第一页为封面页

参赛学生姓名：Aiden Ruipeng Zhou

中学：Western Academy of Beijing

省份：北京市

国家/地区：中国

指导老师姓名：徐葳

指导老师单位：清华大学

论文题目：Flow Matching-based Text-to-Speech for Low-Resource Automatic Speech Recognition Augmentation

Flow Matching-based Text-to-Speech for Low-Resource Automatic Speech Recognition Augmentation

Aiden Ruipeng Zhou

Abstract

Many low-resource languages still lack high-quality Automatic Speech Recognition (ASR) systems, limiting access to information, education, and AI-powered applications. While text-to-speech (TTS) data augmentation has emerged as a promising solution, existing approaches face two critical limitations: (*i*) low-diversity, single-speaker synthesis restricts accent and prosody coverage, leading to Word Error Rate (WER) plateaus; and (*ii*) the absence of principled guidance on dataset preparation, synthetic-to-real data mixing, and training pipeline design makes current solutions inefficient and non-generalizable.

To address these challenges, we propose a unified framework leveraging **Flow Matching-based TTS** for scalable low-resource ASR augmentation. Our approach significantly reduces dependence on costly human-recorded annotations by synthesizing high-quality, accent-diverse speech using minimal reference data. Furthermore, it demonstrates strong cross-lingual transfer, enabling efficient adaptation from high-resource languages to unseen low-resource languages with only limited in-domain audio. We also introduce a mathematical model that predicts WER as a function of real and synthetic data volumes. Validated on two typologically distinct languages—Spanish and Vietnamese—this model shows strong predictive power and broad generalizability.

Applying our framework to Vietnamese, we achieve **state-of-the-art performance** on **Common Voice** (6.55% WER), competitive results on GigaSpeech2 (10.22% WER), and FLEURS (11.41% WER), matching or surpassing industrial-scale systems such as Whisper Large-v3. These findings demonstrate that our framework enables effective, scalable ASR training for underserved languages and provides a principled strategy for optimizing data augmentation at scale.

Keywords: Automatic speech recognition; text-to-speech; low-resource language;

Contents

1 Introduction

Technological progress has deepened the global "digital divide" between developed and underdeveloped regions, particularly in access to speech technologies that power modern AI systems and applications. While high-resource languages such as English enjoy near-perfect automatic speech recognition (ASR) performance, low-resource languages remain severely underserved, limiting equitable access to education, information, and AI-driven tools. For English, ASR systems have achieved word error rates (WER) as low as 1% [1]. However, despite the existence of over 7,000 living languages worldwide [2], WER for many low-resource languages remains as high as 30–50% [3], further exacerbating accessibility gaps.

A major cause of this disparity is the scarcity of high-quality, annotated audio datasets, as collecting and labeling speech data is expensive and time-consuming. One promising direction explored by prior work [4]–[7] is **text-to-speech (TTS) augmentation**, where synthetic audio is generated from text using a TTS system and combined with real speech to train ASR models. While they achieved modest gains, current approaches face three key challenges: **(C1) Insufficient diversity in synthetic augmentation:** Most existing approaches rely on single-speaker or low-variation synthesis, limiting accent, prosody, and speaking-style coverage and resulting in synthetic datasets that fail to generalize. **(C2) Limited scalability of prior methods:** Even with modern generative TTS, prior studies report WER gains plateauing at synthetic-to-real ratios of only ∼1.35:1 [8]. Beyond this threshold, additional synthetic data yields diminishing returns, indicating the need for more linguistically and acoustically diverse synthesis methods. **(C3) Lack of principled strategies for augmented training:** Existing research offers little guidance on how much synthetic data to generate or how best to combine it with real speech. In the absence of predictive models, practitioners must rely on ad-hoc experimentation, which is inefficient, dataset-specific, and difficult to generalize across languages.

In this paper, we propose a unified framework to address these challenges:

1. **Flow Matching-based TTS for ASR augmentation.** To tackle **C1**, we integrate flow matching-based TTS models—a recent generative modeling technique that improves both synthesis quality and accent diversity—into ASR training pipelines for low-resource languages. Our approach enables *one-shot voice cloning*: with only ∼10 seconds of reference audio, we can generate natural speech in any target voice with diverse accents and prosodic styles. Furthermore, these models exhibit strong *cross-lingual transferability*: a system pretrained on millions of hours of high-resource language data can be efficiently adapted to a new language using only dozens [9] to a few hundred hours of in-language recordings.

2. **Substantial improvements in WER scalability.** Addressing **C2**, we demonstrate *continuous WER reductions* even at high synthetic-to-real ratios (4:1 or higher). For Vietnamese ASR, we achieve relative WER reductions of 37.6%, 27.0%, and 34.9% on the Common Voice [10], FLEURS [11], and GigaSpeech2 [12] benchmarks, respectively. Under certain configurations, diversity introduced by synthetic data yielded even more dramatic

improvements, such as a relative 78.1% WER reduction when using attention decoding without rescoring.

3. **A predictive mathematical model for data augmentation.** To solve **C3**, we propose a simple parametric model that predicts WER as a function of real and synthetic data volumes. Validated on Spanish and Vietnamese—two typologically distinct languages—our model achieves R^2 scores of 99.2% and 98.1%, enabling principled, scalable strategies for optimizing synthetic-to-real data ratios during ASR training.

4. **Systematic ablation studies of synthesis-critical factors.** Through controlled experiments, we quantify the impact of speaker diversity and text distribution on ASR performance. Results show that flow matching-based TTS remains robust even when seed speaker diversity is reduced from $2,293$ to 400 identities, but recognition accuracy is highly sensitive to the domain relevance of text used for synthesis.

5. **State-of-the-art results on Vietnamese benchmarks.** Applying our framework to Vietnamese yields a new state-of-the-art WER on Common Voice (6.55%), competitive performance on GigaSpeech2 (10.22%), and near-SOTA results on FLEURS (11.41%), surpassing or matching several industrial-scale systems such as Whisper Large-v3.

