public toolkits (Peng et al., 2023b; Chen et al., 2023a; Hsu et al., 2021). To conduct training on such a scale, we made several optimizations to the open-source ESPnet toolkit, which was originally designed for standard academic-scale experiments. These optimizations will all be made publicly available. More details about these improvements are also reported in Appendix Section A.5.

5 Downstream Evaluation

We examine the capabilities of XEUS in various downstream applications. Section 5.1 evaluates the multilingual capabilities of XEUS in different settings. Section 5.2 evaluates the universality of XEUS' representations for a broad range of speech information, such as emotion and speaker content. Finally, Section 5.3 tests the acoustic representations of XEUS via speech resynthesis. We provide an overview of each downstream setup in its respective subsection. Detailed experimental settings can be found in Appendix Section A.6.

Table 3: Evaluation results on the 10 minute / 1 hour settings of the ML-SUPERB Benchmark in ASR CER (↓) and LID ACC (↑). **Bold** numbers indicate the best model for a task, while underlined numbers indicate second best.

				MONO. ASR	MULTI. ASR		LID	MULTI. ASR + LID		
					Normal	Few-shot	Normal	Normal		Few-shot
Model	Params.	Hours	$SUPERB_s$	CER	CER	CER	ACC	ACC	CER	CER
WavLabLM	316M	40K	700 / 767	39.9 / 32.1	37.8 / 31.3	43.8 / 40.9	71.7 / 81.1	70.8 / 82.2	37.0 / 30.6	43.4 / 40.2
XLS-R 128	316M	436K	707 / 851	39.7 / 30.6	29.3 / 22.0	40.9 / 39.3	66.9 / **87.9**	55.6 / 85.6	28.4 / 22.9	42.1 / 42.4
XLS-R 128	1B	436K	745 / 838	40.5 / 30.9	30.4 / 26.3	39.1 / 38.5	70.9 / 85.8	66.4 / 87.1	28.6 / 25.2	39.2 / 39.5
MMS	316M	491K	795 / 845	33.8 / 30.5	28.7 / 24.0	36.5 / 36.5	62.3 / 84.3	71.9 / 74.3	31.5 / 30.0	30.9 / 29.2
MMS	1B	491K	953 / 948	33.3 / 25.7	21.3 / **18.1**	**30.2 / 30.8**	**84.8** / 86.1	73.3 / 74.8	26.0 / 25.5	**25.4 / 24.8**
w2v-BERT 2.0 v2	580M	4.5M	826 / 916	41.0 / 29.2	24.6 / 20.3	34.3 / 35.3	70.0 / 86.8	83.2 / 90.6	24.2 / 20.3	34.3 / 34.0
XEUS (ours)	577M	1M	**956 / 956**	**30.3 / 25.1**	**21.1** / 20.1	33.4 / 34.1	81.5 / 87.3	**86.4 / 91.3**	**22.9 / 19.6**	32.7 / 32.8

Table 4: FLEURS ASR+LID & JesusFilm ST results.

Model	FLEURS		JesusFilm
	CER ↓	ACC ↑	chrF ↑
XLS-R 1B	9.6	92.5	13.4
MMS 1B	9.2	94.0	15.5
w2v-BERT 2.0 v2	**8.7**	**94.3**	15.1
XEUS (ours)	8.9	93.0	**22.1**

5.1 Multilingual Speech Processing

We primarily compare XEUS with 3 SOTA multilingual SSL models: XLS-R 128 (Babu et al., 2022), MMS (Pratap et al., 2023), and w2v-BERT 2.0 v2 (Barrault et al., 2023b) (Table 1). We use the ML-SUPERB benchmark (Shi et al., 2023a) as our main evaluation, as it tests each models across a diverse range of tasks and languages. We complement these experiments with additional analyses on FLEURS ASR+LID and low-resource ST.

5.1.1 ML-SUPERB

ML-SUPERB benchmarks self-supervised speech representations on a variety of multilingual tasks across 143 languages. ASR performance is evaluated in terms of character error rate (CER ↓), while accuracy (ACC ↑) is used to evaluate LID. The benchmark is split across two data settings for each task: 10 minutes (min.) and 1 hour of data for each language. Each data setting contains 4 tasks: monolingual ASR in 9 languages, multilingual ASR, LID, and joint multilingual ASR+LID. In the multilingual tasks, 5 languages are reserved as a few-shot task, while the other 138 languages have the standard 10 min. / 1 hour of fine-tuning data. An overall $SUPERB_s$ (↑) score for each model in each data setting is calculated following the benchmark rules (Shi et al., 2023a). Further details can be found in Appendix Section A.6.1.

Table 3 shows that XEUS is the overall best performing model, with the highest $SUPERB_s$ score of 956 on both the 10 min. / 1 hour settings. XEUS achieves SOTA results on monolingual ASR with the best scores of 25.1 and 33.3 CER on the 1 hour and 10 min. tracks respectively. On multilingual ASR, XEUS is only outperformed by MMS 1B. For ASR+LID, XEUS achieves the best performance in the normal setting for both data tracks. While XEUS is worse than MMS in few-shot CER, it still achieves reasonable results and outperforms the other SSL models. Overall, XEUS outperforms the parameter-equivalent w2v-BERT 2.0 v2 across all task categories. This is accomplished using only accessible training data, which is 22% the size than that of w2v-BERT 2.0 v2.

5.1.2 FLEURS

FLEURS is a 102-language multilingual ASR benchmark, where each language has around 6-10 hours of training data. In this setting, we use heavier downstream probes that reflect SOTA ASR architectures. We adopt the same setup as (Peng et al., 2023a; Chen et al., 2023c), which remains the SOTA on FLEURS when not using additional labeled data. The downstream model consists of an E-Branchformer encoder paired with a Transformer decoder, totalling 100M parameters. Exact settings are shown in Appendix Section A.6.2.

The results of the FLEURS experiments are shown in the middle section of Table 4. We find that XEUS remains competitive with the SOTA w2v-BERT 2.0 v2 trained on much more data (8.9 vs 8.7 CER), and significantly outperforms both XLS-R and MMS 1B (9.6 and 9.2 CER respectively).

