From Tens of Hours to Tens of Thousands: Scaling Back-Translation for Speech Recognition

Tianduo Wang[♡*] **Lu Xu**[♢] **Wei Lu**[♣] **Shanbo Cheng**[♠♢†]
[♡]Singapore University of Technology and Design, [♠]Nanjing University
[♢]ByteDance Seed [♣]Nanyang Technological University
tianduo_wang@sutd.edu.sg
{xu.lu1,chengshanbo}@bytedance.com
wei.lu@ntu.edu.sg
https://github.com/tianduowang/speech-bt

Abstract

Recent advances in Automatic Speech Recognition (ASR) have been largely fueled by massive speech corpora. However, extending coverage to diverse languages with limited resources remains a formidable challenge. This paper introduces Speech Back-Translation, a scalable pipeline that improves multilingual ASR models by converting large-scale text corpora into synthetic speech via off-the-shelf text-to-speech (TTS) models. We demonstrate that just tens of hours of real transcribed speech can effectively train TTS models to generate synthetic speech at hundreds of times the original volume while maintaining high quality. To evaluate synthetic speech quality, we develop an intelligibility-based assessment framework and establish clear thresholds for when synthetic data benefits ASR training. Using Speech Back-Translation, we generate more than 500,000 hours of synthetic speech in ten languages and continue pre-training Whisper-large-v3, achieving average transcription error reductions of over 30%. These results highlight the scalability and effectiveness of Speech Back-Translation for enhancing multilingual ASR systems.

1 Introduction

Automatic Speech Recognition (ASR) technology has become increasingly important in making digital services accessible across languages and modalities (Baevski et al., 2020; Zhang et al., 2021; Radford et al., 2022). While recent transformer-based architectures have achieved impressive results for high-resource languages, e.g., English and Chinese, many of the world's languages still lack sufficient transcribed speech for training robust ASR models (Pratap et al., 2020a; Babu et al., 2021; Chen et al., 2024). This data scarcity creates a significant barrier to developing effective multilingual speech technologies, particularly affecting communities where manual data collection is resource-intensive or logistically challenging (Costa-jussà et al., 2022; Communication et al., 2023; Pratap et al., 2023).

A natural way to mitigate the data scarcity issue is to leverage high-quality generative models. Recent work has demonstrated successful applications of these models for data augmentation in computer vision (Fan et al., 2023; Azizi et al., 2023), natural language processing (Gunasekar et al., 2023; Li et al., 2024), and speech recognition (Yang et al., 2024). Despite their demonstrated potential, the role of generative models in overcoming data scarcity presents a paradox. These models themselves typically demand vast amounts of labeled data to attain their remarkable capabilities. For instance, Stable Diffusion (Rombach et al., 2022), a leading text-to-image model frequently used for data augmentation (Tian et al., 2023; Trabucco et al., 2023), was trained on millions of labeled images. This reliance prompts a fundamental question: do synthetic data truly alleviate data scarcity in downstream tasks, or do they simply shift the burden of data collection to the pre-training stage of generative models?

Our work investigates whether an off-the-shelf text-to-speech (TTS) model can be trained with limited real transcribed speech data—just tens of hours—to generate synthetic data that enhances multilingual ASR models. To address this challenge, we propose **Speech Back-Translation** (see Figure 1), a scalable method that build large-scale synthetic transcribed speech from text corpora with TTS models. Our results demonstrate that synthetic data, generated by TTS models trained on just tens of hours of labeled audio, can effectively expand small human-labeled datasets to tens of thousands of hours. To assess the quality of this back-translated synthetic dataset, we propose a novel intelligibility-based metric and use it to establish thresholds indicating when

[*]Work done during internship at ByteDance
[†]Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12462–12476
November 4-9, 2025

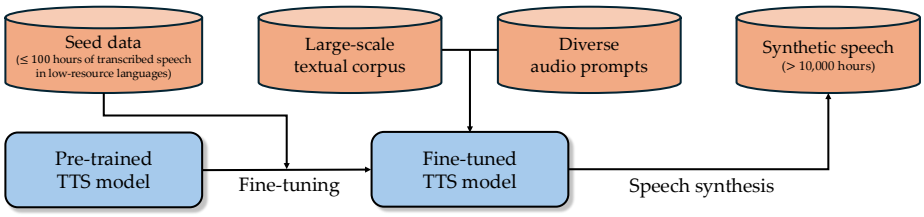

Figure 1: Pipeline of **Speech Back-Translation**. The main objective is to augment limited training data ($\leq$100 hours) for low-resource languages by synthesizing extensive amounts of speech ($>$10,000 hours). Starting from a multilingual TTS model pre-trained with high-resource languages, we fine-tune it on a small set of seed data, then generate synthetic speech by conditioning the fine-tuned model on a large textual corpus and diverse audio prompts.

synthetic speech reliably enhances ASR performance. Finally, we scale Speech Back-Translation to 500K hours across ten languages and continue pre-training Whisper-large-v3[1], one of the state-of-the-art multilingual ASR models. As a result, we observe consistent improvements across all languages, achieving an average reduction of over **30%** in transcription error rates. To summarize, our main contributions are listed as follows:

1. We demonstrate that just tens of hours of real transcribed speech can effectively train TTS models to generate tens of thousands of high-quality synthetic speech, achieving a scaling factor of several hundred.

2. We introduce an intelligibility-based evaluation framework for synthetic speech and establish thresholds to determine when synthetic data reliably benefits ASR performance.

3. We build the largest synthetic speech dataset to date—500K hours spanning ten languages—and use it to further pre-train Whisper-large-v3. This yields an average 30% reduction on transcription error rates, highlighting the scalability of our approach.

2 Background

2.1 Back-Translation

Back-translation is a data augmentation technique originally used in machine translation to expand training data (Sennrich et al., 2016a; Edunov et al., 2018). In a typical setup, a model trained to translate from the target language back into the source language (i.e., a "*reverse*" model) is used to generate synthetic source sentences from real target-language data. These newly created source-target pairs can then be used to train a forward translation model, effectively increasing its exposure to a broader range of textual content. For speech recognition, back-translation offers a mechanism to supplement scarce or imbalanced datasets by leveraging an abundance of target-side text. Here, the "*reverse*" model is typically a text-to-speech model that generates synthetic speech from textual corpora. Integrating this synthetic speech with existing training data allows the model to handle a wider range of speech variability, enhancing recognition performance despite resource constraints.