Contributions. This work advances low-resource ASR in three key ways: (1) We are the first to integrate *flow matching*-based TTS for scalable, accent-diverse data augmentation, enabling strong cross-lingual adaptation with minimal target-language data. (2) We introduce a simple yet effective predictive model for optimizing synthetic-to-real ratios, validated on typologically distinct languages. (3) We provide the most comprehensive evaluation to date of ASR performance under high synthetic-to-real ratios (up to 6:1), achieving new state-of-the-art results on Vietnamese benchmarks.

The remainder of this paper is structured as follows: Section 2 reviews prior work on multilingual ASR and data augmentation. Section 3 details our methodology, including model architectures and training pipelines. Section 4 introduces both real and synthetic datasets we use. Section 5 presents experimental results and analysis, and Section 6 concludes with key findings and future directions.

2 Related Work

Research on low-resource ASR has progressed along two complementary axes: (i) reducing reliance on annotated speech audio via *unsupervised or cross-lingual learning*, and (ii) *expanding effective training data* through augmentation—notably with synthetic speech.

Large-scale cross-lingual learning. Recent advances, particularly from industry, leverage large-scale unsupervised and cross-lingual learning to reduce target-language labeling requirements. Representative systems include *Whisper* [3], which scales weakly supervised training to

5M hours (1M labeled and 4M pseudo-labeled) and achieves strong zero-shot transfer across benchmarks; Google's *USM* [13], pretrained on ∼12M hours spanning 300+ languages with efficient fine-tuning using comparatively modest labeled data; *Universal-1* [14], trained on ∼12.5M hours across multiple languages; and NVIDIA's *Canary*, trained on the ∼1M hours, 25-language *Granary* dataset[15]. These systems reflect a broader trend toward massive multilingual pretraining. However, despite strong performance, their large model sizes and computational demands limit deployment in resource-constrained, on-device, or low-latency scenarios. Furthermore, performance on low-resource languages remains substantially weaker, highlighting the need for complementary approaches.

Unsupervised and self-supervised learning. Midsize speech corpora such as *YODAS* [16] and *GigaSpeech2* [12] facilitate scalable self-supervised and semi-supervised pipelines by offering extensive unlabeled and partially labeled audio. For instance, YODAS provides over 500k hours of multilingual speech across more than 100 languages, with both labeled and unlabeled subsets, while GigaSpeech2 contains 30K hours of automatically transcribed speech in languages such as Thai, Indonesian, and Vietnamese, collected from YouTube. The inherent heterogeneity of these datasets introduces persistent background noise and non-standard speech, even after filtering. Moreover, automated transcription errors introduce label noise, reducing the reliability of supervised signals. For Vietnamese, [17] demonstrates that pretraining on 73K hours of unlabeled data followed by fine-tuning on only 50 hours of labeled speech achieves state-of-the-art performance, underscoring the efficacy of unsupervised learning with large raw corpora. Nevertheless, such approaches remain ineffective for low-resource languages for which large raw datasets are unavailable.

Earlier TTS for data augmentation. Advances in TTS technology have enabled researchers to use synthetic audio for generating automatically labeled data. Previous studies have consistently reported reductions in WER by incorporating TTS-generated training samples. However, obtaining high-quality TTS models for low-resource languages remains a challenge. Early efforts demonstrated ASR improvements only on very small datasets with high baseline WER [4]–[7]. For instance, experiments on West Germanic minority languages [5] used only 168 minutes of synthetic and 24 minutes of real speech—a scale too limited to support broad conclusions. Similarly, [18] used only 99 hours each of real and synthetic data. A larger-scale study[6] observed a minor WER reduction from 8.66% to 7.29% using 480 hours of real data augmented with 1150 hours of synthetic speech.

Generative TTS for data augmentation. Recent advances in generative speech synthesis have significantly mitigated data scarcity in ASR training. Early efforts primarily utilized autoregressive (AR) models based on transformers, such as WaveNet [19], Tacotron [20], Tacotron 2 [21], and FastSpeech [22]. While these achieve high quality, they are limited by slow inference. This has prompted a shift toward non-autoregressive (NAR) methods, particularly diffusion-based and flow-matching TTS models—including Glow-TTS [23], Grad-TTS [24], Seed-TTS [25], and F5-TTS [26]. Such approaches offer notable improvements in naturalness, controllability, and

efficiency, enabling the generation of high-fidelity speech with wide variations in timbre, accent, and prosody across diverse speakers. As a result, synthetic corpora generated by modern TTS systems can now approach the perceptual quality of human-recorded datasets [27]. Empirical studies [8] further demonstrate that ASR models trained mainly on synthetic speech can perform nearly on par with those trained solely on real data, especially when augmented with a small amount of human recordings. However, performance gains exhibit diminishing returns as the volume of synthetic data increases (with a synthetic-to-real ratio up to 1.35:1), highlighting the importance of maximizing both linguistic and acoustic diversity in dataset construction.

Our work builds on these trends in three ways. First, we employ a flow matching-based TTS model, which improves sample quality and training stability while enabling accent-diverse, cross-lingual synthesis from limited in-language data [28]. Second, we introduce a quantitative model relating WER to the volumes of real and synthetic audio, providing a principled way to optimize data mixing and complementing prior empirical scaling studies. Finally, we conduct a comprehensive evaluation of ASR performance under large synthetic-to-real ratios—up to 6:1—on mid-sized datasets such as the 500-hour Common Voice Spanish corpus, demonstrating effectiveness at scales not explored in prior work.

3 Flow-Matching TTS Augmented ASR

3.1 Overall Methodology

Goal. Our objective is to train high-accuracy ASR for a target low-resource language while minimizing human annotation. We achieve this by:
(i) Curating a modest real-speech corpus;
(ii) Fine-tuning a *flow-matching* TTS to the target language with limited in-language audio;
(iii) Synthesizing large, *diverse* multi-speaker speech from a broad-coverage text pool; and
(iv) Mixing real and synthetic speech under a principled schedule to maximize downstream WER gains while avoiding domain leakage and overfitting to TTS artifacts.