5.1.3 Low-Resource Language Coverage

While FLEURS and ML-SUPERB provide comprehensive multilingual benchmarks, their language coverage is far smaller than that of XEUS (102-143 vs 4,057). To understand if XEUS' wide language coverage was effective for languages with small (< 10 hours) amounts of pre-training data, we crawled additional labeled data for evaluation. We randomly chose 3 languages from the Jesus Film

Table 5: Evaluation on the SUPERB Benchmark. (🥇), (🥈), and (🥉) indicate first, second and third best results respectively on the online leaderboard: https://superbbenchmark.org/leaderboard.

Method	Recognition		Detection		Semantics			Speaker			Paralinguistics
	PR↓	ASR↓	KS ↑	QbE ↑	IC ↑	SF (F1) ↑	SF (CER) ↓	SID ↑	ASV ↓	SD ↓	ER ↑
WavLM Large	**3.06** 🥇	3.44 🥉	97.86 🥈	**8.86**	**99.31** 🥇	**92.21** 🥇	**18.36** 🥇	**95.49** 🥇	**3.77** 🥇	3.24 🥈	70.62 🥈
XEUS (ours)	3.21 🥉	**3.34** 🥇	**98.32** 🥇	7.49	98.70	90.05	21.49 🥉	91.70 🥈	4.16 🥉	**3.11** 🥇	**71.08** 🥇

Table 6: Speech resynthesis results on VCTK.

Model	MOS (↑)	WER (↓)	F0 (↓)	MCD (↓)
WavLM Large	3.20	27.8	0.26	4.55
w2v-BERT 2.0 v2	3.21	15.5	0.27	3.92
XEUS (ours)	**3.23**	**10.0**	**0.25**	**3.80**

Project[6] that were not covered in ML-SUPERB and/or FLEURS: Hijazi Arabic, Lumun, and Rajbanshi. This yields around 1.5 hours of speech for each language. Of all existing SSL models, only XEUS covers the former two, while Rajbanshi is covered by both XEUS and MMS. We then train X → English Speech Translation (ST) models for each language (as ASR transcripts were not available) and evaluate using character-level F-score (chrF). More details about the evaluation data and ST settings are shown in Appendix Section A.6.3.

The average chrF scores across all languages are shown in the right of Table 4. XEUS significantly outperforms all other models, likely due to its wider language coverage. The next best model is MMS 1B, which obtains an average chrF of 15.5, while XEUS scores 22.1, a relative improvement of 35%.

5.2 Task Universality

To test how well XEUS encodes other forms of speech content, we benchmark its capabilities on the English-only SUPERB (wen Yang et al., 2021). SUPERB tests self-supervised speech models across 5 broad task categories: Recognition (ASR and Phoneme Recognition (PR)), Detection (Keyword Spotting (KS) and Query By Example (QbE)), Semantics (Intent Classification (IC) and Slot Filling (SF)), Speaker (Speaker Identification (SID), Automatic Speaker Verification (ASV), and Speaker Diarization (SD)), and Paralinguistics (Emotion Recognition (ER)). Metrics for each task are: phoneme error rate (PR), WER (ASR), maximum term weighted value (QbE), F1 and concept error rate (SF), equal error rate (ASV), diarization error rate (SD), and accuracy (KS, IC, SID, and ER). Exact experimental settings are shown in Appendix Section A.6.5. We compare XEUS to WavLM (Chen et al., 2022), the SOTA model on the SUPERB leaderboard for almost all tasks. Our results in Table 5 show that XEUS consistently reaches if not surpasses SOTA scores across a variety of tasks, obtaining the highest score in 4 English-only tasks (KS, SD, ER, ASR), despite its curse of multilinguality (Conneau et al., 2020).

5.3 Acoustic Representation Evaluation

We evaluate XEUS on its acoustic representation quality through the task of speech resynthesis. Here, we compare primarily against w2v-BERT 2.0 v2 as the SOTA multilingual model and WavLM Large as the SOTA English-only model. We train unit-to-speech HiFiGAN vocoders (Kong et al., 2020; Polyak et al., 2021) on the accented-English VCTK (Yamagishi et al., 2019) dataset with a discrete codebook vocabulary size of 100. We evaluate the quality of the resynthesized speech in terms of Mel-Cepstral Distortion (MCD), log-F0 Root Mean Square Error (F0), predicted Mean Opinion Score (Lo et al., 2019) (MOSNet), and Word Error Rate (WER). We obtain WER by transcribing the synthesized speech with a pre-trained Whisper medium (Radford et al., 2023). For each SSL model, we experiment with features extracted at 50% and 75% of the model depth (ie. layer 18 out of 24 in the latter case), and report the best performing configuration in Table 6. More details about this search can be found in Appendix Section A.6. Our results show that resynthesized speech from XEUS is higher quality than that from both WavLM and w2v-BERT 2.0 v2 across all metrics, whether it be perceptual or semantic, showcasing its strong performance in generative tasks along with its SOTA recognition capabilities.

6 Conclusion

This work presents XEUS, an SSL speech encoder trained on over 1 million hours of data across 4,057 languages. As a community contribution, we release a new dataset with 7,413 hours of unlabeled speech data across those 4,057 languages. We also introduce a novel joint dereverberation task for SSL pre-training, which we use to increase the robustness of XEUS. We show that XEUS can achieve

[6]https://www.jesusfilm.org

comparable performance if not outperform other SOTA SSL models on various benchmarks, while having much stronger performance on long-tail languages. To make XEUS reproducible, we will release all training code and configurations, along with model weights. In the future, we hope to extend the downstream use of XEUS to a larger scale.