2.2 Zero-shot Text-to-Speech Model

Zero-shot Text-to-Speech (TTS) models (Wang et al., 2023; Casanova et al., 2024) represent a milestone in speech synthesis, enabling the generation of high-quality speech for previously unseen speakers without additional fine-tuning. These models typically contain the following components:

- **Audio Tokenizer**: Encodes raw acoustic inputs (e.g., mel-spectrograms) into discrete audio tokens, forming the basis for synthesis.
- **Speaker Embeddings**: Contain speaker-specific acoustic features, which are normally extracted from audio clips, enabling zero-shot adaptation to new voices.
- **Decoder-only Transformer**: Processes speaker embeddings alongside textual tokens to generate sequences of audio tokens. The Transformer model is trained in an auto-regressive manner.
- **Vocoder**: Converts the generated audio tokens into waveform audio, producing the final synthesized output.

The synergy of these components allows zero-shot TTS models to generalize effectively to speakers not encountered during training, maintaining high voice similarity and naturalness.

[1] https://hf.co/openai/whisper-large-v3

3 Approach: Speech Back-Translation

In this section, we introduce the proposed Speech Back-Translation (see Figure 1). we first detail how we extend existing TTS models to support new low-resource languages with fine-tuning (Section 3.1). We then describe how we generate large-scale synthetic speech dataset (Section 3.2).

3.1 Fine-tuning with Low-resource Languages

Obtaining high-quality transcribed speech for low-resource languages poses a significant challenge for multilingual ASR training. To address this, we extend existing multilingual TTS models—originally trained on high-resource languages—to new, low-resource languages via targeted fine-tuning with limited data.

Vocabulary Expansion Before fine-tuning, we expand the vocabulary of pre-trained TTS models to accommodate words not encountered during the initial training phase. We employ the Byte-Pair Encoding algorithm (Sennrich et al., 2016b) on textual data from the target language, appending the newly derived subwords to the model's original vocabulary. This approach preserves the integrity of the existing vocabulary while enabling effective representation of new linguistic units.

Limited Data Fine-tuning Given the scarcity of transcribed speech data, we adopt a conservative fine-tuning strategy: we freeze modules responsible for low-level acoustic representations, such as the audio tokenizer and vocoder, while selectively fine-tuning only the transformer part of the TTS model. This ensures stability in fundamental acoustic modeling while effectively adapting linguistic and prosodic mappings to the target language. During fine-tuning, each pair of audio and transcript data is processed by first extracting a speaker embedding $\mathbf{e}$ from the audio clip. Then, we tokenize both the transcript and audio clip, concatenating the S text tokens $\mathbf{x} = [x_1, \ldots, x_S]$ and T audio tokens $\mathbf{y} = [y_1, \ldots, y_T]$ into $\mathbf{z} = [z_1, \ldots, z_{S+T}]$. The training objective minimizes the negative log-likelihood of sequence $\mathbf{z}$ conditioned on the speaker embedding $\mathbf{e}$:

$$\mathcal{L} = -\sum_{t=1}^{S+T} \log\, p\big(z_t \mid z_1, \ldots, z_{t-1}, \mathbf{e}\big), \quad (1)$$

Quality Estimation Evaluating the performance of fine-tuned models is essential before deploying them for large-scale synthetic data generation. Intelligibility—commonly measured as the Word Error Rate (WER) using a robust ASR system—has emerged as the standard metric for assessing synthetic speech quality (Wang et al., 2023; Casanova et al., 2024). Yet this conventional method has two drawbacks: (1) the judge ASR introduces its own errors, particularly in low-resource languages; and (2) absolute WER values are not comparable across languages. To alleviate these issues, we propose a novel metric called *Normalized Intelligibility*, leveraging ASR performance on natural speech as a reference baseline. We use the Fleurs dataset (Conneau et al., 2022), which provides high-quality audio-transcript pairs across 102 languages, and Whisper-large-v3 as our judge ASR system. By synthesizing speech using transcripts from Fleurs, we measure two WER scores for each language: WER on synthetic speech (WER_s) and WER on real speech (WER_r). Normalized Intelligibility (Norm_I) is defined as:

$$\text{Norm\_I} = \exp\left(\frac{\text{WER}_r - \text{WER}_s}{\text{WER}_r}\right) \quad (2)$$

This formulation offers several advantages: (1) it normalizes ASR performance across languages using real speech as a baseline, (2) it enables meaningful cross-language comparisons, and (3) it produces intuitive scores bounded between 0 and e, where higher values reflect better synthetic speech quality relative to natural speech.

3.2 Generating Large-scale Synthetic Speech

Zero-shot TTS converts *text* into *audio* by conditioning on two indispensable inputs: (i) an *audio prompt* that specifies the target voice style and (ii) a *text sentence* that supplies the textual content. Both inputs must therefore be covered at scale and with maximal diversity.

- **Audio Prompts**: We curate around 1 million short audio clips spanning diverse speakers and recording conditions. After strict de-duplication to remove near-identical voices, every retained clip can serve as a style prompt that the TTS model imitates. Details of data sources and filtering are provided in Appendix C.
- **Text Corpus**: To maximize linguistic variety, we sample sentences across various domains, following the data-mixing practices of recent open-source LLMs (Touvron et al., 2023a; Wei et al., 2023). Construction and statistics of the corpus appear in Appendix D.

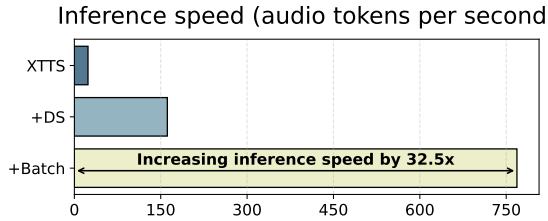

Figure 2: **XTTS inference speed** measured on a single NVIDIA V100-32GB GPU. "DS" refers to DeepSpeed-Inference while "Batch" refers to batch inference. For batch inference, we set batch size to be 16.

Inference Speed-up A key challenge in employing TTS models for large-scale dataset creation is their inference speed. We address this bottleneck using two complementary optimization techniques:

- *DeepSpeed-Inference* (Aminabadi et al., 2022): Involving fused CUDA kernels integration and optimized kernel scheduling, significantly enhancing inference throughput.
- *Batch Inference*: We group multiple sentences with similar lengths using a single audio prompt, then apply tailored attention masks to enable simultaneous generation of multiple utterances in one forward pass.

We evaluate the effectiveness of these techniques using XTTS (Casanova et al., 2024) on a single NVIDIA V100 GPU. As demonstrated in Figure 2, we observe that these optimizations yield a more than **30**× speed-up, making large-scale speech synthesis feasible for our experiments. More details can be found in Appendix A.