Notation. Let $\mathcal{D}_{\text{real}} = \{(x_i, y_i)\}$ denote real audio–text pairs for training; $\mathcal{T}$ a large text pool for the target language; $\mathcal{S} = \{s_k\}$ a "seed" speaker set where each s_k is an 8–12s reference clip with text; and $\mathcal{M}_{\text{TTS}}$ a flow-matching TTS adapted to the target language from a multilingual base. We generate a synthetic set $\mathcal{D}_{\text{syn}} = \{(\tilde{x}_j, \tilde{y}_j)\}$ by conditioning $\mathcal{M}_{\text{TTS}}$ on $(\tilde{y}_j, s_{k(j)})$ with controlled variations (prosody, pace, SNR). The ASR encoder–decoder $\mathcal{M}_{\text{ASR}}$ is trained on $\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{syn}}$ with a curriculum on the synthetic:real ratio.

Stage A — Curate real data and *hold-out* sets. We assemble $\mathcal{D}_{\text{real}}$ from publicly available corpora in the target language, applying text normalization, deduplication, and strict train/dev/test partitioning. To prevent leakage, we *blocklist* test transcripts and their near-duplicates–identified via n-gram Jaccard or fuzzy matching–from any text used for TTS synthe-

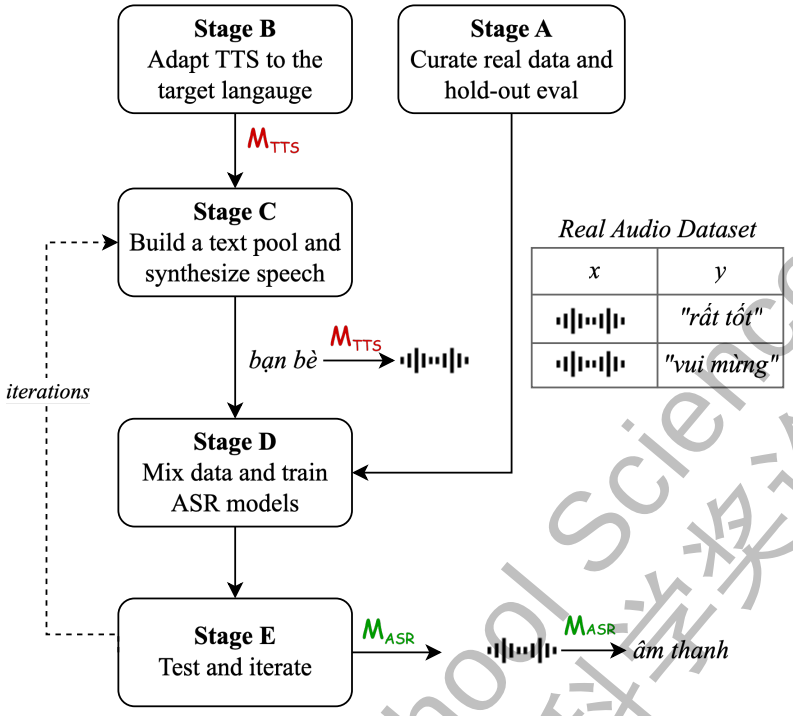

Figure 1: **Flow-Matching TTS Augmented ASR Pipeline.** Overview of the six-stage methodology: (**A**) curate real data and prepare hold-out sets, **B**) adapt a flow-matching TTS model for the target language, (**C**) synthesize diverse speech from a curated text pool, (**D**) mix real and synthetic data under a scheduled ratio, and (**E**) train the Conformer-based ASR model with hybrid CTC/Attention losses.

sis. Acoustic features, vocabulary construction, and the ASR architecture detailed are described in Section 3.2.

Stage B — Adapt a flow-matching TTS to the target language. We begin with a high-capacity, multilingual *flow-matching* TTS model (e.g., the F5-TTS family [26]) and perform parameter-efficient fine-tuning to adapt it to the target language using only a few hundred hours of in-language audio. For experiments, we use two pre-finetuned models for Spanish [29] and Vietnamese [30]. We then construct a *seed speaker bank* $\mathcal{S}$ comprising thousands of 8–12 seconds reference clips spanning diverse accents, regional variations, and recording conditions, sourced from public corpora and carefully curated web materials. The adapted $\mathcal{M}_{\text{TTS}}$ supports: (i) *multi-speaker* synthesis via speaker embeddings derived from $\mathcal{S}$, (ii) controllable prosody and speaking rate, and (iii) robust handling of out-of-vocabulary tokens through grapheme-to-phoneme backoff and character-level conditioning.

Stage C — Text pool and diverse synthesis. We compile $\mathcal{T}$ from heterogeneous sources– including conversational text, subtitles, news articles, novels, and technical documents – aggre-

gating from datasets like WikiNews, WikiSource and Open Subtitles. To ensure both quality and broad linguistic coverage, we apply filtering and stratification based on utterance length (∼2–20 s), punctuation and numeral normalization, domain diversity, and difficulty level. For each sampled $\tilde{y}_j \in \mathcal{T}$, we select a speaker $s_{k(j)} \in \mathcal{S}$ through balanced sampling across attributes such as gender, age, and geographic region, and synthesize $(\tilde{x}_j, \tilde{y}_j)$ with randomized prosodic controls to better capture natural speaker variability. Artificial room impulse responses or additive noise are not introduced, as Stage A recordings already encompass diverse acoustic environments.

Stage D — Mixing policy and training schedule. We train the Conformer-based CTC/Attention ASR model on mixed datasets with a *scheduled* proportion of synthetic data. Let $|\mathcal{D}_{\text{real}}|$ and $|\mathcal{D}_{\text{syn}}|$ represent the amount (in hours) of real and synthetic audio, respectively, and let ρ denote the synthetic-to-real ratio:

$$\rho = \frac{|\mathcal{D}_{\text{syn}}|}{|\mathcal{D}_{\text{real}}|}.$$

Given a midsize real dataset comprising $|\mathcal{D}_{\text{real}}|$ hours of audio, starting from $\rho = 0$, we progressively increase ρ until WER improvements plateau at $|\mathcal{D}_{\text{syn}}| = \rho_{\max} \cdot |\mathcal{D}_{\text{real}}|$. We then randomly sample five subsets from $\mathcal{D}_{\text{real}}$, each containing a portion of the audio:

$$\mathbf{D}_{\text{real}} = \left\{0,\ \frac{1}{5}\mathcal{D}_{\text{real}},\ \frac{2}{5}\mathcal{D}_{\text{real}},\ \frac{3}{5}\mathcal{D}_{\text{real}},\ \frac{4}{5}\mathcal{D}_{\text{real}}\right\}.$$

For each $\mathcal{D}_{\text{real}(i)} \in \mathbf{D}$, $|\mathcal{D}_{\text{syn}(i,j)}| = \rho_j \cdot |\mathcal{D}_{\text{real}(i)}|$ $(\rho_i = 0, 1, ..., \rho_{\max})$ hours of synthetic audio are generated accordingly, and ASR models are trained on mixed sets $\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{syn}}$.

