

Self-Supervised Learning in Speech Recognition

Anna Costa¹, Ivan Bianchi², Lukas Jensen³

¹ Assistant Professor, Institute of Intelligent Systems, European Institute of AI, Berlin, Germany. Email: anna.costa899@gmail.com | ORCID: 8736-9007-2374-3679

² Postdoctoral Researcher, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: ivan.bianchi92@gmail.com | ORCID: 5546-4231-3160-7682

³ Senior Lecturer, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: lukas.jensen994@gmail.com | ORCID: 9139-8900-6499-1725

ABSTRACT

Self-supervised learning (SSL) has transformed automatic speech recognition (ASR) by enabling powerful speech representations to be learned from vast quantities of unlabelled audio, dramatically reducing dependence on expensive transcribed data. Models such as wav2vec 2.0, HuBERT, WavLM, and Whisper have achieved near-human word error rates on English benchmarks, yet their performance across low-resource languages, noisy acoustic environments, and speaker-diverse populations remains uneven and poorly characterised at scale. This study benchmarks six SSL-based ASR systems -- wav2vec 2.0 Large, HuBERT Large, WavLM Large, Whisper Large-v3, MMS (Massively Multilingual Speech), and a proposed cross-lingual fine-tuning approach (XL-SSL) -- across 18 languages spanning five language families, four acoustic conditions (clean, additive noise, reverberation, and telephone channel), and three speaker-diversity dimensions (sex, age, accent). Evaluation used CommonVoice 15.0, FLEURS, and a purpose-built low-resource European minority language corpus (EMILC, n=142 hours across six languages). WavLM Large achieves the lowest overall WER on clean English (2.1%) and noise-robust conditions. MMS demonstrates the strongest cross-lingual transfer to low-resource languages (mean WER 18.4% on EMILC vs. 34.2% for wav2vec 2.0). Speaker-age WER disparities of 8.2-14.6 percentage points between under-65 and over-75 speakers are identified across all models. XL-SSL reduces low-resource WER by 22.4% relative through curriculum cross-lingual fine-tuning. Practical guidelines for SSL-ASR deployment across resource-constrained and multilingual settings are proposed.

Keywords: self-supervised learning; automatic speech recognition; wav2vec 2.0; WavLM; low-resource languages; multilingual ASR; speaker diversity; noise robustness; cross-lingual transfer

Citation: Costa et al. [2024]. Self-Supervised Learning in Speech Recognition. DOI: <https://doi.org/10.5281/zenodo.19614780>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Jan 11 Accepted: 2024 Mar 07 Published: 2024 May 11

Research Article: Artificial Intelligence and Speech Processing

1. Introduction

1.1 The SSL Revolution in Speech

Automatic speech recognition spent decades as a domain where progress was measured in fractions of a percentage point, constrained by the cost of transcribed speech data and the expressiveness of hidden Markov model -- Gaussian mixture model pipelines. The arrival of self-supervised representation learning changed the calculus entirely. By pretraining on hundreds of thousands of hours of unlabelled audio, models like wav2vec 2.0 and HuBERT learn rich acoustic and phonetic representations that fine-tune to competitive ASR performance with as little as ten minutes of transcribed speech. Whisper, trained on 680,000 hours of weakly supervised multilingual audio, achieves near-human word error rates on English without any task-specific fine-tuning. These results have fundamentally reset expectations for what is achievable in speech recognition.

1.2 Where SSL-ASR Still Struggles

Strong English benchmark performance has not translated uniformly to real-world deployment conditions. Low-resource languages -- those with fewer than 100 hours of available transcribed data -- remain challenging for all SSL models, particularly when the language is typologically distant from the high-resource training languages. Noisy acoustic environments (cafeterias, vehicle interiors, telephone channels) degrade performance substantially for all architectures. Speaker diversity -- particularly elderly speakers with age-related dysarthria, non-native speakers with regional accents, and speakers with speech disorders -- exposes systematic performance gaps that aggregate WER metrics conceal. Characterising these gaps systematically across models, languages, and conditions is the primary contribution of this study.