4 Experimental Setup

4.1 ASR Backbone Models

Our experiments leverage Whisper models (Radford et al., 2022), a family of multilingual ASR models pre-trained on 680,000 hours of labeled speech data, as the backbone. The models are available in five sizes: Tiny (39M), Base (74M), Small (244M), Medium (769M), and Large (1.5B). Further training details are provided in Appendix E.

4.2 Zero-shot TTS Models

We employ two state-of-the-art zero-shot TTS models in our experiments: XTTS (Casanova et al., 2024) and ChatTTS (2noise, 2024). XTTS supports 16 languages, covering a range of language families and resource levels, while ChatTTS only supports Chinese and English. More details about these two models can be found in Appendix B.

Model	WER↓		
	vi	cs	hu
Whisper-medium	25.4	22.5	27.8
+ Real-only	22.8	15.6	16.9
+ Speech BT	**19.0**	**10.3**	**13.2**
Whisper-large	24.5	19.9	23.8
+ Real-only	19.9	12.5	13.9
+ Speech BT	**16.0**	**9.1**	**11.1**

Table 1: **WER results for low-resource languages on Common Voice**. The "Real-only" rows indicate models trained only on tens of hours of real audio, while the "Speech BT" rows present performance achieved when expanding training data to 10K hours using our method.

4.3 Languages

Our experiments span ten languages across diverse language families and resource levels. Following Whisper's training data distribution, we categorize them based on the relative resource availability:

- High (≥10K hours): English (en), Chinese (zh), French (fr), German (de), Spanish (es)
- Mid (1K∼10K hours): Dutch (nl), Italian (it)
- Low (≤1K hours): Vietnamese (vi), Czech (cs), Hungarian (hu)

Of these languages, XTTS supports all except Vietnamese. To enable Vietnamese support, we fine-tune XTTS with 100 hours of transcribed speech sampled from viVoice[2], a high-quality dataset derived from YouTube.

4.4 Datasets

Most of our experiments use Common Voice data (Ardila et al., 2019), chosen for its high quality and broad language coverage, and it also serves as the primary training corpus for XTTS. To assess generalization, we additionally evaluate our ASR models on Voxpopuli (Wang et al., 2021) and Multilingual LibriSpeech (Pratap et al., 2020b).

5 Results

We begin by demonstrating the effectiveness of our approach in scaling limited real training data to tens of thousands of hours using synthetic speech (Section 5.1). We then evaluate the models' multilingual performance and examine their generalization to out-of-domain data (Section 5.2). Next,

[2] https://hf.co/datasets/capleaf/viVoice

Model	Common Voice (In-Domain)				Voxpopuli (Out-of-Domain)			
	High	Mid	Low	Avg. Δ	High	Mid	Low	Avg. Δ
Whisper-medium	11.5	10.6	25.2	-	11.3	21.8	23.4	-
+ Real-only	9.0	8.0	17.6	-4.0	11.0	20.9	19.9	-1.4
+ Speech BT	**8.5**	**6.1**	**11.1**	-6.6	**10.0**	**19.4**	**13.3**	-4.1
Whisper-large	10.5	9.1	21.9	-	11.4	20.3	18.1	-
+ Real-only	8.7	7.2	15.4	-5.0	10.7	19.3	16.2	-1.2
+ Speech BT	**6.6**	**5.2**	**10.7**	-6.3	**9.5**	**17.7**	**12.5**	-3.3

Table 2: **Comparison of Whisper models' WER across in-domain and out-of-domain data.** Adding 3,800 hours of Common Voice data (Real-only) provides strong in-domain gains but limited out-of-domain improvements, whereas scaling synthetic Speech BT data to 160,000 hours achieves robust gains across both domains.

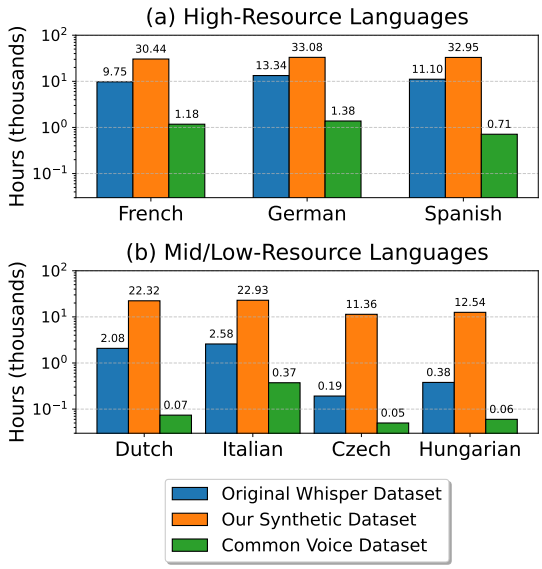

Figure 3: **Comparison of dataset sizes across seven languages** (log-scale y-axis). Languages are categorized by resource availability in the Whisper dataset: (a) high-resource, (b) mid and low-resource groups.

we analyze the relationship between TTS quality and ASR performance using our fine-tuned TTS model (Section 5.3), and explore strategies for optimally leveraging limited in-domain real data (Section 5.4). Finally, we scale the synthetic corpus to 500K hours and compare our results with prior work (Section 5.5).

5.1 From Tens of Hours to Tens of Thousands

We first assess the effectiveness of our approach by expanding the amount of training data for three low-resource languages—Vietnamese (vi), Czech (cs), and Hungarian (hu)—from mere tens of hours to ten thousand hours. As a baseline, we sample real audio in amounts matching the data originally used for TTS training (100 hours for vi, 50 hours for cs, and 60 hours for hu)[3]. Table 1 compares these "Real-only" models against models enhanced with our "Speech BT" method for both Whisper-medium and Whisper-large. Consistently across all three languages, Speech BT provides substantial gains in WER, underscoring the effectiveness of augmenting limited real speech with large-scale synthetic data.

5.2 Multilingual Performance and Out-of-Domain Generalization

To evaluate the effectiveness and scalability of our approach in a multilingual setting, we generated 160,000 hours of synthetic speech spanning seven languages at varying resource levels: French, German, and Spanish (high-resource); Dutch and Italian (mid-resource); Czech and Hungarian (low-resource). As a baseline, we also collected 3,800 hours of transcribed speech from Common Voice as the training data. Figure 3 compares our synthetic dataset with the original Whisper Dataset and Common Voice. Our synthetic dataset provides substantially more training hours than the original Whisper dataset for each language: a 3-fold increase for high-resource languages, a 10-fold increase for mid-resource languages, and a 40-fold increase for low-resource languages. While both Whisper training data and Common Voice exhibit substantial resource imbalance across languages (with high-resource languages having significantly more data than mid and low-resource ones), our Speech BT dataset maintains a more uniform distribution. This balanced allocation across language resources enables more equitable training, addressing a key limitation of naturally collected datasets.