7 Limitations:

While the overall pre-training corpus of XEUS contains over 4,000 languages, many of these languages have less than 1 hour of speech data. While we are able to show that the presence of this small amount of data is still beneficial for these languages in downstream tasks (Section 5.1.3), the performance is still likely much worse than the performance on higher-resourced languages. Furthermore, due to the efforts required to collect and manually clean evaluation data, we only test on a subset of these truly low-resource languages in Section 5.1.3. While XEUS is one step towards speech recognition or translation systems for these tail languages, much work is still required before these tools can be deployed to end users.

Due to the large number of tasks and domains that our evaluation covers, we mostly focus on relatively lightweight benchmarks such as the SUPERB suite and perform limited hyperparameter tuning. While this allows for fair comparisons between different models, the evaluation does not use the large-scale fine-tuning common in SOTA settings for downstream tasks.

8 Broader Impact and Ethics:

Broader Impact: In this work, we introduce XEUS, a new large-scale multilingual SSL speech encoder. Unlike previous foundation models that focus on a single aspect, XEUS obtains SOTA performance across a diverse range of *both* tasks *and* languages, further pushing towards the goal of truly universal models. By releasing both our data and model checkpoints, our goal is to provide foundations for more multilingual research, particularly in domains such as robust ASR and speech enhancement where evaluation is typically done solely on English.

Another major goal of our work is to democratize the development of speech foundation models. We believe that training infrastructure remains a significant barrier to entry for SSL research. This has two main aspects: *software infrastructure* and *training configurations*. Current speech toolkits such as ESPnet (Watanabe et al., 2018) and SpeechBrain (Ravanelli et al., 2021) focus on academic scale experiments, while general frameworks such as HuggingFace and Fairseq (Wang et al., 2020) are more limited in their implementation of different tasks and SOTA methods. By integrating our changes into ESPnet, our optimizations can allow users to scale speech models for other tasks such speaker representation learning. In the latter aspect, we note that the availability of training recipes and configurations pose the other major barrier to entry. Due to the computational cost of training, the development of foundation models poses a risk that is too high for most academic labs, as a single failed training run can be disastrous for the lab's budget. However, this can be mitigated by publishing training recipes and hyperparameters known to work well. The benefits of this is most visible in the OWSM project (Peng et al., 2023b, 2024b,a), where each successive work reported lower and lower computational expenses.

Ethics: We recognize the delicate nature of speech data, particularly when it involves the languages of indigenous and marginalized communities. Many authors of this work have long-term collaborations with indigenous communities and researchers. We are careful to crawl and release data only from sources that contain permissive licenses of the source data to avoid potential cases of misuse and violations of language ownership. For data sources that do not clearly indicate their usage/distribution terms, we obtained explicit permission from the corresponding stakeholders (such as in the case of the Global Recordings Network in Section 3.2). To follow the data's access conditions, we release all of our data under non-commercial licenses.

We partially anonymize our crawled datasets by removing speaker names from the meta-data. However, we do not alter the content of the speech itself. As such, we urge users of our released data to respect the privacy of the speakers and not attempt to identify them. It is also possible that portions of the speech content may be offensive in particular settings. With the diversity of over 4000 languages, it is likely that there are statements or views that are normative in one culture but offensive in another.

While encoder-only speech models like XEUS have limited uses without any task-specific fine-tuning, the downstream models that are created after such processes are prone to the biases and misuse cases that all machine learning models are

vulnerable to. For example, XEUS' capabilities in speech generation can be used for misinformation via audio deepfakes, which is an unintended use case of this model.

Acknowledgement

This work used the Bridges2 at PSC and Delta NCSA computing systems through allocation CIS210014 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, supported by National Science Foundation grants 2138259,2138286, 2138307, 2137603, and 2138296.

References

aidatatang_200zh, a free Chinese Mandarin speech corpus by Beijing DataTang Technology Co., Ltd.

Afroz Ahamad, Ankit Anand, and Pranesh Bhargava. 2020. AccentDB: A database of non-native English accents to assist neural speech recognition. In *LREC 2020*.

AI4Bharat. 2020. NPTEL2020: Indian English Speech Dataset.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, Phil Wang, and Samuel Weinbach. 2023. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *LREC 2020*, pages 4218–4222.

Peter K. Austin and Julia Sallabank, editors. 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge University Press.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML 2022*, pages 1298–1312.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS 2020*, volume 33.

Jeong-Uk Bang et al. 2020. KsponSpeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023a. SeamlessM4T-massively multilingual & multimodal machine translation. *arxiv:2308.11596*.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023b. Seamless: Multilingual expressive and streaming speech translation. *arxiv:2312.05187*.

Dan Berrebbi, Jiatong Shi, Brian Yan, Osbel López-Francisco, Jonathan Amith, and Shinji Watanabe. 2022. Combining spectral and self-supervised features for low resource speech recognition and translation. In *Interspeech 2022*.

Alan W Black. 2019. CMU Wilderness Multilingual Speech Dataset. In *ICASSP 2019*.

Hui Bu et al. 2017. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *O-COCOSDA*.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern Quechua. *ISI-NLP*.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. Springer.

Heng-Jui Chang and James Glass. 2023. R-spin: Efficient speaker and noise-invariant representation learning with acoustic pieces. *arXiv preprint arXiv:2311.09117*.

Xuankai Chang, Takashi Maekaku, Pengcheng Guo, Jing Shi, Yen-Ju Lu, Aswin Shanmugam Subramanian, Tianzi Wang, Shu-wen Yang, Yu Tsao, Hung-yi Lee, and Shinji Watanabe. 2021. An exploration of self-supervised pretrained representations for end-to-end speech recognition. In *ASRU 2021*, pages 228–235.

Guoguo Chen et al. 2021. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu,

Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE JSTSP*.

William Chen, Xuankai Chang, Yifan Peng, Zhaoheng Ni, Soumi Maiti, and Shinji Watanabe. 2023a. Reducing Barriers to Self-Supervised Learning: HuBERT Pre-training with Academic Compute. In *Interspeech 2023*.