Stage E — Losses, decoding, and rebalancing. Training follows the hybrid CTC/attention objective (Section 3.2). During development, we evaluate four decoders (attention-only, attention-rescoring, CTC greedy, CTC prefix beam). If WER gains saturate, we (i) increase text/domain diversity in $\mathcal{T}$, (ii) rebalance speaker strata with underrepresented accents, or (iii) modestly *lower* $\rho_{\max}$ to mitigate TTS-overfitting. This *data-first* loop complements model-side tuning.

We ensure a few pratical safeguards as below. (i) *Leakage control*: Exclude any development/test text and enforce near-duplicate filtering. (ii) *Domain balance*: Maintain at least 20–30% in-domain text while retaining diverse out-of-domain data for generalization. (iii) *Reproducibility*: Resample speaker–text assignments across synthetic datasets and retrain models for variance estimation.

3.2 End-to-End ASR using Conformer and Hybrid CTC/Attention

We first extracted 80-channel filter banks features from audio recordings, computed using a 25ms window and a 10ms stride. We applied global Cepstral Mean and Variance Normalization (CMVN) to normalize the features, enhancing the performance of the ASR system under varying

recording conditions. CMVN mitigates channel/microphone effects and compensates for speaker volume variations, while normalized distribution also contributes to more stable gradients during model optimization.

Utilizing Byte Pair Encoding (BPE) [31], we decompose words in audio transcripts into recombinable subword units. This approach enables direct subword output in our end-to-end ASR system, bypassing the alignment complexities inherent in phoneme-to-word conversion. Although a shared multilingual BPE vocabulary is feasible, we deliberately construct language-specific vocabularies for distinct ASR tasks (e.g., Spanish and Vietnamese) to isolate experimental variables. Each BPE vocabulary contains 4,000-5,000 subword tokens.

Extracted audio features and tokenized transcripts are sent to an end-to-end ASR system comprising an encoder and a decoder module. The encoder utilizes a standard Conformer [32] architecture, while the decoder employs a hybrid CTC-attention loss function [33] during training.

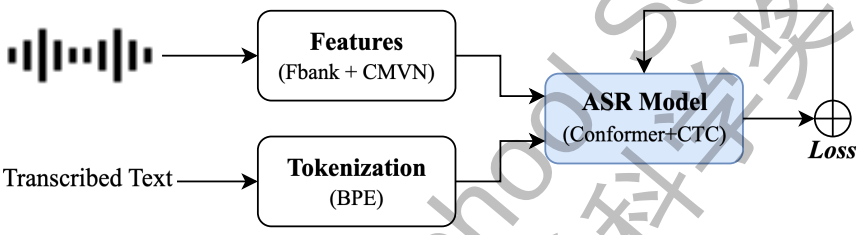

Figure 2: Training Flow of a Conformer-based ASR Model

Following the original Conformer, the encoder consists of 12 Conformer blocks, each containing an 8-head self-attention module, a convolution module, and a feed-forward module with Swish activation.

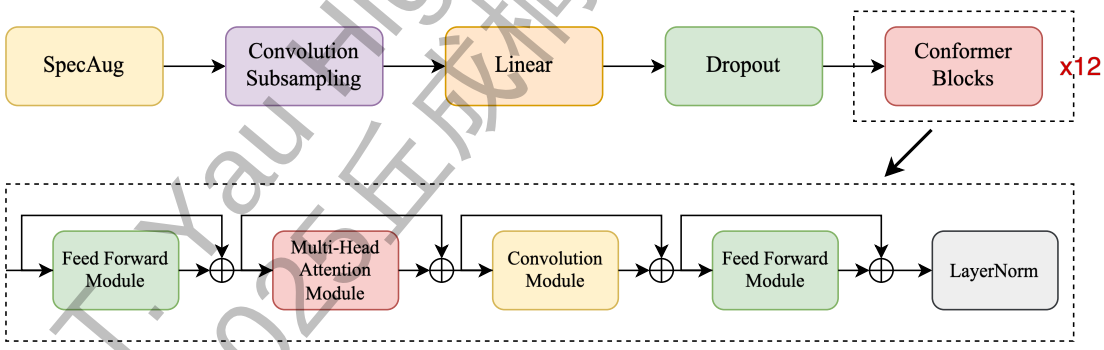

Figure 3: Conformer Architecture

The hybrid CTC-attention loss is employed in decoder stage during training, leveraging the CTC module's faster convergence while benefiting from the attention module's enhanced robust-

ness. The combined loss function is defined as below, where λ is set as 0.3.

$$L_{\text{combined}} = \lambda L_{\text{CTC}} + (1 - \lambda) L_{\text{attention}}$$

When applying the trained model to speech recognition, multiple key strategies are available for hybrid CTC-attention decoding: attention-only decoding, attention rescoring, CTC greedy decoding, and CTC prefix beam search. We conducted comparative analysis of these approaches, with experimental details provided in Section 5.

3.3 Flow Matching-based Text-to-Speech Engine

We employ the F5-TTS synthesis engine [26] to generate transcribed synthetic audio. F5-TTS is a non-autoregressive text-to-speech system utilizing flow matching with a Diffusion Transformer architecture. The base F5-TTS model was trained on 95,000 hours of carefully filtered English and Chinese speech data derived from the multilingual Emilia dataset.

Although the base F5-TTS model exclusively supports Chinese and English speech generation, such flow matching-based models learn underlying, universal representations of human speech rather than superficial language-specific rules. When sufficiently pretrained on one or two languages (e.g., English and Chinese in the base F5-TTS model), it acquires core capabilities for constructing natural speech. By fine-tuning the base F5-TTS model with a few hundreds hours of target-language audio data (e.g., Spanish or Vietnamese), high-quality text-to-speech models for new languages can be rapidly developed.