[3]Training data for vi comes from viVoice, whereas data for cs and hu are sampled from Common Voice.

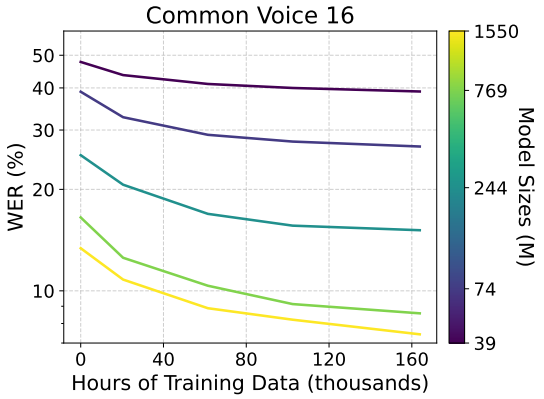

Figure 4: **Whisper's performance improves consistently with larger models and more training data.** We train five sizes of Whisper models with up to 160,000 hours of data and conduct evaluation on Common Voice 16. We report averaged WER across seven languages.

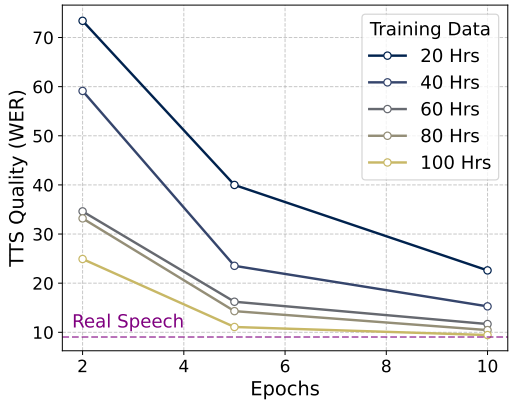

Figure 5: **Impact of training data quantity and epochs on Vietnamese TTS quality.** The purple dashed line shows the WER of natural speech from Fleurs.

Out-of-Domain Generalization Table 2 shows a detailed comparison of model performance in both in-domain (Common Voice) and out-of-domain (Voxpopuli) scenarios. Training with only real transcribed speech from Common Voice (Real-only) yields clear in-domain improvements for both Whisper-medium and Whisper-large (4.0% and 5.0% average WER reduction, respectively), but the generalization to out-of-domain data is limited (just 1.4% and 1.2% average reduction). In contrast, supplementing real data with Speech BT significantly enhances both in-domain (6.6% for Whisper-medium, 6.3% for Whisper-large) and out-of-domain performance (4.1% and 3.3%, respectively). This clearly demonstrates that our synthetic data not only improves model robustness within-domain but also enhances generalization capabilities across diverse domains.

Scalability with Model and Data Size To further assess scalability, we train five Whisper model variants—tiny, base, small, medium, and large—using the same data mentioned above. Figure 4 presents the averaged WER across all seven languages for each model size at increasing scales of training data up to 160,000 hours. The results show two clear trends. First, adding more training data consistently lowers WER across all model sizes. Second, larger models achieve substantially lower WER at each data scale. These scaling trends suggest that our Speech BT approach effectively improves multilingual ASR performance across different model and data scales.

5.3 TTS Quality vs ASR Performance

We now investigate extending TTS support to a new language, Vietnamese, using limited amounts of transcribed speech data. To explore how the quantity of training data impacts TTS model performance, we sampled datasets at increments of {20,40,60,80,100} hours and trained each for up to 10 epochs. The results, shown in Figure 5, clearly indicate that performance consistently improves as the amount of training data and the number of epochs increase. Specifically, the model trained on the 100-hour dataset reaching a WER of 10% in the end, which closely approaches the baseline WER for natural speech.

Next, we analyze the relationship between TTS model quality and downstream ASR performance. We selected several checkpoints from the fine-tuned TTS models, varying by the amount of training data and the number of epochs. For each checkpoint, we generated 100 hours of synthetic speech and subsequently used it to train Whisper-medium. We then measured the resulting changes in WER (denoted as ΔWER, where negative values indicate improvement) on the Common Voice dataset. The correlation between each checkpoint's normalized intelligibility score (see Equation 2) and ASR performance is illustrated in Figure 6. Our analysis reveals a strong correlation between TTS intelligibility scores and ASR performance improvements. Notably, we identified a critical intelligibility threshold around 0.01, serving as a clear inflection point. Below this threshold, TTS-generated speech leads to increased WER, degrading ASR performance by up to 2 points. Conversely, once

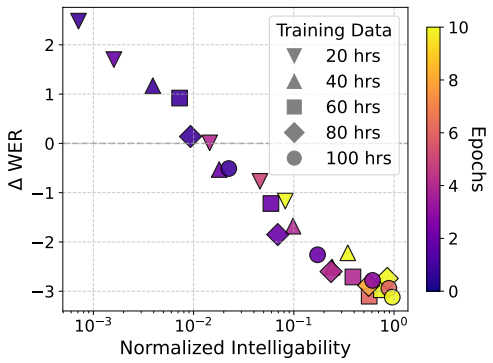

Figure 6: **Relationship between TTS quality and ASR performance.** Higher TTS intelligibility correlates with greater ASR improvement.

Models	WER↓
Whisper-medium	25.4
+ in-domain fine-tune	21.6
Approach 1	
+ Synthetic data pre-train	21.2
+ in-domain fine-tune	20.4
Approach 2	
+ Synthetic data pre-train	20.1
Approach 3	
+ Synthetic data pre-train	**18.6**

Table 3: Vietnamese WER performance on Common Voice using different approaches for leveraging limited in-domain data.

the threshold is surpassed, synthetic speech consistently enhances ASR accuracy, with greater intelligibility corresponding to more pronounced reductions in WER. This underscores the importance of achieving a minimum TTS quality level for effective ASR data augmentation. Additionally, the volume of TTS training data significantly influences the ability to surpass this intelligibility threshold. Models trained on larger datasets generally achieve higher intelligibility scores and yield greater ASR performance gains. However, we observe diminishing returns as normalized intelligibility approaches 1.0, where WER reductions stabilize around 3 percentage points. This finding suggests that while adequate training data is essential to cross the quality threshold, further improvements in ASR performance may plateau beyond a certain point.

