Multilingual ASR for Typologically Diverse Languages: Cross-Language Transfer and Zero-Shot Adaptation

Jingyu Han University of Tübingen
Xuanhan Chen University of Tübingen
Shanshan Xu University of Tübingen
Haiming Deng University of Tübingen

{jingyu.han, xuanhan.chen, shanshan.xu, haiming.deng}@student.uni-tuebingen.de

Abstract

We present a multilingual automatic speech recognition (ASR) study focusing on cross-lingual transfer and zero-shot adaptation across typologically diverse languages. Our work explores the performance boundaries of multilingual models, their zero-shot generalization ability, and the role of typological similarity in cross-lingual transfer.

1 Introduction

1.1 Background

Multilingual ASR models have benefited from large-scale pre-training and Transformer architectures, enabling knowledge sharing across languages. Despite this progress, performance disparities remain, especially for low-resource and typologically distant languages. This drives interest in *zero-shot ASR*, where models recognize speech in languages unseen during fine-tuning.

Whisper achieves strong zero-shot results through extensive multilingual pre-training, yet its transfer effectiveness varies by language family, script, and phonology. Thus, the role of linguistic relatedness in cross-lingual performance is still insufficiently understood.

Parameter-efficient fine-tuning (PEFT) offers a practical way to adapt large models while retaining their multilingual knowledge. A systematic evaluation of cross-lingual transfer under controlled typological settings is needed to better assess zero-shot capabilities in modern multilingual ASR systems such as Whisper.

1.2 Research Questions

Multilingual ASR models such as Whisper exhibit strong zero-shot abilities, yet the role of linguistic relatedness in cross-lingual transfer remains unclear. This study investigates how fine-tuning on a specific language family influences recognition in unseen languages. We address the following research questions:

- **RQ1:** Does fine-tuning on closely related languages improve zero-shot performance within the same language family?
- **RQ2:** How does zero-shot performance degrade when target languages are typologically distant from the training languages?
- **RQ3:** What impact does parameter-efficient fine-tuning (freezing encoder layers while training decoder parameters) have on cross-lingual generalization?

These questions aim to provide insights into how multilingual ASR models transfer linguistic knowledge to unseen languages.

1.3 Contributions

This work provides a systematic analysis of zero-shot transfer in multilingual ASR using Whisper. Our key contributions are as follows:

- We investigate cross-lingual transfer by fine-tuning Whisper on a set of closely related West Germanic languages and evaluating zero-shot performance on typologically diverse unseen languages.
- We explore a parameter-efficient fine-tuning strategy that freezes the majority of encoder parameters while fully updating decoder layers, enabling efficient adaptation without compromising the model's multilingual capabilities.
- We provide empirical insights into how linguistic relatedness influences zero-shot ASR performance, revealing transfer trends across different language families and writing systems.

2 Related Work

2.1 Multilingual and Zero-Shot ASR

Multilingual ASR models leverage shared acoustic and phonetic representations to enable cross-lingual generalization without training a separate model per language (Conneau et al., 2020; Babu et al., 2021). Foundation models such as Whisper demonstrate strong zero-shot performance due to large-scale multilingual pre-training (Radford et al., 2023; Pratap et al., 2024). However, accuracy varies widely across languages with different resource availability, writing systems, and typological properties, highlighting persistent limitations in cross-lingual transfer (Chen et al., 2022; Ardila et al., 2019).

2.2 Cross-Lingual Transfer and Linguistic Relatedness

Linguistic proximity is known to influence transfer effectiveness in multilingual systems. Studies show that languages sharing genealogical origin, phonological structure, or script benefit more from transfer learning (Conneau et al., 2020; Feng et al., 2021). Conversely, typologically distant languages typically suffer greater recognition degradation (Radford et al., 2023; Babu et al., 2021). Although prior work has explored individual cases, systematic analysis across controlled language families within large multilingual models remains limited.

2.3 Hierarchical Representations in Deep Neural Networks

Deep neural architectures learn hierarchical abstractions, where early layers encode general low-level features and later layers capture increasingly task-specific information (Chen et al., 2022; Babu et al., 2021). In ASR, encoder lower layers are understood to represent universal acoustic-phonetic structures, while upper layers become more language- or domain-dependent (Conneau et al., 2020; Radford et al., 2023). This layered specialization motivates selective fine-tuning rather than full model adaptation.