Fine-tuning essentially guides the model to adapt its acquired universal speech generation capabilities to the specific patterns of new languages. This approach is grounded in two key principles:

- Linguistic universals and shared acoustic properties: All human languages utilize identical articulatory organs (vocal cords, tongue, lips, palate, etc.) to produce speech. Consequently, phonemes – the fundamental acoustic units – and their acoustic properties exhibit substantial cross-linguistic commonality. Many phonemes are identical or highly similar across languages. The pretrained model has effectively learned to generate these shared phonemes.

- Advantages of the flow matching architecture: The model inherently learns the conditional probability distribution $p(\mathbf{x}|\mathbf{c})$ where $\mathbf{x}$ denotes the speech waveform and $\mathbf{c}$ comprises conditioning inputs (text, speaker information, etc.). Pretrained on English and Chinese corpora, the model develops precise generative control through conditional parameters (phoneme sequences, speaker embeddings). When fine-tuning for new languages, the core objective becomes accurate projection of novel phoneme sequences (including language-specific phonemes) and prosodic patterns onto the model's established universal acoustic feature space.

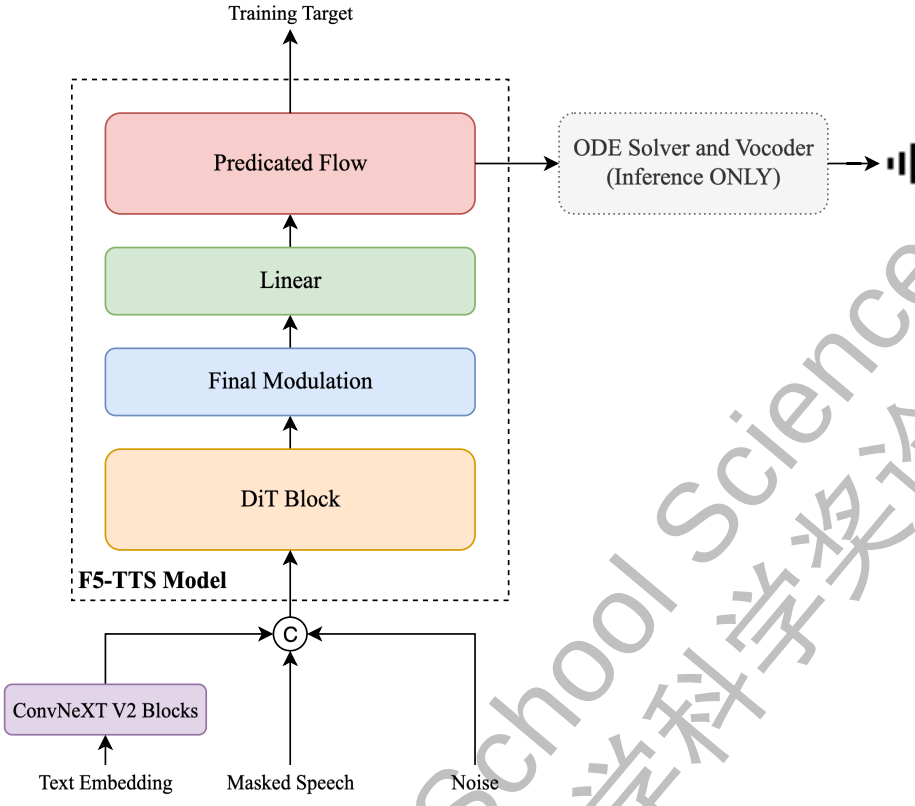

Figure 4: F5-TTS Model: Using ODE Solver and Vocoder During Inference

In cross-lingual TTS adaptation scenarios, flow matching-based systems demonstrate significant superiority over previous Diffusion Model approaches. While Diffusion Models excel at synthesizing high-prosody audio within their training language domain, flow matching enables efficient cross-lingual transfer through parameter reuse: most bottom-layer parameters encoding universal speech features remain frozen during fine-tuning. Recent research [9] shows effective adaptation of the F5-TTS base model to new languages by updating only 1.72% of total parameters. This substantially reduces adaptation costs, enabling stable training of new language TTS models with merely dozens to hundreds of hours of target-language data. Crucially, our systematic application of flow-matching's cross-lingual transferability yields substantially greater relative WER improvements in experiments (Section 5) than prior ASR augmentation studies using earlier Diffusion-based TTS engines [8].

3.4 Open Source Tools and Models

To train the model described in Section 3.1 and implement the methodology outlined in Section 3.2 to develop Spanish and Vietnamese TTS engines, we used the following open source tools.

3.4.1 WeNet

WeNet [33] is an open-source speech recognition toolkit designed with production readiness as its core principle. The framework implements a standard Transformer/Conformer architecture, incorporating multi-head attention mechanisms, positional encoding, feed-forward networks (for Transformer), and convolutional modules (in Conformer model). By leveraging this toolkit, we eliminated the need for implementing Conformer models from scratch, thereby enabling focused investigation into how blending authentic and synthetic speech data impacts recognition accuracy (i.e. word error rate).

We configure the training pipeline to employ PyTorch-implemented AdamW optimization with dynamic learning rate scheduling (linear warmup followed by cosine decay). This configuration accelerates convergence while reducing GPU memory overhead.

3.4.2 F5-TTS and its Multilingual Derivatives.

The F5-TTS training code is publicly available on GitHub [34], with pretrained base models accessible via HuggingFace [35]. Leveraging its exceptional cross-lingual adaptation capabilities (Section 3.3), numerous high-quality multilingual TTS engines have been fine-tuned from this base. For our study, we utilize two such derivatives: a Spanish TTS model (F5-Spanish [29], 218 hours audio) and a Vietnamese model (EraX-Smile-UnixSex-F5 [30], about 1000 hours audio), both adapted from the F5-TTS foundation.

To evaluate synthetic audio quality, we conducted a randomized sample assessment of 200 synthesized utterances per language. These were benchmarked against reference audio generated by public TTS services (e.g., Google Translate) using identical transcripts. Perceptual evaluation revealed better quality in Spanish model outputs compared to Vietnamese counterparts.