William Chen, Jiatong Shi, Brian Yan, Dan Berrebbi, Wangyou Zhang, Yifan Peng, Xuankai Chang, Soumi Maiti, and Shinji Watanabe. 2023b. Joint prediction and denoising for large-scale multilingual self-supervised learning. In *ASRU 2023*.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. 2023c. Improving massively multilingual ASR with auxiliary CTC objectives. In *ICASSP 2023*.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. 2022. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*. PMLR.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *ASRU 2021*.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL 2020*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau et al. 2022. FLEURS: Few-shot learning evaluation of universal representations of speech. In *SLT 2022*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *NeurIPS 2022*.

Ethnologue. 2017. How many languages in the world are unwritten?

Daniel Galvez, Greg Diamos, Juan Manuel Ciro Torres, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The People's Speech: A large-scale diverse English speech recognition dataset for commercial usage. In *NeurIPS 2021*.

J.J. Godfrey et al. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *ICASSP 1992*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML 2006*, pages 369–376.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*.

Inga R. Helgadóttir, Anna Björk Nikulásdóttir, Michal Borský, Judy Y. Fong, Róbert Kjaran, and Jón Guðnason. 2019. The Althingi ASR system. In *Interspeech 2019*.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018*. Springer.

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 TALSP*.

Kuan-Po Huang, Yu-Kuan Fu, Tsu-Yuan Hsu, Fabian Ritter Gutierrez, Fan-Lin Wang, Liang-Hsuan Tseng, Yu Zhang, and Hung-yi Lee. 2023. Improving generalizability of distilled self-supervised speech processing models under distorted settings. In *SLT 2022*, pages 1112–1119.

IARPA. The Babel Program.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-Light: A benchmark for ASR with limited or no supervision. In *ICASSP 2020*.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J Han, and Shinji Watanabe. 2023. E-branchformer: Branchformer with enhanced merging for speech recognition. In *SLT 2023*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR 2015*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *ICASSP 2017*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *NeurIPS 2020*, 33:17022–17033.

Xinjian Li, Florian Metze, David R. Mortensen, Alan W Black, and Shinji Watanabe. 2022. ASR2K: Speech Recognition for Around 2000 Languages without Audio. In *Interspeech 2022*.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. YODAS: Youtube-oriented dataset for audio and speech. In *ASRU 2023*.

Alexander H Liu, Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and Jim Glass. 2024. DinoSR: Self-distillation and online clustering for self-supervised speech representation learning. *NeurIPS 2024*, 36.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. 2023. LLM360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion. In *Interspeech 2019*, pages 1541–1545.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2010. SEAME: a Mandarin-English code-switching speech corpus in south-east Asia. In *Interspeech 2010*.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. VoxtLM: Unified Decoder-Only Models for Consolidating Speech Recognition, Synthesis and Speech, Text Continuation Tasks. In *ICASSP 2024*, pages 13326–13330. IEEE.

Dianwen Ng, Ruixi Zhang, Jia Qi Yip, Zhao Yang, Jinjie Ni, Chong Zhang, Yukun Ma, Chongjia Ni, Eng Siong Chng, and Bin Ma. 2023. De'hubert: Disentangling noise in a self-supervised model for robust speech recognition. In *ICASSP 2023*, pages 1–5.

Jumon Nozaki and Tatsuya Komatsu. 2021. Relaxing the conditional independence assumption of CTC-based ASR by conditioning on intermediate predictions. In *Interspeech 2021*, pages 3735–3739.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. 2021. SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. In *Interspeech 2021*.

Vassil Panayotov et al. 2015. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP 2015*.

Douglas B Paul and Janet Baker. 1992. The design for the Wall Street Journal-based CSR corpus. In *Workshop on Speech and Natural Language*.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding. In *ICML 2022*.

Yifan Peng, Kwangyoun Kim, Felix Wu, Brian Yan, Siddhant Arora, William Chen, Jiyang Tang, Suwon Shon, Prashant Sridhar, and Shinji Watanabe. 2023a. A comparative study on e-branchformer vs conformer in speech recognition, translation, and understanding tasks. In *Interspeech 2023*.

Yifan Peng, Yui Sudo, Muhammad Shakeel, and Shinji Watanabe. 2024a. OWSM-CTC: An Open Encoder-Only Speech Foundation Model for Speech Recognition, Translation, and Language Identification. *arXiv preprint arXiv:2402.12654*.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, et al. 2024b. OWSM v3. 1: Better and Faster Open Whisper-Style Speech Models based on E-Branchformer. *arXiv preprint arXiv:2401.16658*.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee weon Jung, Soumi Maiti, and Shinji Watanabe. 2023b. Reproducing Whisper-style training using an open-source toolkit and publicly available data. In *ASRU 2023*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Interspeech 2021*.

Matt Post et al. 2013. Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus. In *IWSLT 2013*.

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

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML 2023*.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Chandan K.A. Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. 2021. INTERSPEECH 2021 Deep Noise Suppression Challenge. In *Interspeech 2021*.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. Multilingual TEDx corpus for speech recognition and translation. In *Interspeech 2021*.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The Edinburgh international accents of English corpus: Towards the democratization of English ASR. In *ICASSP 2023*.

Jiatong Shi, Jonathan D Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021a. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *AmericasNLP 2021*.

Jiatong Shi, Jonathan D Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021b. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yolóxochitl Mixtec. In *EACL 2021*.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023a. ML-SUPERB: Multilingual Speech Universal PERformance Benchmark. In *Interspeech 2023*.

Jiatong Shi, William Chen, Dan Berrebbi, Hsiu-Hsuan Wang, Wei-Ping Huang, En-Pei Hu, Ho-Lam Chuang, Xuankai Chang, Yuxun Tang, Shang-Wen Li, Abdelrahman Mohamed, Hung-Yi Lee, and Shinji Watanabe. 2023b. Findings of the 2023 ML-SUPERB challenge: Pre-training and evaluation over more languages and beyond. In *ASRu 2023*.

Jiatong Shi, Hirofumi Inaguma, Xutai Ma, Ilia Kulikov, and Anna Sun. 2024. Multi-resolution HuBERT: Multi-resolution speech self-supervised learning with masked unit prediction. In *ICLR 2024*.