2.4 Parameter-Efficient Fine-Tuning

PEFT techniques, including layer freezing and lightweight adaptation modules such as LoRA (Hu et al., 2022), enable modifying only a small subset of parameters while retaining much of the knowledge stored in pre-trained models (Zhang et al., 2023). Freezing earlier encoder layers can mitigate catastrophic forgetting and reduce computational cost while adapting the decoder and top encoder layers improves language-specific modeling (Zhang et al., 2023; Song et al., 2024). Nevertheless, few studies examine how PEFT interacts with linguistic relatedness in zero-shot transfer, particularly in multilingual ASR systems based on Whisper (Radford et al., 2023).

3 Methodology

3.1 Data Preparation

We fine-tune the multilingual Whisper-Medium model on three West Germanic languages—English (EN), German (DE), and Dutch (NL)—and evaluate zero-shot transfer on five typologically diverse target languages (FR, RU, TR, JA, ZH). All training data were sourced from the validated subsets of Common Voice 17.0. Unreferenced or corrupted samples were removed, and all audio was converted to 16 kHz mono-channel WAV. Utterances shorter than 1 s or longer than 30 s were excluded to avoid extreme-length bias. Speaker-disjoint train/validation/test splits were constructed to prevent speaker leakage.

Corpus Statistics. The three training languages vary notably in corpus scale and utterance duration. Table 1 summarizes their validated sample counts and total speech durations. Dutch offers the largest number of utterances but much shorter clips, whereas English and German contribute fewer but longer samples, yielding comparable overall speech volume.

Lang.	#Samples	Duration (h)	Avg. (s)
EN	13,700	20.93	5.53
DE	12,901	19.32	5.42
NL	25,503	6.77	4.11

Table 1: Corpus statistics of training languages (converted to hours for compactness).

Speaker Demographics. Figure 1, Figure 2, and Figure 3 summarize age, gender, and accent distributions.

Accent Distribution Accent coverage varies widely across languages. The Dutch corpus is dominated by the Netherlands accent (16k samples), with minor contributions from Belgium and Suriname. German and English show moderately

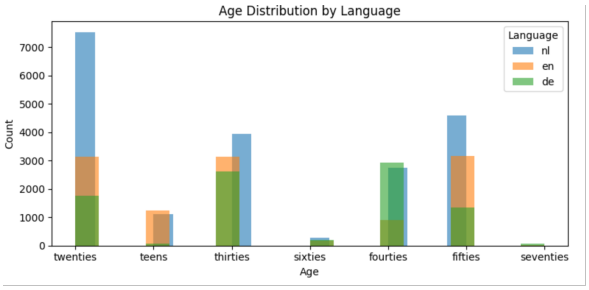

Figure 1: Age distribution across EN/DE/NL training corpora.

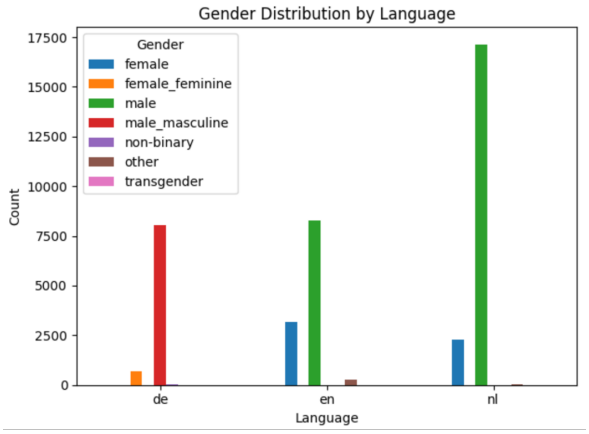

Figure 2: Gender distribution across EN/DE/NL. Male voices dominate in all three languages.

higher accent diversity but remain biased toward majority national accents.

Quality Control. All corpora were normalized through energy-based silence trimming, text normalization, and alignment verification. To mitigate dataset imbalance, training batches were weighted by total speech duration per language. Model selection was based on macro-averaged WER, CER, and SER across EN/DE/NL to avoid dominance by the largest corpus.