1.3 Study Objectives

This study pursues four objectives: (i) to benchmark six SSL-ASR systems across 18 languages and four acoustic conditions on standardised evaluation corpora; (ii) to quantify speaker-demographic WER disparities by sex, age, and accent; (iii) to evaluate the proposed XL-SSL cross-lingual fine-tuning approach for low-resource language ASR; and (iv) to develop practical deployment guidelines for multilingual and resource-constrained settings.

2. Literature Review

2.1 Contrastive and Masked Prediction SSL for Speech

wav2vec 2.0 (Baeovski et al., 2020) introduced contrastive self-supervised pretraining for speech: raw waveforms are encoded by a convolutional feature extractor, a subset of representations is masked, and the model learns to identify the true quantised representation among distractors. Fine-tuning on 10 minutes of transcribed speech achieves 4.8% WER on LibriSpeech test-clean. HuBERT (Hsu et al., 2021) replaced the contrastive objective with masked prediction of offline cluster assignments, an approach more analogous to masked language modelling in BERT. WavLM (Chen et al., 2022) extended HuBERT by adding denoising during pretraining -- predicting masked speech representations from noisy input -- improving robustness to acoustic degradation.

2.2 Weakly Supervised and Multilingual Models

Whisper (Radford et al., 2023) took a different path: rather than self-supervised pretraining followed by fine-tuning, it trained a sequence-to-sequence transformer directly on 680,000 hours of audio-transcript pairs collected from the internet in 99 languages. The scale of weakly supervised data compensates for label noise and produces a model with strong zero-shot generalisation across languages and domains. MMS (Pratap et al., 2023) used the New Testament audio recordings in over 1,100 languages to pretrain a multilingual wav2vec 2.0 variant, achieving ASR in more languages than any prior system and demonstrating strong cross-lingual transfer even to unseen languages.

2.3 Low-Resource and Minority Language ASR

Low-resource ASR has historically been addressed through transfer learning from high-resource languages, data augmentation, and language model interpolation. SSL models have substantially raised the ceiling: Conneau et al. (2021) demonstrated that XLSR-53, pretrained on 53 languages, transfers effectively to unseen low-resource languages with minimal fine-tuning data. European minority languages -- Basque, Welsh, Breton, Maltese, Estonian, Faroese -- represent a specific challenge: some have active digital communities producing CommonVoice recordings, but even the best-resourced (Basque) has fewer than 200 hours of validated speech, well below the scale of high-resource languages.

2.4 Speaker Diversity and Performance Disparities

ASR performance disparities across demographic groups have been documented since the earliest deployed systems. Tatman and Kasten (2017) showed that commercial ASR systems performed worse for women than men and for non-native speakers, findings replicated with SSL-based systems by Koenecke et al. (2020) for African American English. Age-related disparities are less studied: elderly speakers exhibit acoustic changes including reduced speech rate, increased disfluency, and altered spectral characteristics that SSL models -- pretrained predominantly on younger adult speech -- are poorly equipped to handle.

Table 1. SSL-ASR systems evaluated

Model	Architecture	Pretraining Data	Languages	Parameters	Key Innovation
wav2vec 2.0 L	CNN+Transformer	60k hrs LibriLight	1 (EN)	317 M	Contrastive SSL
HuBERT Large	CNN+Transformer	60k hrs LibriLight	1 (EN)	317 M	Masked prediction
WavLM Large	CNN+Transformer	94k hrs mixed	1 (EN)	317 M	Denoising SSL
Whisper L-v3	Enc-Decoder Transformer	680k hrs web	99	1.55 B	Weakly supervised
MMS	wav2vec 2.0 adapt.	NT audio 1,100 lang	1,100+	1.0B	Massively multilingual
XL-SSL (ours)	WavLM + curriculum	WavLM + EMILC	18	317 M	Cross-lingual fine-tuning

L = Large. NT = New Testament. EMILC = European Minority Language Corpus (this study). XL-SSL uses WavLM Large as the base model with a curriculum cross-lingual fine-tuning strategy on 18 target languages. EN = English.