Anna Slizhikova et al. 2020. Russian Open Speech To Text (STT/ASR) Dataset.

Inc Smule. 2019. DAMP-MVP: Digital Archive of Mobile Performances - Smule Multilingual Vocal Performance 300x30x2.

Per Erik Solberg and Pablo Ortiz. 2022. The Norwegian parliamentary speech corpus. In *LREC 2022*, Marseille, France.

Jörgen Valk and Tanel Alumäe. 2021. VOXLINGUA107: A dataset for spoken language recognition. In *SLT 2021*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS 2017*.

VoxForge. VoxForge.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. *arxiv:2010.05171*.

Changhan Wang et al. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *ACL 2021*.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Interspeech 2018*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech 2021*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Junichi Yamagishi et al. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit.

Guanrou Yang, Ziyang Ma, Zhisheng Zheng, Yakun Song, Zhikang Niu, and Xie Chen. 2023. Fast-HuBERT: an efficient training framework for self-supervised speech representation learning. In *ASRU 2023*.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, Lei Xie, and Yonghong Yan. 2022. Open source MagicData-RAMC: A rich annotated mandarin conversational (RAMC) speech dataset. In *Interspeech 2022*, pages 1736–1740.

Yue Yin, Daijiro Mori, et al. 2023. ReazonSpeech: A Free and Massive Corpus for Japanese ASR.

Binbin Zhang et al. 2022. WenetSpeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022*.

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

Qiu-Shi Zhu, Long Zhou, Jie Zhang, Shu-Jie Liu, Yu-Chen Hu, and Li-Rong Dai. 2023. Robust data2vec: Noise-robust speech representation learning for asr by combining regression and improved contrastive learning. In *ICASSP 2023*, pages 1–5.

A Appendix

A.1 Data

For brevity, we provide an overview of these datasets in Table 11 and refer readers to the original papers for more details (Bu et al., 2017; Chen et al., 2021; Panayotov et al., 2015; O'Neill et al., 2021; Hernandez et al., 2018; Pratap et al.; Zhang et al., 2022; Carletta, 2007; Ardila et al., 2020; Godfrey et al., 1992; Conneau et al., 2022; Bang et al., 2020; Yang et al., 2022; Wang et al., 2021; Chen et al., 2023b; Li et al., 2023; Kahn et al., 2020; Galvez et al., 2021; Valk and Alumäe, 2021; Helgadóttir et al., 2019; Lyu et al., 2010; Solberg and Ortiz, 2022; Shi et al., 2021a,b; Cardenas et al., 2018; Sanabria et al., 2023; Ahamad et al., 2020; Berrebbi et al., 2022).

Over 80% of the pre-training data is derived from two corpora: YODAS (Li et al., 2023) and VoxPopuli (Wang et al., 2021). However, both of these sources largely consist of European languages. To add more linguistic diversity, we also smaller scale multilingual corpora such as include BABEL (IARPA), Googlei18n (Chen et al., 2023b), VoxLingua (Valk and Alumäe, 2021), and FLEURS (Conneau et al., 2022). While some may argue that FLEURS and/or BABEL was originally designed to be out-of-domain evaluation, we note that many works now include it as an in-domain dataset for both supervised and unsupervised training (Peng et al., 2023b; Pratap et al., 2023; Chen et al., 2023b; Zhang et al., 2023).

We complement this selection of multilingual corpora with various language-specific data to boost the representation of languages that are underrepresented in large web-scale datasets. This includes many indigenous languages, such as Quechua (Cardenas et al., 2018), Mixtec (Shi et al., 2021b), and Totonac (Berrebbi et al., 2022).

Finally, we also make an effort to support speech outside of mainstream accents and voices. For example, we include code-switching (Lyu et al., 2010), accented data (AI4Bharat, 2020; Sanabria et al., 2023; Ahamad et al., 2020), and even singing voices (Smule, 2019).

A.2 Ablations

To understand the effectiveness of E-Branchformer and WavLM denoising in the multilingual setting, we conducted several SSL experiments at a smaller scale. We sampled 7,000 hours of data from our 1 million hour SSL dataset and trained a 95M parameter HuBERT Base model on that data (ML-HuBERT 7K). Then, we trained a variant with the addition of the acoustic denoising task and another with the Transformer layers replaced with E-Branchformer layers. We then benchmarked each model on ML-SUPERB using the same settings as Section 5.1.1. Our results show meaningful gains achieved with the addition of the WavLM objective and E-Branchformer architecture.

Table 7: Ablations on the ML-SUPERB 1-hour set.

Model	Multi.ASR + LID		
	CER	CER (Few-Shot)	ACC
ML-HuBERT 7K	36.9	**40.5**	78.5
+ Denoising	30.5	41.2	84.3
+ E-Branchformer	**29.0**	40.9	**85.2**

A.3 Dereverberation

We perform a preliminary investigation of on the effectiveness of our proposed SSL dereverberation task (Section 4.2) by training two HuBERT Base models (Hsu et al., 2021) on LibriSpeech 960 (Panayotov et al., 2015). The first model corresponds to the vanilla HuBERT setting without any form of data augmentation, while the second model is trained with our proposed dereverberation technique. We fine-tune each model on 10-hours of LibriLight ASR (Kahn et al., 2020), where we use performance on test-clean and test-other as a proxy for comparing performance on clean and noisy data respectively. Results are in Table 8, with improvements across both test sets. The benefits are more significant on the noisier test-other, with relative reductions of 6.9% Word Error Rate (WER), indicating the effectiveness of our technique for enhancing model robustness. Performance on test-clean is also improved by 3.1%.

Table 8: Investigation on effectiveness of dereverberation augmentation on English ASR by WER.

Model	test-clean	test-other
HuBERT	13.1	22.6
+ Dereverberation	**12.7**	**21.1**

A.4 Pre-Training Details

Table 9 provides a breakdown of XEUS' the computational costs, measured in CPU and GPU hours. We used AMD EPYC 7763 processors for CPU-based jobs. For most GPU tasks, we use a mix of

Table 9: Computation budget of XEUS in CPU/GPU hours. Reported numbers for formatting are in CPU hours, while all other stages are measured in GPU hours.