5.4 Effective Utilization of Limited In-Domain Transcribed Audio

The experiments in Section 5.3 are essentially conducted under out-of-domain conditions, as the TTS models are trained on the viVoice dataset but the final ASR performance are evaluated with the Common Voice dataset. Notably, we identified only about three hours of transcribed audio in Common Voice Vietnamese available for training. This prompts an important research question: how can we effectively leverage such a small but valuable amount of in-domain data? We propose three methods for effectively utilizing the in-domain data to enhance model performance:

- **Approach 1:** Pre-train Whisper on large-scale synthetic data followed by supervised fine-tuning using the limited in-domain data.
- **Approach 2:** Prompt the fine-tuned Vietnamese TTS model with in-domain audio clips for speech synthesis.
- **Approach 3:** Further fine-tune the Vietnamese TTS model with in-domain data before synthesizing speech.

For all three approaches, we utilize fine-tuned XTTS checkpoints trained on 100 hours of transcribed speech and generate 1,000 hours of synthetic speech for pre-training Whisper-medium. Table 3 summarizes the resulting WERs. As a baseline, we first fine-tune Whisper-medium on three hours of in-domain data, which alone reduces WER from 25.4 to 21.6 and demonstrates the effectiveness of even limited domain adaptation. By combining synthetic speech pre-training on 1,000 hours and subsequent in-domain fine-tuning (Approach 1), we obtain further WER reduction to 20.4. However, the most pronounced improvements arise from leveraging in-domain data *within* the TTS pipeline itself (Approach 2 and 3). Simply prompting a fine-tuned XTTS model with in-domain audio achieves a WER of 20.1, already outperforming the synthetic pre-training baseline. Adding a dedicated fine-tuning stage for XTTS on the three hours of Common Voice audio (Approach 3) yields the best overall WER of 18.6—a 27.0% relative improvement over the 25.4 baseline. This underscores the value of adapting both the TTS and ASR models to the target domain, especially for low-resource languages like Vietnamese. In summary, leveraging a small in-domain dataset—through synthetic pre-training, language-specific fine-tuning, and TTS-based in-domain adaptation—proves highly effective for improving ASR performance under real-world low-resource conditions.

Model	Size	Common Voice			Voxpopuli			MLS	
		High	Mid	Low	High	Mid	Low	High	Mid
SeamlessM4T-medium	1.2B	13.3	12.8	24.4	10.7	20.0	12.6	8.0	13.0
Whisper-large-v2	1.5B	11.4	8.1	19.9	9.8	19.5	16.3	6.3	11.5
Whisper-large-v3	1.5B	10.1	5.9	15.6	12.6	28.6	14.4	5.3	10.2
+ Real-only (15K Hrs)	-	8.6	4.9	12.5	7.9	17.1	10.6	5.0	9.4
+ Speech-BT (500K Hrs)	-	**7.8**	**4.3**	**8.3**	**7.6**	**16.2**	**8.0**	**4.4**	**7.6**

Table 4: **Multilingual ASR performance on various benchmarks.** Results are averaged for each language resource category. Word Error Rate (WER) is reported for all languages except Chinese, which is measured with Character Error Rate (CER). All results are normalized with Whisper Normalizer (Radford et al., 2022).

5.5 Scaling to 500,000 Hours

Building on insights from our previous analysis, we now push the limits of multilingual ASR training with Speech Back-Translation. Starting from the baseline approach in Section 5.2, we implement several key enhancements:

Training Data Expansion We expand coverage to ten languages by incorporating three additional ones—English, Chinese, and Vietnamese. We also extend the amount of real speech: in addition to Common Voice (Ardila et al., 2019), we include real transcribed speech from Multilingual LibriSpeech (Pratap et al., 2020b), Voxpopuli (Wang et al., 2021), and viVoice, bringing the total amount of real data to 15,000 hours. Most significantly, we scale our synthetic speech dataset to 500,000 hours—a volume more than thirty times larger than the real data. The statistics of training data is illustrated in Appendix G.

Backbone Model and Baselines We adopt Whisper-large-v3, one of the state-of-the-art multilingual ASR models with 1.5B parameters, as our backbone model. For comparison, we include two ASR models with similar sizes—SeamlessM4T-medium (Communication et al., 2023) and Whisper-large-v2 (Radford et al., 2022)—as our baselines for their competitive performance and wide language coverage.

Results We evaluate both our models and baseline models on three benchmarks, Common Voice, Voxpopuli, and Multilingual LibriSpeech (MLS), and present the results in Table 4. We report the averaged results for each language category. Results demonstrate a clear performance trajectory, training Whisper-large-v3 with 15K hours of real audio consistently improves performance across all benchmarks, while augmenting with 500K hours of Speech-BT data yields further substantial gains, achieving state-of-the-art results across all language categories. On average across all benchmarks, our full model achieves a 30% error rate reduction over the base Whisper-large-v3. Breaking this down by language groups, high-resource and mid-resource languages achieve 26% and 30% improvements respectively, while low-resource languages achieve a remarkable 46% improvement. These findings indicate that augmenting real data substantially with our synthetic Speech BT data contributes significantly to advancing multilingual ASR systems, with particular benefits for traditionally underserved language communities. Detailed per-language results can be found in Appendix H.

6 Conclusion

This work introduced Speech Back-Translation, a scalable approach to address the persistent challenge of data scarcity in multilingual ASR. Our method demonstrates that TTS models trained on merely tens of hours of transcribed speech can generate hundreds of times more synthetic data of sufficient quality to significantly improve ASR performance. The large-scale implementation across ten languages with 500,000 hours of synthetic speech yielded an average 30% reduction in Whisper-large-v3's transcription error rates, confirming the effectiveness and scalability of our approach. Speech Back-Translation challenges the need for massive human-labeled datasets by effectively scaling limited data, making advanced speech recognition more accessible across diverse languages. Future work could extend to extremely low-resource languages, refine language-specific metrics, and combine with other augmentation techniques.

Limitations

While our approach demonstrates significant improvements in multilingual ASR performance, several limitations should be noted.

First, the synthetic speech data generated through TTS models may not fully capture the acoustic complexity present in real-world environments, particularly in scenarios with background noise, multiple speakers, or variable recording conditions. This limitation could impact model robustness when deployed in settings with poor signal-to-noise ratios or challenging acoustic environments.

Second, although we introduce an intelligibility-based metric for assessing synthetic speech quality, this assessment framework may not comprehensively capture all relevant aspects of speech that could influence ASR training effectiveness. Future work could explore additional quality metrics that consider factors such as prosody and emotional expression.