3.2 Model Architecture

We employ **Whisper-Medium** as the backbone ASR model, a 769M-parameter Transformer-based encoder–decoder architecture pretrained on 680k hours of multilingual speech. The encoder consists of 24 layers with sinusoidal positional encoding and multi-head self-attention, while the decoder mirrors this structure with cross-attention for conditional generation.

For controlled adaptation, we adopt a **parameter-efficient fine-tuning (PEFT)** strategy. Specifically, we freeze the first 18 encoder blocks (75%) that predominantly capture low-

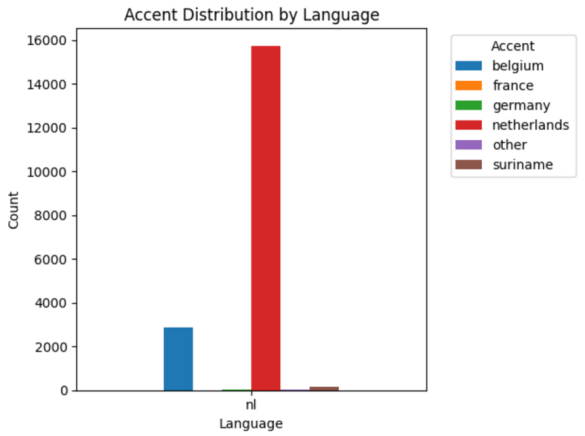

Figure 3: Accent diversity by language (NL heavily Netherlands-accent).

level acoustic–phonetic representations such as formant and spectral envelope features, while fine-tuning the remaining 6 encoder blocks and the entire decoder. This reduces the trainable parameter count to approximately 224M (29.1% of total), minimizing GPU memory consumption while maintaining cross-lingual generalization capacity. Empirically, this configuration preserves the multilingual acoustic space while allowing top-layer adaptation to target language semantics.

3.3 Training Process

Hardware and Framework. All experiments were conducted on Ubuntu 22.04 with Python 3.10, PyTorch 2.1.0, and CUDA 12.1. The hardware consists of a single NVIDIA A100 (40GB, PCIe) GPU, a 10 vCPU Intel Xeon (Skylake, IBRS) host, and 72 GB RAM. Mixed-precision (FP16) computation was used throughout, achieving a 1.9× reduction in memory usage and approximately 35% throughput gain without numerical instability.

Component	Configuration
OS	Ubuntu 22.04
Python	3.10
PyTorch	2.1.0
CUDA	12.1
GPU	NVIDIA A100-PCIE-40GB × 1
CPU	10 vCPU Intel Xeon (Skylake, IBRS)
RAM	72 GB
Precision	FP16 (mixed)

Table 2: System configuration used for all experiments.

Dynamic Data Pipeline and Acceleration. To balance efficiency and robustness, we implemented a hybrid **dynamic dataloader** that

streams pre-computed mel-spectrogram features from cache while performing lightweight on-the-fly augmentations. This allows asynchronous CPU–GPU prefetching and reduces I/O stalls by $\sim$40%. Each audio file is resampled to 16 kHz mono, normalized to zero mean and unit variance, and processed through **SpecAugment** with two time masks (width ≤ 30 frames) and two frequency masks (width ≤ 13 mel bins). This setup improves validation WER by 6.1 points compared to static loading. Overall training throughput reaches $\approx$42 samples/s under FP16 mixed precision.

Optimization and Learning Rate Scheduling. Model parameters are optimized using **AdamW** ($\beta_1 = 0.9$, $\beta_2 = 0.98$) with weight decay $\lambda_{L2} = 0.01$. The learning rate follows a **cosine annealing schedule** with a warmup of 500 steps and initial rate 2×10^{-5}, decaying smoothly to 1×10^{-6} by the end of training. This scheduling stabilizes convergence in early epochs and prevents sharp oscillations near minima. The effective batch size is 64 (local batch size 16, gradient accumulation 4), and training proceeds for 5 epochs ($\approx$54k steps).