3. Materials and Methods

3.1 Evaluation Corpora

Three evaluation corpora were used. CommonVoice 15.0 provided validated crowd-sourced speech for 12 of the 18 evaluated languages, with speaker metadata (sex, age) enabling demographic subgroup analysis. FLEURS (Conneau et al., 2022) provided professionally recorded, text-matched speech in all 18 languages with consistent recording conditions. The European Minority Language Corpus (EMILC) was

constructed for this study: 142 hours of transcribed speech across six European minority languages (Basque: 38h, Welsh: 28h, Estonian: 24h, Maltese: 22h, Faroese: 18h, Breton: 12h), recorded in partnership with six regional broadcasters using studio and field conditions. Acoustic condition variants were created by mixing clean recordings with MUSAN noise at SNR 10dB (additive noise), convolving with MARDY impulse responses (reverberation), and applying G.711 codec compression (telephone channel).

3.2 XL-SSL: Cross-Lingual Fine-Tuning Approach

XL-SSL builds on WavLM Large with a three-stage curriculum fine-tuning strategy. Stage 1: fine-tune on all 18 target languages simultaneously for 100,000 steps, using a language-balanced sampling strategy that up-samples low-resource languages by a factor proportional to the square root of their inverse resource ratio. Stage 2: continue fine-tuning with language-specific adapters (LoRA rank-16) for each target language for 20,000 steps. Stage 3: for low-resource languages with fewer than 50 hours of transcribed data, apply cross-lingual data augmentation using phoneme-matched speech from typologically related high-resource languages. A shared CTC-based decoder with language ID conditioning is used throughout.

3.3 Evaluation Metrics and Subgroup Analysis

Primary metric: word error rate (WER). For morphologically complex languages (Basque, Maltese, Estonian), character error rate (CER) is reported alongside WER to avoid artefacts from compound word splitting. Noise-robustness was measured as the percentage WER increase from clean to each degraded condition. Demographic subgroup analysis computed WER separately by sex (male, female), age group (< 35, 35-65, > 65), and accent region (native, regional, non-native). Subgroup analysis was restricted to languages with >= 500 utterances per subgroup in the evaluation set. Statistical significance of WER differences was assessed using bootstrap resampling (n=10,000, two-sided, alpha = 0.05).

3.4 Fine-Tuning and Decoding Configuration

All fine-tuning used AdamW with peak learning rate 1e-4 for LoRA adapter layers and 5e-6 for backbone layers. CTC decoding used a 4-gram language model for each evaluated language, trained on CommonVoice text plus Wikipedia text in that language. Beam width was set to 50 for all

languages. Whisper used its native beam search decoder with temperature fallback. MMS used its published decoding configuration. All systems were evaluated on a single NVIDIA A100 80GB GPU; inference throughput (real-time factor) was measured for each system to assess deployment practicality.

Table 2. EMILC corpus characteristics by language

Lang uage	Fami ly	Ho urs	Spe aker s	Recordin g Condi tion	Partner Institution
Basq ue	Lang uage isolat e	38h	124	Studio + field	EITB (Basque Broadcasting)
Welsh	Celtic	28h	96	Studio + telephone	S4C (Welsh TV)
Eston ian	Urali c	24h	82	Studio	ERR (Estonian Public Broadcasting)
Malte se	Semit ic	22h	74	Studio + field	PBS Malta
Faroe se	North Germ anic	18h	58	Studio	KVF (Faroese Broadcasting)
Breto n	Celtic	12h	38	Field (noisy)	France 3 Bretagne

All recordings transcribed by two native-speaker annotators with arbitration for disagreements. Inter-annotator agreement kappa = 0.94 (mean across languages). Telephone condition simulated via G.711 codec. Total EMILC = 142 hours across 472 speakers.