Stage	Hours
Data Preparation	100,000
Pseudo-labelling	15,000
Ablations	2,100
Scaling	200
Pre-Training	63,000

Nvidia A40 46GB and Nvidia A100 40GB GPUs. For pre-training the final model, we exclusively used A100s.

We consumed around 100K CPU hours for the data stage. The bulk of this usage came from processing the YODAS dataset, which originally consisted of document-level WAV files at various sampling rates. We had to convert each file into 16kHz audio samples and then segment each into utterance-level pieces. Another large source of CPU consumption came from downloading each dataset and unarchiving them onto the disk, which became a non-trivial effort for larger datasets such as VoxPopuli.

Obtaining the HuBERT pseudo-labels required a large amount of compute, totalling 15,000 GPU hours. While one may argue that this process could have been avoid by using another SSL objective, such as data2vec (Baevski et al., 2022) or w2v-BERT (Chung et al., 2021), we believed that this cost was worth the guaranteed stability during large-scale training. Self-distilling SSL objectives like data2vec are prone to representation collapse (Baevski et al., 2022; Liu et al., 2024), while wav2vec derived models require a codebook diversity loss (Baevski et al., 2020; Chung et al., 2021). As our experimental runs were limited, we believed the simple HuBERT objective would be the easiest to scale.

The ablations described in Sections A.3 and A.2 consumed a total of 2,100 GPU hours, while the hyperparameter tuning required for scaling (masking ratio, learning rate, warmup steps) consumed 200 GPU hours.

The largest source of compute usage came from the final pre-training phase, which consumed 63,000 GPU hours across a total of 40 days. We use bfloat16 mixed precision with Flash Attention v2 (Dao et al., 2022) for faster pre-training. To improve convergence, for the first 3,000 steps we include an additional intermediate cross entropy SSL loss (Yang et al., 2023), an initial masking probability of 0.65, and no data augmentation. This intermediate loss is applied to the 10th encoder layer and is weighted by a factor of 0.3. Afterwards, we remove the intermediate loss for greater efficiency, increase the masking probability to 0.8, and enable the above augmentation methods. We use the Adam optimizer with 32,000 warmup steps and a peak learning rate of 0.0003.

A.5 Engineering Details

This section expands upon the engineering optimizations that we made on the ESPnet toolkit mentioned in Section 4.4. We organize it into two main components: GPU communication and batching. Quantitative analyses of these optimizations are reported in Table 10.

Table 10: Quantitative results of our engineering improvements. We report percent increases in relative throughput and percent decreases in relative memory usage after our optimizations.

Optimization	Throughput (↑)	Memory (↓)
GPU Synchronization	120%	-
Batch Optimization	60%	113%

GPU Communication: As the dataset size increases, scaling to a larger number of GPUs is important to finish pre-training in a reasonable time. However, this leads to non-trivial overhead due to communication costs across multiple compute nodes. During our training, we found and removed 2 unnecessary GPU synchronization steps within the ESPnet training code. We also disabled synchronization in iterations without an optimization step (when using gradient accumulation), further improving runtime. These changes are particularly impactful in compute clusters that lack Infiniband support for inter-node communication, which is common in more academic-oriented data centers.

Batching: The batching method has a large impact on both the memory efficiency and throughput of model training. Ideally, utterances of similar length should be batched together to reduce the amount of padding. This can be complicated during multi-node training, especially when there is large amounts of variance between utterance lengths. We found that ESPnet had issues over-allocating data to batches, which had remained hidden as it only noticeable given a sufficiently large batch size and number of GPUs (such as our case). Fixing this

issue more than halved our memory usage given a fixed batch size. We also found that ESPnet randomizes the batches across GPUs regardless of the sequence length. This means that one GPU may process a few utterances that are 30 seconds long, while another may process many utterances less than a second long. As the GPUs need to be synchronized during the backwards pass, this sequence length mismatch causes unnecessary waiting. By enforcing length-aware batch distribution, we are able to improve the model throughput by 60%.

A.6 Experimental Setups

This section details the hyperparameters we searched for our experiments, which we then used to obtain and report the best performing results for each model.

A.6.1 ML-SUPERB:

The ML-SUPERB benchmark is designed to be a lightweight downstream probe of the multilingual representation quality of SSL models. The benchmark is split across two data settings: 10 minutes and 1 hour of data for each language. Each data setting has 4 tasks: monolingual ASR across 9 languages, multilingual ASR, LID, and joint multilingual ASR+LID. In the multilingual ASR tasks, 5 languages are reserved as few-shot task with 5 examples per language, while the remaining 138 languages have the standard 10 min. / 1 hour of fine-tuning data. An overall SUPERB_s score for each model u is calculated relative to the performance of the best performing model for each task t and filterbank-based features. This is obtained with the following formula:

$$\text{SUPERB}_s(u) = \frac{1000}{T}\sum_{t}^{T}\frac{1}{M_t}\sum_{m}^{M_t}\frac{s_{t,m}(u) - s_{t,m}(\text{filterbank})}{s_{t,m}(\text{SOTA}) - s_{t,m}(\text{filterbank})} \quad (1)$$

Where M_t is the set of metrics for task t, such that $s_{t,m}(u)$ yields the score of model u on metric m in M_t for task t. SOTA represents the best model for a given metric of each task. Since our model sets new multiple SOTA results, we re-calculate this score for each model following Shi et al. (2023a,b).

The downstream probe is a 2-layer Transformer encoder trained using CTC (Graves et al., 2006) loss. It has a hidden size of 256, a feed forward size of 1024, and 8 attention heads. Each task has a fixed number of training steps and enforces a constant learning scheduler with the Adam (Kingma and Ba, 2015) optimizer. The only hyperparameter we adjust during the evaluation is the learning rate, testing values of 0.00004, 0.0001, and 0.0004.