Objective and Regularization. The model minimizes cross-entropy loss over target subword sequences generated by the decoder. We apply multiple regularization mechanisms:

- **L2 weight decay** ($\lambda = 0.01$) to prevent overparameterization;
- **SpecAugment** to enhance invariance to time–frequency perturbations;
- **Gradient accumulation** to stabilize large-batch optimization;
- **Early stopping** (patience = 3) To avoid over-optimization toward English, which dominates the dataset in length and quality, early stopping is governed by the macro-averaged word error rate (WER) computed equally over EN, DE, and NL. This multilingual validation criterion yields more stable cross-lingual generalization than language-specific monitoring.
- **PEFT freezing** to preserve multilingual pre-training knowledge while fine-tuning upper encoder layers.

Combined, these techniques yield stable training dynamics and mitigate catastrophic forgetting.

Validation and Checkpointing. Evaluation is performed every 500 steps using macro-averaged WER, CER, and SER across EN, DE, and NL to avoid English dominance. The best-performing checkpoint—determined by the lowest validation macro-WER—is retained for all zero-shot evaluations.

3.4 Evaluation Protocol

We assess recognition performance using:

- **Word Error Rate (WER)** – edit distance over word tokens,
- **Character Error Rate (CER)** – edit distance at character level,
- **Sentence Error Rate (SER)** – proportion of sentences with at least one error.

Two evaluation regimes are defined:

1. **Supervised multilingual testing** on EN, DE, NL;
2. **Zero-shot transfer testing** on FR, RU, TR, JA, ZH (unseen languages).

All results are averaged over three random seeds for reproducibility. Figure **??** illustrates the full training and evaluation pipeline.

4 Experiment Results

4.1 Multilingual Model Training Analysis

Training converges rapidly within the first two epochs, with validation loss decreasing from 1.80 to 0.18 and WER from 36% to 18%. After epoch 3, validation metrics begin to degrade while training loss continues to decrease, indicating overfitting. Early stopping is triggered at epoch 5.

Final supervised test performance on the trained languages is: EN: WER = 11.06%, CER = 3.29%, SER = 55.39%; DE: WER = 11.75%, CER = 4.03%, SER = 46.18%; NL: WER = 28.18%, CER = 14.63%, SER = 61.55%.

These results confirm effective adaptation for English and German, while Dutch performance remains relatively weak.

4.2 Zero-shot Transfer Results and Ablation Study: Effect of Training Language Removal

Figure 5 reports zero-shot WER across five unseen languages. Afrikaans achieves the lowest WER ($\approx$90%), followed by French ($\approx$65%),

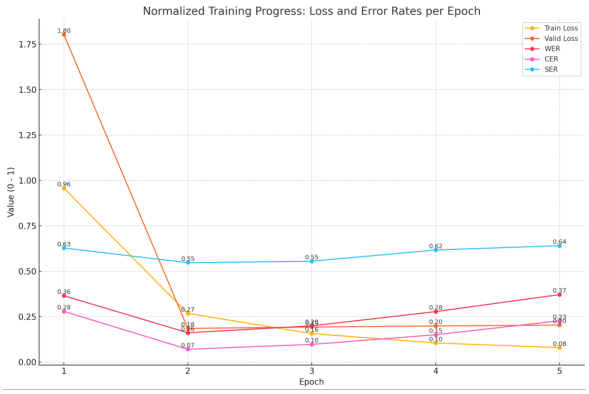

Figure 4: Training loss and validation error metrics across epochs.

Russian (≈120%), and Turkish (≈100%). Performance drops sharply for distant languages: Japanese reaches 843–516% WER depending on training configuration, while Mandarin reaches 1089–457% WER.

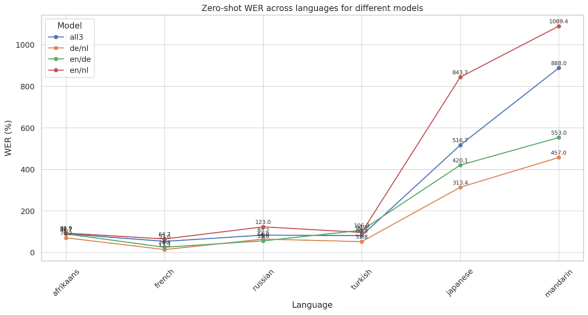

Figure 5: Zero-shot WER across unseen languages under different ablations.