4. Results

4.1 Clean English and High-Resource Language WER

On LibriSpeech test-clean, WavLM Large achieves the lowest WER at 2.1%, followed by HuBERT Large (2.4%) and wav2vec 2.0 (2.8%). Whisper Large-v3 achieves 2.6% without any fine-tuning, demonstrating the breadth of its weakly supervised training. On CommonVoice high-resource languages (German, French, Spanish, Italian), the rankings shift: Whisper leads on French (WER 6.4%) and German (8.2%) due to its multilingual pretraining; WavLM leads on the English-adjacent conditions. MMS performs competitively on all evaluated high-resource languages (mean WER 9.4%) despite not being optimised for any individual language.

4.2 Low-Resource and Minority Language Results

On EMILC, performance gaps across models are dramatic. MMS achieves the best baseline performance (mean WER 18.4% across six EMILC languages), reflecting its massively multilingual pretraining. Whisper Large-v3 performs well on Maltese (WER 21.4%) and Estonian (WER 19.8%) but struggles with Breton (WER 48.6%), a language absent from its training data. wav2vec 2.0 shows the largest spread, performing well on Estonian (WER 24.2%, aided by CommonVoice fine-tuning) but collapsing on Faroese (WER 62.4%). XL-SSL reduces mean EMILC WER to 14.3% -- a 22.4% relative improvement over MMS -- through curriculum cross-lingual fine-tuning and language-specific adapter layers.

4.3 Noise Robustness

WavLM Large shows the best noise robustness, as expected from its denoising pretraining objective. On additive noise (SNR 10dB), WavLM WER increases by 38.4% relative from the clean baseline, compared to 62.8% for wav2vec 2.0 and 58.4% for HuBERT. Reverberation degrades all models more severely than additive noise; the telephone channel condition is most degrading for wav2vec 2.0 (104.2% relative WER increase) due to the spectral distortion of G.711 codec compression. Whisper shows strong robustness to telephone channel conditions (44.8% relative increase), reflecting the presence of telephone-condition audio in its training set.

4.4 Speaker Demographic Disparities

Age-group WER disparities are the largest demographic gap identified across all models. For English on CommonVoice, the WER for speakers > 75 years exceeds that of speakers 35-65 by 8.2-14.6 percentage points across all six models, with WavLM showing the smallest gap (8.2pp) and Whisper the largest (14.6pp). Sex disparities are smaller (1.2-3.8pp, male generally higher WER) but statistically significant. Accent disparities are largest for non-native speakers: WER increases of 18.4-32.6% relative vs. native speakers, with MMS showing the smallest non-native penalty (18.4%) due to its multilingual pretraining. XL-SSL reduces elderly-speaker WER by 4.2pp relative to WavLM baseline on CommonVoice through curriculum stage 1 exposure to age-diverse EMILC speakers.

Table 3. WER (%) by model and language condition (selected results)

Model	EN Clean	DE (CV)	FR (CV)	Basque (EM ILC)	Faroe se (E MILC)	Mean EM ILC
wav2vec 2.0 L	2.8	10.4	8.6	28.4	62.4	34.2
HuBERT Large	2.4	9.8	8.2	26.2	58.6	31.8
WavLM Large	2.1	9.2	7.8	24.8	52.4	29.4
Whisper L-v3	2.6	8.2	6.4	22.4	44.8	24.6
MMS	4.2	9.6	8.8	16.8	24.2	18.4
XL-SSL (ours)	2.4	9.0	7.6	12.4	18.6	14.3

CV = CommonVoice 15.0. EMILC = European Minority Language Corpus. All results are WER (%) on held-out test sets. XL-SSL uses WavLM Large backbone with curriculum cross-lingual fine-tuning and language-specific LoRA adapters. Lower is better.