4 Training Datasets

4.1 Real Audio Datasets

Dataset	Language	Train (Hours)	Validation (Hours)
Common Voice	Spanish	˜500	4
Bud500	Vietnamese	˜500	˜50
LSVSC	Vietnamese	˜80	˜10
VLSP 2020	Vietnamese	˜80	˜10

Table 1: Annotated Human Speech Datasets for ASR

For Spanish ASR experiments, we use the publicly available Common Voice dataset, which contains approximately 500 hours of training data, along with 4 hours each for development and test splits. For Vietnamese, we leverage the Bud500 dataset [36], which offers a comparable

training size (˜500 hours). Bud500 covers diverse topics—including podcasts, travel, literature, and food—while capturing a wide range of accents from Vietnam's Northern, Southern, and Central regions.

To benchmark Vietnamese ASR performance against state-of-the-art (SOTA) systems, we additionally evaluate on two widely recognized public datasets, LSVSC [37] and VLSP 2020 [38]. Both are noted for their high transcription accuracy and are frequently used in the Vietnamese ASR research community. The characteristics of all real-speech datasets are summarized in Table 1.

4.2 Synthetic Datasets

The synthesis process was *Stage C* in Section 3.1. Two inputs are required:

1. A "seed set" of audio samples from real human speakers, from real speakers, each contributing 8–12 seconds of recorded speech paired with the corresponding text. For Spanish, the seed set includes 2,180 speakers; for Vietnamese, 4,808 speakers were used.

2. A large pool of target-language text, segmented into short sentences suitable for synthesis.

Dataset Type	Language	"Seed" Speakers	Synthetic Audio (Hours)
TTS	Spanish	2180	˜2500
TTS	Vietnamese	4808	˜2500

Table 2: Synthesized Speech Datasets for ASR

Spanish. For Spanish, we utilize the VoxForge Spanish Corpus [39], which contains read speech from 2,180 speakers (1,713 male, 467 female). One reference clip per speaker is selected to form the seed set. A diverse Spanish text corpus is compiled from online sources, including movie subtitles, official documents, novels, and TED transcripts. After cleaning and normalization, the corpus is segmented into short sentences averaging approximately 100 characters, resulting in ∼4M entries. Each entry is synthesized into speech using the F5-Spanish model.

Vietnamese. Due to the lack of large, publicly available datasets with sufficient speaker diversity, we collect 4,808 seed recordings (about 10 seconds each) from 186 YouTube channels. The Vietnamese text corpus is constructed similarly to the Spanish pipeline, combining diverse open-domain sources, normalized and segmented before synthesis.

For experiments in Section 5, we utilize subsets or the entirety of the synthesized speech from the datasets summarized in Table 2.

5 Experiment Results

We evaluate all models using Word Error Rate (WER) as the primary metric. All our ASR models are Conformer-based (Section 3.2) and uses one of four decoding strategies: attention-

only decoding, attention rescoring, CTC greedy decoding, and CTC prefix beam search.

5.1 A Mathematical Model for WER with Real and Synthetic Data

We investigate the relationship between WER and the ratio of real to synthetic training data using Spanish Common Voice V21 as the benchmark.

5.1.1 Effectiveness of Synthetic Data

Figure 5 shows how increasing synthetic data affects ASR performance. Starting with ∼500 hours of Common Voice Spanish (Section 4.1), we gradually augment the training set with 0.5×, 1×, 2×, 4×, and 6× synthetic audio (Section 4.2). Moderate augmentation consistently reduces WER but yields diminishing returns. Beyond a certain point, excessive synthetic data slightly degrades performance, likely due to overfitting on less diverse TTS-generated speech.

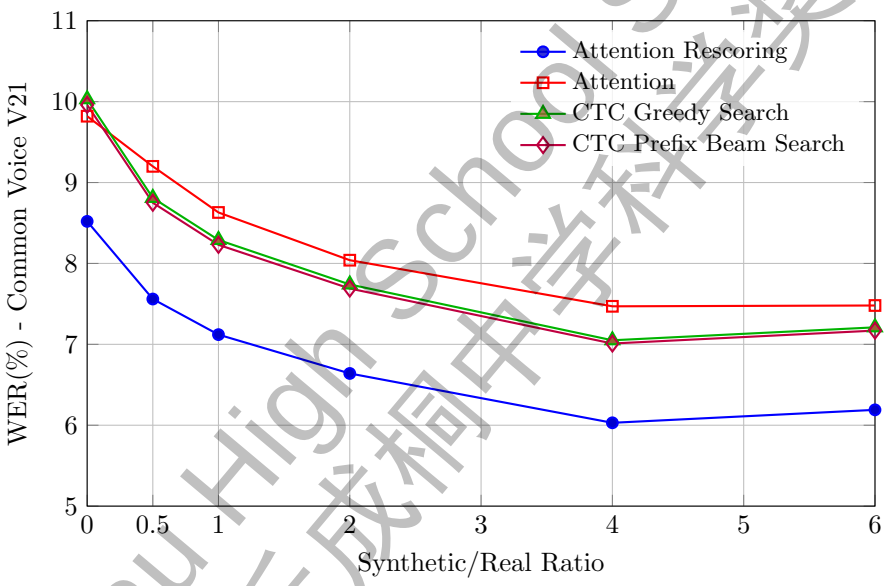

Figure 5: WER Trends When Adding Synthetic Audio to the Training Set

We further evaluate models trained exclusively on synthetic datasets of varying sizes (Figure 6). WERs are much higher compared to mixed training but still decrease as the amount of synthetic data grows.

Table 3 highlights that even when 95% of training data is synthetic and only 5% real (˜75 hours), performance improves substantially—achieving over 10 percentage points of WER reduction. This demonstrates the importance of small real-speech anchors.