Third, our experimental validation is primarily based on two TTS models (XTTS and ChatTTS), which may not represent the full spectrum of TTS capabilities and limitations. A more comprehensive evaluation across a broader range of TTS systems could provide additional insights into the generalizability of our approach and identify potential TTS-specific biases or artifacts.

Lastly, while we demonstrate the scalability of our method by generating 500,000 hours of synthetic speech, our language coverage remains limited to ten languages, with nine already supported by existing TTS models. Further research is needed to validate our approach's effectiveness in other low-resource languages, particularly those with distinct phonological characteristics or limited linguistic resources.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative, (AISG Award No: AISG-NMLP-2024-005), and Ministry of Education, Singapore, under its Academic Research Fund (AcRF) Tier 2 Programme (MOE AcRF Tier 2 Award No. : MOE-T2EP20122-0011). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the National Research Foundation, Singapore, or Ministry of Education, Singapore.

References

2noise. 2024. Chattts. https://github.com/2noise/ChatTTS.

Reza Yazdani Aminabadi, Samyam Rajbhandari, Minjia Zhang, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, and Yuxiong He. 2022. Deepspeed- inference: Enabling efficient inference of transformer models at unprecedented scale. *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15.

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*, 1912.06670.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. 2023. Synthetic data from diffusion models improves imagenet classification. *ArXiv*, 2304.08466.

Matthew Baas and Herman Kamper. 2021. Voice conversion can improve asr in very low-resource settings. In *Proceedings of Interspeech*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Miguel Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of Interspeech*.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, 2006.11477.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chen Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, Lu Gao, Yi Guo, Minglun Han, Ting Han, Wenchao Hu, Xinying Hu, Yuxiang Hu, Deyu Hua, Lu Huang, Ming Huang, Youjia Huang, Jishuo Jin, Fanliu Kong, Zongwei Lan, et al. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *ArXiv*, 2407.04675.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation*.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn B. Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of ACL*.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *ArXiv*, 2406.04904.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. Towards robust speech representation learning for thousands of languages. *ArXiv*, 2407.00837.

Shanbo Cheng, Zhichao Huang, Tom Ko, Hang Li, Ningxin Peng, Lu Xu, and Qini Zhang. 2024. Towards achieving human parity on end-to-end simultaneous speech translation via llm agent. *ArXiv*, 2407.21646.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady ElSahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *ArXiv*, 2312.05187.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, et al. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, 2207.04672.

Frederico Santos de Oliveira, Edresson Casanova, Arnaldo Candido J'unior, Anderson da Silva Soares, and Arlindo R. Galvão Filho. 2023. Cml-tts a multilingual dataset for speech synthesis in low-resource languages. *ArXiv*, 2306.10097.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of EMNLP*.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. 2023. Scaling laws of synthetic images for model training ... for now. In *Proceedings of CVPR*.

Heting Gao, Kaizhi Qian, Junrui Ni, Chuang Gan, Mark A. Hasegawa-Johnson, Shiyu Chang, and Yang Zhang. 2024. Speech self-supervised learning using diffusion model synthetic data. In *Proceedings of ICML*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio C'esar Teodoro Mendes, Allison Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, S. Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuan-Fang Li. 2023. Textbooks are all you need. *ArXiv*, 2306.11644.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Proceedings of NeurIPS*.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. *ArXiv*, 2407.05361.

Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of NMT@ACL*.

Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. 2023. Speech translation with large language models: An industrial practice. *ArXiv*, 2312.13585.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *ArXiv*, 2010.05646.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of ICLR*.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, et al. 2024. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *ArXiv*, 2402.13064.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. 2024. G-dig: Towards gradient-based diverse and high-quality instruction data selection for machine translation. In *Proceedings of ACL*.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Y. Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020a. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. *ArXiv*, 2007.03001.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Mamdouh Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui

Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1, 000+ languages. *ArXiv*, 2305.13516.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020b. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, 2012.03411.

Krishna C Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, et al. 2024. Less is more: Accurate speech recognition & translation without web-scale data. *ArXiv*, 2406.19674.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *ArXiv*, 2212.04356.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2019. Zero: Memory optimizations toward training trillion parameter models. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of CVPR*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of ACL*.

Hubert Siuzdak. 2023. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *ArXiv*, 2306.00814.

Jenthe Thienpondt and Kris Demuynck. 2023. Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings. *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. 2023. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *ArXiv*, 2306.00984.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, 2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, 2307.09288.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. 2023. Effective data augmentation with diffusion models. *ArXiv*, 2302.07944.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Miguel Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *ArXiv*, 2101.00390.

Chengyi Wang, Sanyuan Chen, Yu Wu, Zi-Hua Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *ArXiv*, 2301.02111.

Tianduo Wang, Shichen Li, and Wei Lu. 2024. Self-training with direct preference optimization improves chain-of-thought reasoning. In *Proceedings of ACL*.

Tianduo Wang and Wei Lu. 2022. Differentiable data augmentation for contrastive sentence representation learning. In *Proceedings of EMNLP*.

Tianduo Wang and Wei Lu. 2023. Learning multi-step reasoning by solving arithmetic tasks. In *Proceedings of ACL*.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xue Gang Wu, Lunan Liu, Wenjun Cheng, et al. 2023. Skywork: A more open bilingual foundation model. *ArXiv*, 2310.19341.

Guanrou Yang, Fan Yu, Ziyang Ma, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. Enhancing low-resource asr through versatile tts: Bridging the data gap. *ArXiv*, 2410.16726.

Heiga Zen, Viet Dang, Robert A. J. Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Z. Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Proceedings of Interspeech*.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *Proceddings of ICASSP*.

Yu Zhang, Daniel S. Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, Zongwei Zhou, Bo Li, Min Ma, William Chan, Jiahui Yu, Yongqiang Wang, Liangliang Cao, Khe Chai Sim, Bhuvana Ramabhadran, Tara N. Sainath, Franccoise Beaufays, Zhifeng

Chen, Quoc V. Le, Chung-Cheng Chiu, Ruoming Pang, and Yonghui Wu. 2021. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 16:1519–1532.

A Inference Optimization Details

We accelerate inference by integrating DeepSpeed-Inference (Aminabadi et al., 2022) into the TTS pipeline. DeepSpeed's *deep fusion* merges multiple tiny CUDA launches into a single, highly optimized kernel that combines element-wise operations, matrix multiplications, transpositions, and reductions. Merging these operations reduce kernel-invocation overhead and off-chip memory traffic, translating into noticeably lower latency and higher throughput. We compound these gains with *batch inference*. Input sentences are grouped by language and length, then paired with a single audio prompt that supplies the target voice. Custom attention masks mark prompt–text boundaries, allowing the TTS model to synthesize multiple utterances concurrently. This batching strategy reduces redundant computations and GPU idle time, dramatically improving overall inference efficiency.