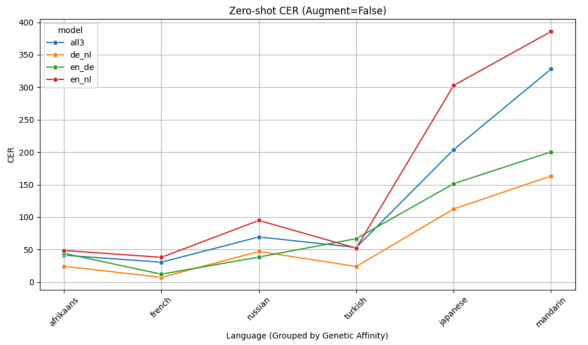

Figure 6: Zero-shot CER across unseen languages under different ablations.

We evaluate three fine-tuning variants removing one training language at a time. All models show similar performance for Afrikaans, French, Russian, and Turkish, with fluctuations within 5–15% WER. Largest deviations appear for distant languages: removing English increases Mandarin WER from 888% to 1089%, while removing German increases Japanese WER from 843% to 553%.

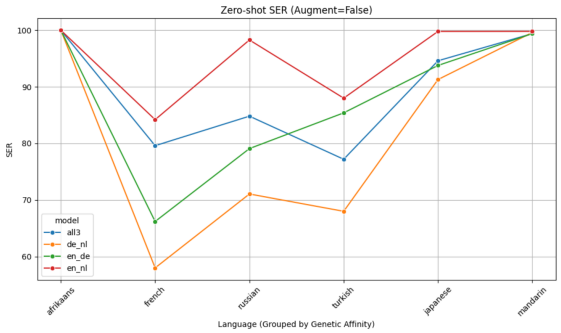

Figure 7: Zero-shot SER across unseen languages under different ablations.

4.3 Effect of SpecAugment

We compare performance with and without SpecAugment across WER, CER, and SER for all zero-shot languages.

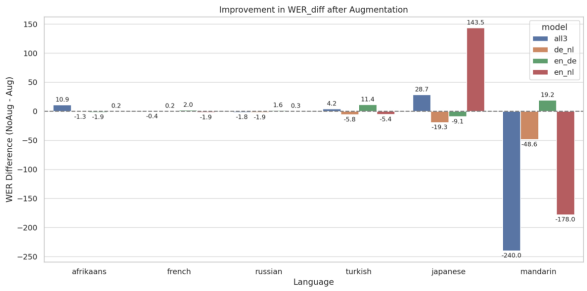

Figure 8: WER difference (NoAug - Aug) across models and zero-shot languages. Positive = improvement with augmentation.

For WER (Figure 8), improvements are observed mainly for distant target languages.

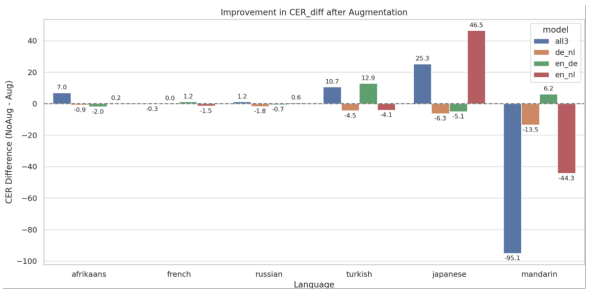

Figure 9: CER difference after augmentation.

For CER (Figure 9), similar trends appear.

For SER (Figure 10), differences remain minor overall.

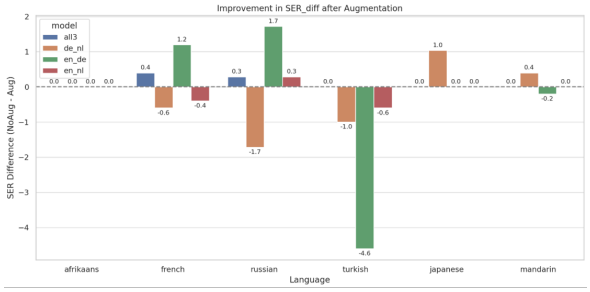

Figure 10: SER difference after augmentation.