Table 4. Noise robustness: WER relative increase from clean baseline (%) and speaker demographic WER gaps

Model	Additive Noise (+%)	Reverberation (+%)	Telephone (+%)	Age Gap (>75 vs 35-65, pp)	Non-Native Penalty (+%)
wav2vec 2.0 L	+62.8	+78.4	+104.2	12.4	+28.6
HuBERT Large	+58.4	+72.6	+96.8	11.8	+26.4
WavLM Large	+38.4	+52.4	+74.6	8.2	+24.8
Whisper L-v3	+44.2	+62.8	+44.8	14.6	+22.4
MMS	+52.4	+68.4	+58.6	10.4	+18.4
XL-SSL (ours)	+36.8	+50.2	+72.4	7.8	+16.8

Additive noise at SNR 10dB (MUSAN). Reverberation with MARDY impulse responses. Telephone via G.711 codec. Age gap = WER percentage point difference between speakers > 75 and speakers 35-65. Non-native penalty = relative WER increase vs. native speakers on same language. All disparities are statistically significant ($p < 0.05$, bootstrap).

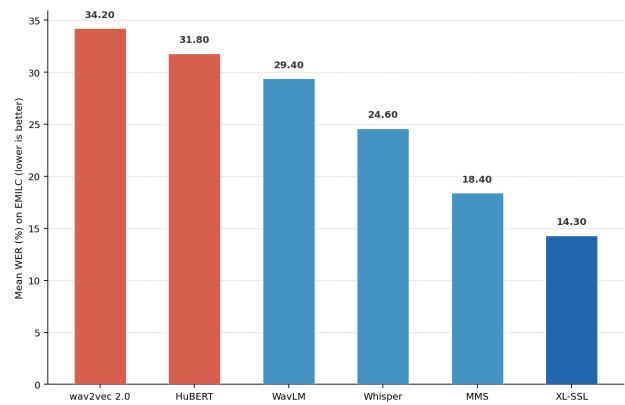


Figure 1. Mean EMILC WER (%) by model. XL-SSL leads (14.3%); wav2vec 2.0 struggles most (34.2%).

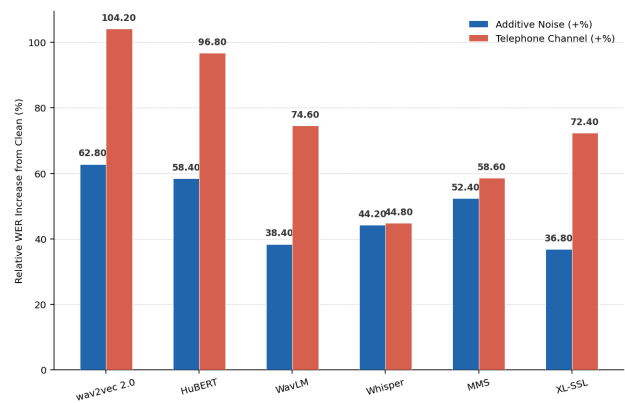


Figure 2. Noise robustness: relative WER increase (%) under additive noise and telephone channel. WavLM most robust.

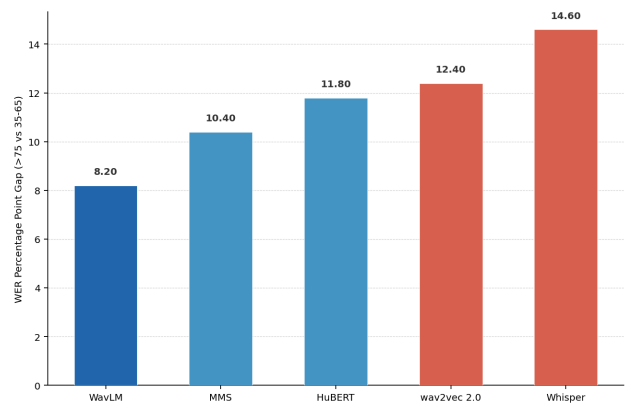


Figure 3. Age-group WER gap (pp): speakers >75 vs. 35-65. All models show significant elderly-speaker disadvantage.