B XTTS vs ChatTTS

In this section, we present a comparative analysis of XTTS and ChatTTS for generating synthetic audio in Chinese and English. Table 5 summarizes the architectural details of both models. As the XTTS's training data mainly come from Common Voice, we treat Common Voice 16 as the in-domain dataset and Fleurs as the out-of-domain dataset for evaluation.

Performance Comparison We synthesize speech from 100K Chinese and English sentences using both models and train Whisper-medium to assess the effectiveness of these synthetic datasets. As shown in Figure 7 (a) and (b), XTTS outperforms ChatTTS on in-domain Chinese data, whereas ChatTTS excels on out-of-domain Chinese data. For English, XTTS achieves a WER of 4.0%, surpassing ChatTTS's 4.4%. These trends highlight each model's distinct strengths in handling language-specific characteristics.

TTS Quality Comparison To understand the performance difference, We assess the TTS quality with our proposed normalized intelligibility metric for both models. As shown in Figure 7, XTTS achieves superior intelligibility in English

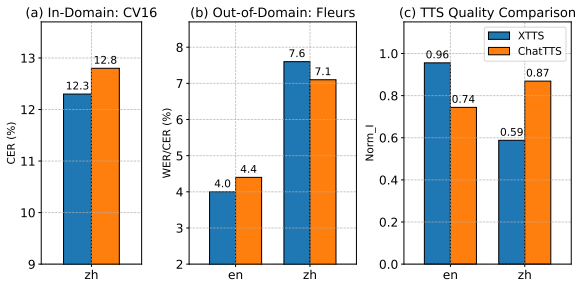

Figure 7: Comparison of Whisper-medium ASR performance on in-domain (CV16) and out-of-domain (Fleurs) test sets, as well as TTS quality, when training with synthetic Chinese and English speech generated by XTTS and ChatTTS.

(0.96 vs 0.74) while ChatTTS excels in Chinese (0.87 vs 0.59). Nevertheless, XTTS performs better on in-domain Chinese data, suggesting that while ChatTTS produces more intelligible speech in general, XTTS is more effective within the specific domain represented by Common Voice 16. This discrepancy may be attributed to domain-matched acoustic patterns and speaking styles that XTTS models more accurately. Meanwhile, on out-of-domain data (Fleurs), ChatTTS's superior general intelligibility dominates, leading to stronger performance. In English, XTTS demonstrates higher intelligibility and more robust ASR results compared to ChatTTS. Overall, these findings underscore how a TTS model's domain alignment and language-specific strengths can influence synthetic data quality and downstream ASR performance.

C Audio Prompt Details

We collect a diverse set of audio clips from various sources to serve as audio prompts for our TTS models. To prevent redundancy in voice characteristics, we extract speaker embeddings from each reference clip using the ECAPA2 speaker encoder (Thienpondt and Demuynck, 2023) and remove duplicates by comparing their cosine similarity, applying a threshold of 0.8. Table 6 summarizes the sources of these audio clips.

D Textual Data Details

Our textual corpus is sourced from a wide range of domains. Since some sources include sequences that are too long for TTS synthesis, we first segment the text using a sentencizer. We then filter out sentences that are either too short, too long, or contain an excessive number of non-alphabetic charac-

Model	Transformer			Vocabulary		Vocoder	Parameters	Lang
	Layers	Width	Heads	Text	Audio			
XTTS	30	1,024	16	6,681	1,024	Hifi-GAN (2020)	467M	16
ChatTTS	20	768	12	21,178	626	Vocos (2023)	280M	2

Table 5: Architecture details of XTTS and ChatTTS.

Dataset	Num. Clips
Emilia (He et al., 2024)	560K
CommonVoice (Ardila et al., 2019)	230K
WenetSpeech (Zhang et al., 2022)	102K
CML-TTS (de Oliveira et al., 2023)	92K
LibriTTS (Zen et al., 2019)	10K
Total	994K

Table 6: **Audio prompt distribution.** The audio clips used for voice cloning comes from various sources.

ters. To reduce redundancy, we perform sentence-level de-duplication. A detailed breakdown of our corpus sources is provided below.

Wikipedia Wikipedia is a collaborative online encyclopedia containing millions of articles, serves as a valuable source of high-quality natural text, therefore has been widely used for training language models (Touvron et al., 2023a,b).

WMT (Barrault et al., 2019) We also collected textual data from the training split of WMT19 translation task, which is a widely-used training data source in machine translation research.

Books Our Books dataset is sourced primarily from Project Gutenberg, a digital library of public domain literature. Book-level de-duplication is performed to ensure the quality and uniqueness of the corpus.

Europarl (Koehn, 2005) Europarl is a parallel corpus created for training machine translation systems, containing aligned text in European languages extracted from European Parliament proceedings. We utilize 8th version of the dataset.

SkyPile (Wei et al., 2023) SkyPile is a large-scale Chinese dataset containing approximately 150 billion tokens, curated specifically for pre-training large language models. The corpus is compiled from diverse Chinese web pages across the public internet and undergoes rigorous quality control, including thorough document-level de-duplication and content filtering.

E Training Details

Whisper We train Whisper using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1e-8$) with a weight decay of 0.01. We use constant learning rate $7e-6$ after 5% warm-up steps. To optimize distributed training, we leverage DeepSpeed ZeRO-2 (Rajbhandari et al., 2019). Additionally, we concatenate short audio clips—up to Whisper's 30-second input limit—to improve efficiency. Unless otherwise specified, our batch size is 128. In experiments presented in Section 5.2, we increase it to 768, while in Section 5.5 experiment, we further increase it to 1,024. For evaluation, we generate transcripts with greedy decoding.

XTTS Before fine-tuning, we expand the model's text vocabulary by incorporating 2,000 additional Vietnamese tokens by running Byte-Pair Encoding algorithms over Vietnamese textual data. We used the AdamW optimizer $\beta_1 = 0.9$, $\beta_2 = 0.96$, and $\epsilon = 1e-8$ with weight decay 0.01, and a learning rate of 5e-6. The batch size is set to 32.