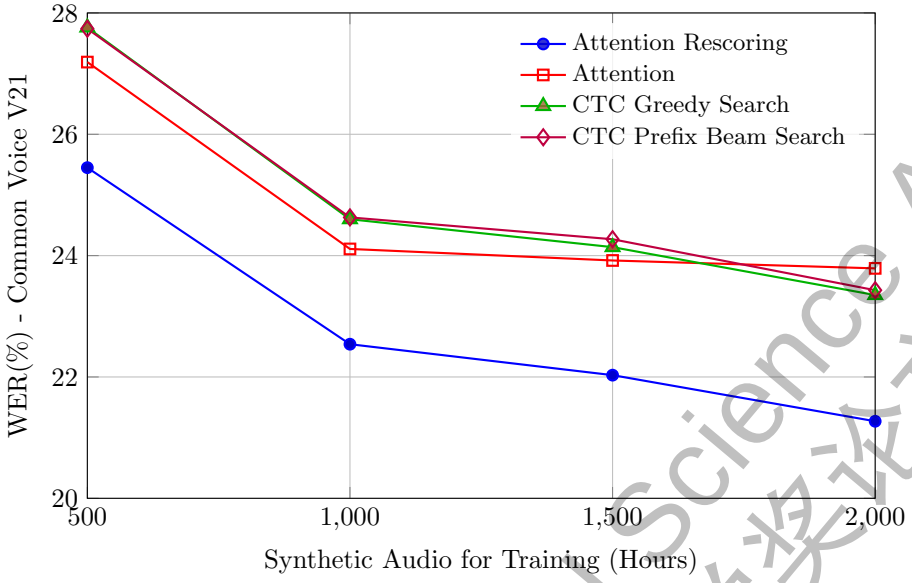

Figure 6: WER Trends When Training Exclusively on Synthetic Audio

Decoding Method	**100% Synthetic**	**5% Real + 95% Synthetic**
Attention	23.92	10.76
Attention Rescoring	22.03	9.16
CTC Greedy Search	24.14	10.51
CTC Prefix Beam Search	24.27	10.45

Table 3: Impact of Adding 5% Real Speech to Synthetic Training Sets

5.1.2 The Mathematical Model

Building on these findings, we hypothesize that WER systematically depends on the amounts of real and synthetic data used for training. We propose the following parametric function:

$$\text{WER}(X,Y) = \frac{A}{e^{EX} \cdot e^{FY}} + \frac{B}{e^{EX}} + \frac{C}{e^{FY}} + D,$$

where X and Y denote the hours of real and synthetic data. Model parameters A to F are optimized via least squares, and fit quality is measured using R^2.

	Spanish	**Vietnamese**
Real Speech Audio	0-500 hours	0-500 hours
Synthetic Speech Audio	0-1000 hours	0-1800 hours
WER test points	35	45

The model achieves excellent predictive accuracy, with $R^2 = 99.2\%$ for Spanish and $R^2 = 98.1\%$ for Vietnamese (Table 4), suggesting robustness across typologically distinct languages.

Decoding Method	**Spanish**	**Vietnamese**
Attention	0.992	0.975
Attention Rescoring	0.991	0.981
CTC Greedy Search	0.991	0.975
CTC Prefix Beam Search	0.991	0.969

Table 4: R^2: the Goodness-of-Fit of WER(X, Y)

Figure 7 and Figure 8 show the matching between the surfaces of the mathematical model (blue and green) and the test points (dots in red and orange).

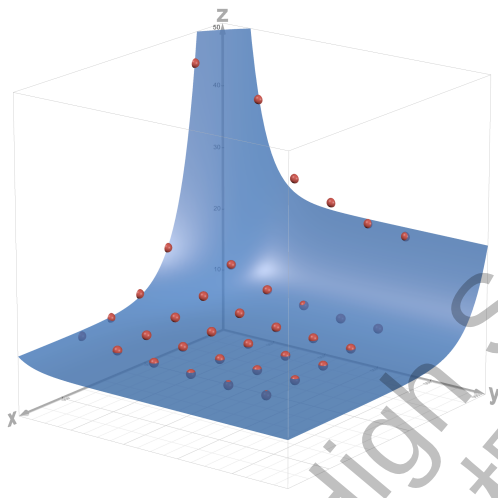

Figure 7: Goodness-of-Fit: Spanish

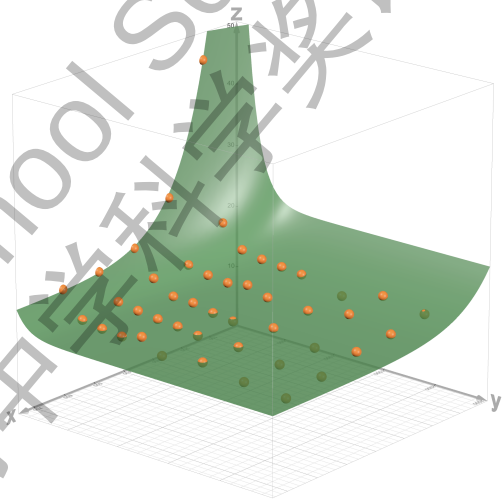

Figure 8: Goodness-of-Fit: Vietnamese

5.2 Ablation Study: Impact of Synthesis Critical Factors

We analyze two synthesis-related factors that affect ASR performance: (i) the number of distinct seed speakers used for TTS, and (ii) the distribution of text used for generating synthetic audio.

5.2.1 Effect of Speaker Diversity

We train models on 500 hours of real speech and 1000 hours of synthetic audio, varying seed speakers from 400 to 2293 (Table 5).

Results demonstrate that even after substantially reducing the number of distinct speaker vocal identities used for synthesis from 2,293 to just 400, the impact on Automatic Speech Recognition (ASR) performance remains remarkably minimal. Although a certain amount of

Speaker#	Attention Rescoring	Attention	CTC Greedy Search	CTC Prefix Beam Search
2293	6.64	8.04	7.74	7.69
1600	6.50	8.04	7.58	7.57
1200	6.52	7.75	7.59	7.55
800	6.56	7.82	7.65	7.60
400	6.58	8.18	7.62	7.60

Table 5: ASR Performance (WER) Comparison with Varying Numbers of "Seed Speakers"

speaker variety is essential, this finding also suggests that in practical applications of ASR for low-resource languages, it is feasible to collect seed audio from an *acceptably* small number of speakers—each contributing as little as around 10 seconds of recording—and subsequently use a TTS model to generate large quantities of synthetic speech for ASR training. Moreover, seed recordings can be conveniently sourced from publicly available materials such as films, TV programs, and online video platforms, reducing data acquisition costs.

5.2.2 Effects of Text Distribution

We examine four text sources—TED Talks, subtitles, books/news, and Common Voice transcripts—and synthesize 500 hours from each. When combined with 500 hours of real speech, in-domain text consistently yields the best results (Table 6).