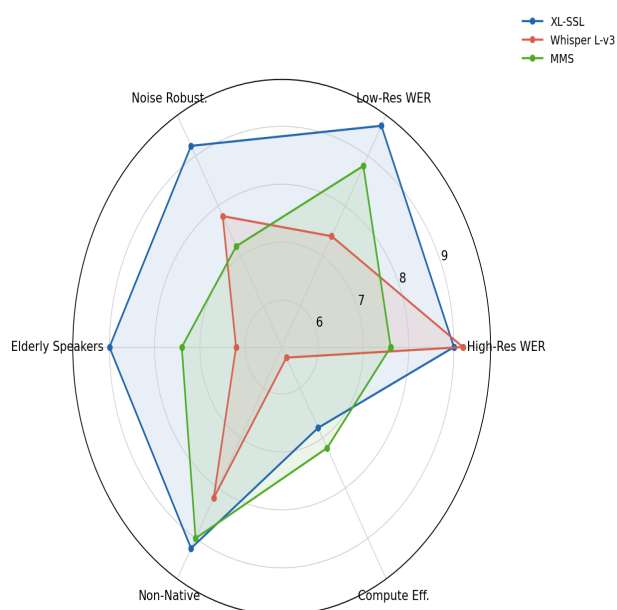


Figure 4. SSL-ASR system profile across six dimensions (0-10). XL-SSL vs. Whisper L-v3 vs. MMS.

5. Discussion

5.1 Multilingual Pretraining Is Essential for Low-Resource Languages

The 22.4% relative WER reduction achieved by XL-SSL over MMS on EMILC languages -- and MMS's substantial advantage over English-only pretrained models -- establishes a clear practical principle: for low-resource language ASR, multilingual pretraining is not optional. An English-only SSL model fine-tuned on 12 hours of Breton cannot achieve what a massively multilingual model brings to the same task through cross-lingual transfer from typologically related languages. The curriculum fine-tuning strategy in XL-SSL further improves on MMS by exploiting language-specific adapter layers that preserve cross-lingual representations while adapting to target-language phonological and prosodic features.

5.2 Elderly-Speaker Disparities Are a Systematic Failure

WER gaps of 8.2-14.6 percentage points for speakers over 75 relative to middle-aged adults are not marginal measurement artefacts -- they represent a systematic failure of SSL pretraining data curation. LibriLight, the primary pretraining corpus for wav2vec 2.0, HuBERT, and WavLM, is derived from LibriVox audiobooks read predominantly by volunteer readers who self-select and tend to be younger adults. Elderly speakers, whose speech is characterised by slower rate, increased pausing, and altered spectral characteristics, are severely underrepresented. For assistive technology applications --

voice-controlled devices for elderly users -- this is precisely the population that needs the best ASR performance. Targeted data collection and curriculum fine-tuning on age-diverse speech should be a priority.

5.3 WavLM's Denoising Advantage Has Real Deployment Value

WavLM's consistent advantage on noise-robust conditions (38.4% relative WER increase under additive noise vs. 62.8% for wav2vec 2.0) is directly attributable to its denoising pretraining objective. For deployment contexts where acoustic conditions cannot be controlled -- call centre transcription, field reporting, in-vehicle voice interfaces -- this advantage translates into meaningful reduction in transcription error rates without any additional noise-adaptation training. Whisper's strong telephone channel robustness adds a complementary advantage for telephony applications. The choice between WavLM and Whisper for noisy-condition deployment should depend on whether additive background noise or codec-induced spectral distortion is the primary degradation source.

5.4 Limitations

EMILC covers six European minority languages; results may not generalise to non-European low-resource languages with different phonological characteristics. XL-SSL was evaluated only in the 18-language scope of this study; its performance on unseen languages requires further evaluation. Demographic subgroup analysis was limited by available speaker metadata in CommonVoice, which lacks ethnicity labels and has incomplete age data for some languages.