A.6.2 FLEURS:

We adopt an identical setup to (Peng et al., 2023a) for the FLEURS evaluations, which remains the SOTA design when not using additional labeled data. The downstream model consists of a 12 layer E-Branchformer encoder paired with a 6 layer Transformer decoder. The SSL model remains frozen, and a weighted sum of its layer-wise inputs are input into the downstream encoder. Each encoder layer has a convolution kernel of 31, 8 attention heads, a hidden size of 512, and a feed forward size of 2048. Each decoder layer also has 8 attention heads, a hidden size of 512, and a feed forward size of 2048. The model is trained with the joint CTC/attention (Watanabe et al., 2017) objective with a CTC weight of 0.3. To better model the diversity in language, the encoder is trained with self-conditioned CTC (Nozaki and Komatsu, 2021; Chen et al., 2023c) using the LID+ASR labels. Inference is performed using joint CTC/attention decoding with a language model, using a beam size of 10, a CTC weight of 0.3, and language model weight of 0.4. The model is trained using the Adam optimizer with a Noam-style learning rate scheduler (Vaswani et al., 2017) and a peak learning rate of 0.002. We keep all of these hyperparameters consistent across each SSL model, only varying the batch size when memory limitations are encountered.

A.6.3 JesusFilm ST:

We collected the ST data from `jesusfilm.org`, which contains 2 hour long video dramas about the life of Jesus that are parallel in various languages. Each video contains multiple male and female speakers that appear throughout the drama. We downloaded the audio for Rajbanshi, Hijazi Arabic, and Lumun, while the ST labels are derived from the captions of the English audio. We process the data by splitting the 2-hour long videos sequentially, such that the first 70% of the movie is used as the training set, the next 15% is used as the development set, and the final 15% is used as the test set. We use this method instead of random splitting to minimize the potential content and speaker overlap between the data splits. We manually clean the test set by filtering out segments with non-speech descriptions such as laughter or

background music cues. For reference, the XEUS pre-training data contains 7 hours of Rajbanshi, 3 hours of Hijazi Arabic, and 0.5 hours of Lumun.

To improve model convergence, we re-use the FLEURS models from Section A.6.2 and individually fine-tune them on ST for each language pair. We keep the fine-tuning parameters identical across all models, using a constant learning rate of 0.0001 with the Adam optimizer and a fixed batch size of 32. Inference is performed with a beam size of 10.

A.6.4 Speech Resynthesis:

Speech resynthesis experiments are performed on the VCTK dataset. For each SSL model we compare, we experiment with features extracted at 50% and 75% of the SSL model depth, which are standard settings used in discrete unit speech generation (Barrault et al., 2023a; Maiti et al., 2024). For the 19-layer XEUS model, this means we test layers 10 and 14. For the 24-layer WavLM and w2v-BERT 2.0 v2 models, we trial layers 12 and 18. The features are then clustered at the frame-level via K-means to obtain the discrete units. We use a fixed value of $K = 100$. We then train unit-to-speech HiFiGAN vocoders for each set of discrete units. Each vocoder is trained to generate 16 kHz speech for 150K steps. We use identical hyperparameters that correspond to the default VCTK settings in `https://github.com/kan-bayashi/ParallelWaveGAN` for each trial. We then select the best performing layer for each model based off of the MOSNet score of the resynthesized speech on the development set, and report its performance on the test set.

A.6.5 SUPERB:

The SUPERB Benchmark tests SSL models on a diverse selection of tasks. The SSL model is frozen and a learned weighted sum of its layer-wise inputs are fed into the respective downstream probe fine-tuned for each task. To limit the hyperparameter search space, we only tune the learning rate and batch size. Otherwise, we use the same settings as the original benchmark (wen Yang et al., 2021). For each task, the batch size is set to the maximum amount that can fit within a 40GB GPU. For the learning rate, we begin with the settings used by Chen et al. (2022) and conduct a sweep of those values multiplied by [0.25, 0.5, 1.0, 2.0, 4.0].