F Related Work

F.1 Synthetic Data for Multilingual ASR

Recently we have witnessed the application of synthetic data in various domains and modalities, e.g., contrastive representation learning (Wang and Lu, 2022; Tian et al., 2023), math reasoning (Wang and Lu, 2023; Wang et al., 2024). Our work focuses on improving multilingual ASR models using synthetic audio generated by zero-shot TTS models, with particular emphasis on low-resource languages. This research builds upon previous efforts that address data scarcity through synthetic data generation. Bartelds et al. (2023) demonstrated that both self-training and TTS-generated data can effectively overcome data availability limitations in resource-scarce languages. Their work specifically examined four languages: Gronings, West-Frisian, Besemah, and Nasa, showing significant improvements in ASR performance. Baas and Kamper (2021) explored voice conversion (VC) models

for data augmentation in low-resource languages. Their key finding was that a VC system trained on a well-resourced language like English could generate effective training data for previously unseen low-resource languages. More recently, Gao et al. (2024) proposed using diffusion models to generate high-quality synthetic audio for self-supervised pre-training. The authors suggest that diffusion models are particularly adept at capturing complex speech structures from real audio, making the synthetic data especially valuable for self-supervised learning tasks.

F.2 Text-Based Back-Translation

Back-Translation (Sennrich et al., 2016a; Edunov et al., 2018) is originally proposed machine translation (Sennrich et al., 2016b; Pan et al., 2024) to augment the limited parallel training corpus from the large amount of monolingual textual data. It is designed to translate the target-language data into the source language, generatin additional synthetic parallel data that boosts overall translation quality (Sennrich et al., 2016a). This method capitalizes on monolingual text resources, which are more abundant than parallel corpora, thereby increasing model robustness and reducing overfitting. Subsequent work has explored variants of back-translation such as iterative back-translation, filtering synthetic data by quality, and domain adaptation strategies (Edunov et al., 2018; Hoang et al., 2018). In addition, dual learning frameworks have incorporated back-translation and forward-translation jointly for unsupervised and semi-supervised machine translation scenarios (He et al., 2016; Lample et al., 2018). These developments underscore the broader impact of synthetic data in enhancing model performance, even where labeled data are sparse.

F.3 Speech Translation

Beyond text-based machine translation, speech translation deals with converting audio signals in one language to either text or audio in another language, frequently via cascading automatic speech recognition and machine translation modules or through end-to-end systems (Cheng et al., 2024; Huang et al., 2023). One persistent challenge in this domain, especially for lower-resource languages, is the scarcity of paired audio-transcript data. A widely used approach to address this limitation is to create pseudo-labeled data by transcribing existing audio and then translating the resulting transcripts (Communication et al., 2023; Puvvada et al., 2024). A natural future direction for our Speech Back-Translation approach could be extended to speech translation tasks by synthesizing speech from existing parallel corpora.

Language	Amount (Hrs)	
	Real	Synthetic
English	3,951	75,159
French	2,486	94,822
German	3,706	90,782
Spanish	1,674	47,745
Chinese	204	37,910
Dutch	1,525	41,095
Italian	839	38,069
Czech	119	33,312
Hungarian	156	33,492
Vietnamese	104	13,444
Total	14,864	505,830

Table 7: Statistics of the training data in our 500K-hour experiment.

G 500K-Hour Training Data Statistics

The detailed statistics of training data used in our 500K-hour scaling up experiments are presented in Table 7.

H Additional 500K-Hour Scaling Results

In this section, we show detailed results for each languages from Section 5.5. The results for Multilingual Librispeech (MLS), Voxpopuli, and Common Voice 16 are presented in Table 8, Table 9, and Table 10 respectively. Additionally, we make comparisons with state-of-the-art multilingual ASR models: SeedASR (Bai et al., 2024), SeamlessM4T (Communication et al., 2023), Canary (Puvvada et al., 2024), and Whisper-large and Whisper-large-v2 (Radford et al., 2022).

Model	Size	High				Mid	
		en	fr	de	es	nl	it
SeedASR	-	4.1	5.1	-	3.8	-	-
SeamlessM4T-medium	1.2B	9.8	7.9	8.9	5.4	13.6	12.3
Canary	1.0B	5.1	4.4	4.7	3.4	-	-
Whisper-large	1.5B	7.2	8.8	7.4	5.3	11.1	14.1
Whisper-large-v2	1.5B	6.8	7.4	6.4	4.6	10.0	12.9
Whisper-large-v3	1.5B	5.3	5.6	6.0	4.0	10.4	9.9
+ Real-only (15K Hrs)	-	5.5	5.1	5.7	3.5	10.2	8.5
+ Speech BT (500K Hrs)	-	5.2	4.3	4.9	3.0	8.5	6.7

Table 8: Performance comparison across languages on Multilingual LibriSpeech (MLS).

Model	Size	High				Mid		Low	
		en	fr	de	es	nl	it	cs	hu
SeamlessM4T-medium	1.2B	8.2	11.8	14.0	8.8	17.2	22.8	11.0	14.1
Canary	1.0B	6.0	9.2	10.7	7.0	-	-	-	-
Whisper-large	1.5B	8.1	10.5	15.2	8.5	17.6	22.9	17.7	18.4
Whisper-large-v2	1.5B	7.9	10.4	13.1	7.9	15.8	23.2	14.3	18.3
Whisper-large-v3	1.5B	9.7	10.4	19.7	10.6	24.9	32.3	12.4	16.3
+ Real-only (15K Hrs)	-	6.0	8.9	9.6	7.0	12.5	21.7	9.5	11.7
+ Speech BT (500K Hrs)	-	5.6	8.4	9.0	7.4	12.5	19.8	7.6	8.3

Table 9: Performance comparison across languages on Voxpopuli.

Model	Size	High					Mid		Low		
		en	fr	de	es	zh	nl	it	cs	hu	vi
SeamlessM4T-medium	1.2B	11.3	14.5	12.1	9.8	18.7	15.2	10.4	14.4	34.8	24.1
Canary	1.0B	8.6	6.9	5.1	4.4	-	-	-	-	-	-
Whisper-large	1.5B	12.2	15.0	8.9	7.6	17.3	8.1	10.1	19.9	23.8	24.5
Whisper-large-v2	1.5B	11.7	13.7	7.8	6.9	16.9	6.9	9.3	16.5	20.3	22.8
Whisper-large-v3	1.5B	10.7	11.8	6.5	5.5	16.1	4.9	6.9	10.9	15.3	20.5
+ Real-only (15K Hrs)	-	9.7	8.7	5.9	4.4	14.3	4.3	5.5	9.2	11.4	16.9
+ Speech BT (500K Hrs)	-	8.8	7.3	5.0	4.2	13.6	3.7	4.9	5.2	6.0	13.6

Table 10: Performance comparison across languages on Common Voice 16.