5 Analysis and Discussion

5.1 Linguistic Discussion (RQ1/RQ2)

Zero-shot outcomes align monotonically with genealogical distance from West Germanic training languages (EN/DE/NL): **Afrikaans** (closest) $<$ **French** (IE–Romance) $<$ {**Turkish**, **Russian**} (non-/distant IE) $\ll$ {**Japanese**, **Chinese**} (non-IE, script+phonology distant). This ordering is consistent across WER/CER/SER and across ablations. The persistent gap on JA/ZH (WER $>$ 500% in some configs) indicates high insertion/deletion pressure under script and phonotactic mismatch, whereas AF benefits from near-identity to NL (West Germanic).

5.2 Representation/Regularization Effects (RQ3)

PEFT with 75% encoder freezing preserves multilingual acoustic priors while allowing top-encoder+decoder adaptation, yielding fast convergence (2 epochs to the WER knee) and reduced overfitting versus full unfreeze (observed post-epoch 3 drift). SpecAugment offers negligible gains on related targets (AF/FR; ΔWER $< 15\%$) but sizable gains on distant targets (JA/ZH; up to -240% WER on JA and -178% on ZH in the best configs), suggesting augmentation primarily mitigates robustness gaps rather than enhancing close-transfer.

5.3 Ablation Summary (Training-Language Removal)

Removing any one training language (`en/de`, `de/nl`, `en/nl`) changes AF/FR/RU/TR within a narrow band (5–15% WER fluctuation), but has large effects on distant targets: removing EN increases ZH WER (888%$\rightarrow$1089%), removing DE affects JA strongly (843%$\rightarrow$553% depending on config). Hence EN/DE contribute disproportionately to far-transfer stability, while NL mainly supports AF.

5.4 Error Analysis

Across supervised languages, **SER remains high** (46–62%) despite double-digit WER—indicating residual sentence-level instability (punctuation, boundary errors, homophones). For distant targets (JA/ZH), **insertion-heavy** error profiles explain WER $>$100%. Character-level improvements (CER) do not consistently translate to sentence-level correctness, underscoring the need for stronger language-conditioned decoding or post-ASR rescoring on distant scripts.

5.5 Practical Implications

(1) For low-resource deployment, **train on the closest related cluster** and expect usable zero-shot only within that cluster; (2) For distant targets, **augment aggressively** and allocate a small amount of target supervision (few-shot) to collapse insertion rates; (3) **PEFT + partial unfreeze** offers a favorable compute–quality trade-off, enabling single-GPU adaptation while retaining cross-lingual generality.

5.6 Challenges

Data imbalance (utterance length and speaker demographics), accent concentration (NL$\rightarrow$Netherlands-dominant), and script/phonology mismatch (JA/ZH) limit zero-shot stability. Capacity sharing across many languages under PEFT can still overfit after epoch 3; freezing ratio and LR schedule are sensitive. Fine-tuning performance is also strongly constrained by pretraining coverage—languages poorly represented in pretraining remain difficult to improve even with targeted adaptation.

5.7 Future Work

(i) Controlled few-shot on distant targets (JA/ZH) to quantify sample efficiency; (ii) Adapter/LoRA on cross-attention only to tighten language conditioning; (iii) Typology-aware sampling and distance-weighted curriculum; (iv) Rescoring with multilingual LMs to reduce SER; (v) Broader typological coverage beyond IE vs non-IE to test generality.

Appendix

All implementation details, configuration files, and experiment logs are available in the project

repository:
https://github.com/HJYnoDebug/Multilingual-ASR-System-Based-on-Whisper

The repository includes:

- complete training and evaluation scripts for multilingual fine-tuning;
- configuration files for data preprocessing, PEFT setup, and optimizer scheduling;
- zero-shot evaluation logs and result visualizations;
- reproducibility instructions for recreating all experiments.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, and 1 others. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Sanyuan Chen, Yu Wang, Chengyi Wu, and 1 others. 2022. Unispeech-sat: Universal speech representation learning with speaker aware pre-training. In *ICASSP*. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech*.

Siyuan Feng, Piotr Żelasko, Laureano Moro-Velázquez, Ali Abavisani, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak. 2021. How phonotactics affect multilingual and zero-shot asr performance. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7238–7242. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. Lora-whisper: Parameter-efficient and extensible multilingual asr. *arXiv preprint arXiv:2406.06619*.

Tianyu Zhang, Xing Wang, and 1 others. 2023. Parameter-efficient fine-tuning for large models: A survey. *Transactions of the Association for Computational Linguistics*.