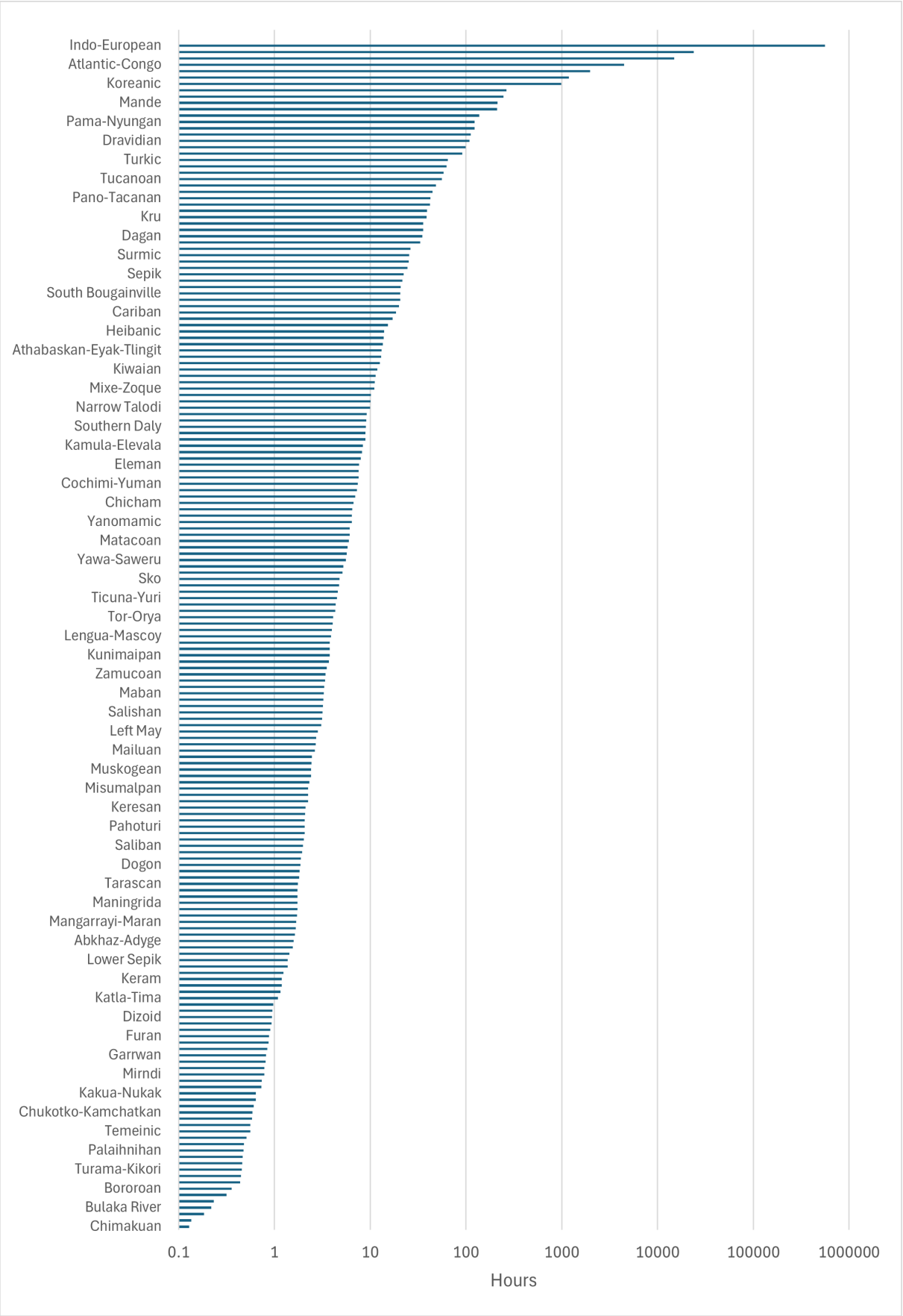

Figure 3: Distribution of data between the 189 language families in the XEUS pre-training data. We use Glottolog (https://glottolog.org/) to automatically map each ISO3 code to a language family.

Table 11: Overview of datasets used for pre-training. The language column indicates the language used in monolingual datasets and the number of languages in multilingual datasets. **Bolded** dataset names indicate new corpora we will release.

Dataset	License	Language(s)	Domain	Hours
YODAS (Li et al., 2023)	CC BY 3.0	140	Variety	422K
VoxPopuli (Wang et al., 2021)	CC BY-NC 4.0	23	Legal	400K
LibriLight (Kahn et al., 2020)	MIT	English	Audiobook	60K
MLS (Pratap et al.)	CC BY 4.0	8	Audiobook	44K
People's Speech (Galvez et al., 2021)	CC-BY 4.0	English	Variety	30K
WeNetSpeech (Zhang et al., 2022)	CC BY 4.0/SA	Mandarin	Variety	22K
Russian Open STT (Slizhikova et al., 2020)	CC-BY-NC	Russian	Variety	20K
NPTEL2020 (AI4Bharat, 2020)	CC	Indian English	Talk	15K
Reazonspeech (Yin et al., 2023)	Apache 2.0	Japanese	Television	15K
Common Voice 13 (Ardila et al., 2020)	CC0-1.0	92	Read	13K
GigaSpeech (Chen et al., 2021)	Apache 2.0	English	Variety	10K
VoxLingua (Valk and Alumäe, 2021)	CC BY 4.0	107	Variety	6800
MMS-unlab v2	CC BY-NC 4.0	4,023	Religious	6700
SPGI (O'Neill et al., 2021)	-	English	Finance	5000
Fisher (Post et al., 2013)	LDC	English	Conversation	2000
Googlei18n (Chen et al., 2023b)	Varies	34	Variety	1328
BABEL (IARPA)	IARPA Babel License	17	Conversation	1000
FLEURS (Conneau et al., 2022)	CC BY 4.0	102	News	1000
KSponSpeech (Bang et al., 2020)	MIT	Korean	Conversation	970
LibriSpeech (Panayotov et al., 2015)	CC BY 4.0	English	Audiobook	960
MagicData (Yang et al., 2022)	CC BY-NC-ND 4.0	Mandarin	Conversation	755
mTEDx (Salesky et al., 2021)	CC BY-NC-ND 4.0	8	Talk	753
Jesus Dramas	CC BY-NC 4.0	430	Religious	643
Althingi (Helgadóttir et al., 2019)	Apache 2.0	Icelandic	Legal	542
TEDLIUM3 (Hernandez et al., 2018)	CC BY-NC-ND 3.0	English	Talk	500
VoxForge (VoxForge)	GPL	8	Read	235
AISHELL (Bu et al., 2017)	Apache 2.0	Mandarin	Read	200
SEAME (Lyu et al., 2010)	LDC	Codeswitching	Conversation	192
DAMP-MVP (Smule, 2019)	Smule Research Data License	English	Singing	150
NorwegianParl. (Solberg and Ortiz, 2022)	CC0	Norwegian	Legal	140
AIDATATANG (aid)	CC BY-NC-ND 4.0	Mandarin	Read	140
AMI (Carletta, 2007)	CC BY 4.0	English	Meetings	100
Nahuatl (Shi et al., 2021a)	CC BY-NC-SA 3.0	Nahuatl	Conversation	82
WSJ (Paul and Baker, 1992)	LDC	English	Read	81
Mixtec (Shi et al., 2021b)	CC BY-NC-SA 3.0	Mixtec	Conversation	70
WikiTongues	CC BY-NC 4.0	700	Conversation	70
Siminchik (Cardenas et al., 2018)	CC BY-NC-ND 3.0	Quechua	Radio	50
Edinburgh Accent (Sanabria et al., 2023)	CC-BY-SA	Accented English	Conversation	40
VCTK (Yamagishi et al., 2019)	CC BY 4.0	English	Read	25
AccentDB (Ahamad et al., 2020)	CC BY-NC 4.0	Indian English	Read	20
Totonac (Berrebbi et al., 2022)	CC BY-NC-SA 3.0	Totonac	Monologue	17