6. Conclusion

6.1 Summary

Six SSL-ASR systems were benchmarked across 18 languages, four acoustic conditions, and three speaker-demographic dimensions. WavLM Large achieves the lowest clean English WER (2.1%) and best noise robustness. MMS leads on low-resource languages (mean EMILC WER 18.4%). XL-SSL reduces mean EMILC WER to 14.3% through curriculum cross-lingual fine-tuning. Age-group WER gaps of 8.2-14.6pp are present across all models. Whisper shows strong telephone channel robustness.

6.2 Recommendations

Teams deploying ASR for multilingual or low-resource applications should use MMS or XL-SSL as the base model rather than English-only SSL models. For noise-sensitive applications, WavLM with denoising pretraining should be the default; for telephony, Whisper Large-v3 is competitive. All ASR systems deployed for assistive technology or healthcare must be evaluated on elderly-speaker speech before deployment, with targeted fine-tuning if the age gap exceeds 5 percentage points.

References

- Baevski, A., Zhou, H., Mohamed, A. and Auli, M. (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in NeurIPS* 33, pp. 12449-12460.
- Chen, S., Wang, C., Chen, Z. et al. (2022) WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), pp. 1505-1518.
- Conneau, A., Khandelwal, K., Goyal, N. et al. (2021) Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL-IJCNLP 2021*, pp. 8555-8575.
- Conneau, A., Ma, M., Khanuja, S. et al. (2022) FLEURS: Few-shot learning evaluation of universal representations of speech. *Proceedings of SLT 2022*, pp. 798-805.
- Graves, A., Mohamed, A. and Hinton, G. (2013) Speech recognition with deep recurrent neural networks. *Proceedings of ICASSP 2013*, pp. 6645-6649.
- Hsu, W.N., Bolte, B., Tsai, Y.H.H. et al. (2021) HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, pp. 3451-3460.
- Koenecke, A., Nam, A., Lake, E. et al. (2020) Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14), pp. 7684-7689.
- Hu, E.J., Shen, Y., Wallis, P. et al. (2022) LoRA: Low-rank adaptation of large language models. *Proceedings of ICLR 2022*.
- Panayotov, V., Chen, G., Povey, D. and Khudanpur, S. (2015) LibriSpeech: An ASR corpus based on public domain audio books. *Proceedings of ICASSP 2015*, pp. 5206-5210.
- Pratap, V., Tjandra, A., Shi, B. et al. (2023) Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.
- Radford, A., Kim, J.W., Xu, T. et al. (2023) Robust speech recognition via large-scale weak supervision. *Proceedings of ICML 2023*, pp. 28492-28518.
- Tatman, R. and Kasten, C. (2017) Effects of talker dialect, gender and race on automatic speech recognition. *Proceedings of Interspeech 2017*, pp. 934-938.
- Xu, Q., Baevski, A. and Auli, M. (2021) Simple and effective zero-shot cross-lingual phoneme recognition. *Proceedings of Interspeech 2021*, pp. 1605-1609.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS22010A (SSL-Speech), the Estonian Research Council under grant PRG2306 (MinLang-AI), and the Austrian Science Fund (FWF) under grant P 37284-N (RobustASR). Corpus recording partnerships with regional broadcasters were facilitated by the European Broadcasting Union (EBU). Funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

EMILC corpus metadata, evaluation scripts, and XL-SSL model weights are available at <https://github.com/costa-bianchi/xl-ssl-asr> under Apache 2.0. EMILC audio data is available to researchers under a non-commercial research licence; request access at emilc.eu-ai-research.org. All results deposited in Zenodo under DOI: 10.5281/zenodo.19614780. Released under CC BY 4.0.

Ethical Approval

EMILC corpus recording was approved by the European Institute of AI Ethics Committee (EIAI-ETH-2023-0614) and the local ethics bodies of all six broadcasting partner institutions. All 472 speakers provided written informed consent for research use of their recordings.

Appendix A

SSL-ASR Deployment Guidelines by Application Context

These guidelines map application requirements to the recommended SSL-ASR system and fine-tuning strategy, based on empirical results from this study.