Dataset	Attention Rescoring	Attention	CTC Greedy Search	CTC Prefix Beam Search
CV+ted	7.68	9.19	8.89	8.84
CV+subtitle	7.67	9.17	9.04	9.01
CV+book	7.74	8.81	8.91	8.88
CV+cv	**7.12**	**8.63**	**8.29**	**8.23**

Table 6: ASR Performance (WER) with Different Text Distributions for Synthetic Data

The results confirm that the choice of text distribution significantly affects model accuracy. Synthetic audio derived from in-domain transcripts (CV+CV) yields the lowest WER, improving recognition accuracy by up to 0.6 absolute points compared to out-of-domain text. We attribute this to improved coverage of domain-specific vocabulary and phonetic patterns, enhancing the model's generalization and robustness to target test sets.

5.2.3 Key Insights

This ablation study highlights two practical findings:

- **Speaker diversity is less critical:** For flow-matching TTS; a small number of seed speakers suffices for generating large, effective synthetic datasets.

- **Text distribution dominates:** Using in-domain transcripts for synthesis substantially boosts ASR performance.

Together, these results suggest that low-resource ASR systems can be built efficiently by focusing on domain-relevant text while keeping seed speaker requirements minimal, improving accessibility for underrepresented languages.

5.3 Application on a Low-Resource Language (Vietnamese)

Model	Parameter (M)	Human Speech (Hours)	Common Voice	FLEURS	Giga Speech2
Whisper large-v3 (OpenAI)	1542	1M+4M[1]	13.74	**8.59**	17.94
Whisper base (OpenAI)	72	-	44.07	40.41	39.88
MMS L1107 (Meta AI)	964	49K+55K[2]	43.88	55.35	46.62
GigaSpeech2 small	68	6039	18.81	13.50	14.72
GigaSpeech2 large	152	6039	14.43	11.59	12.83
Google USM	-	-	12.46	11.75	13.38
Azure Speech CLI 1.37.9	-	-	10.21	11.88	11.78
Ours					
Bud500+0H	121	500	16.37	18.72	16.04
Bud500+2000H	121	500	10.20	13.65	10.44
Bud500+LSVSC+VLSP	121	660	9.22	13.57	11.57
Bud500+LSVSC+VLSP+2000H	121	660	**6.55**	11.41	**10.22**

Table 7: WER(%) Comparison on Three Benchmarks

Building on the findings in Section 5.1.1, which demonstrated that flow matching-based TTS synthesis effectively improve Spanish ASR performance even at a 4:1 synthetic-to-real ratio, we applied the same data generation and training methodology to Vietnamese.

Our approach is evaluated on three widely adopted benchmarks: Common Voice, FLEURS, and GigaSpeech2. Table 7 presents the WER results of our models alongside several existing state-of-the-art systems. During the decoding process, attention rescoring was applied to compute the WER values reported in Table 7. All our models are based on the same Conformer architecture and hyperparameter configuration as detailed in Section 3.2, and differ only in the composition of training data.

We begin with the 500-hour Bud500 Vietnamese dataset. Adding 2000 hours of synthetic audio (Bud500+2000H) yields significant WER reductions but exhibits diminishing returns beyond this point. To further improve performance, we incorporate two additional public Vietnamese datasets, LSVSC and VLSP, resulting in a combined 660-hour real dataset. Training with this expanded dataset plus 2000 hours of synthetic audio (Bud500+LSVSC+VLSP+2000H) achieves

[1] 1M labeled and 4M pseudo-labeled
[2] 49K labeled and 55K unlabeled

state-of-the-art WER performance on Common Voice (6.55%), competitive results on GigaSpeech2 (10.22%) and FLEURS (11.41%). For comparison, we also report results for the same 660-hour combined real dataset without synthetic augmentation (Bud500+LSVSC+VLSP).

We note that the FLEURS dataset [11] primarily consists of Wikipedia-based audio-text pairs, while our real and synthetic datasets are mostly colloquial in nature. This domain mismatch likely contributes to the relative performance gap observed on FLEURS.

As shown in Table 7, the ASR model trained on the real Bud500 dataset achieves a relatively high yet acceptable WER when using attention rescoring for decoding. However, experimental results indicate that the same model exhibits a near-complete performance breakdown on the FLEURS and GigaSpeech2 datasets when decoded with attention only (Table 8). In contrast, the model trained on the augmented dataset Bud500+2000H (which includes synthetic data) demonstrates more stable WER performance across different decoding strategies, including Attention Only.

Dataset	**Common Voice**	**FLEURS**	**GigaSpeech2**
Bud500+0H	21.23	58.55	47.74
Bud500+2000H	12.13	19.37	14.28

Table 8: WER(%) of ASR Models using Attention Only Decoding

6 Conclusion

In this work, we presented a **flow matching-based Text-to-Speech (TTS) data augmentation framework** for improving low-resource Automatic Speech Recognition (ASR). By leveraging the cross-lingual transferability and high-fidelity synthesis capabilities of *flow-matching generative models* (e.g., F5-TTS), our approach enables the creation of *diverse, multi-speaker, accent-rich synthetic corpora* from limited reference audio. This reduces dependence on expensive human annotations while substantially boosting ASR performance in low-resource settings.

Extensive experiments on **Spanish** and **Vietnamese** demonstrate three key findings:

1. **High-quality synthetic audio drives consistent WER gains** up to synthetic-to-real ratios of 4:1–6:1, outperforming prior TTS-based augmentation studies that plateau near 1.35:1.

2. Our proposed **WER prediction model** captures the quantitative relationship between training data composition and recognition accuracy, achieving $\mathbf{R^2 \geq 0.98}$ across typologically distinct languages. This provides a principled tool for optimizing augmentation strategies in multilingual ASR.

3. Through ablation studies, we show that while speaker diversity has limited impact, **textual coverage and domain matching** are critical for downstream performance, enabling

tailored augmentation for domain-specific applications.

Applied to **Vietnamese ASR**, our framework achieves **state-of-the-art WER performance** on Common Voice (6.55%), while delivering competitive performance on FLEURS (11.41%) and GigaSpeech2 (10.22%) with industrial systems like Whisper Large-v3, despite using only a fraction of the training data and computational resources.

Looking forward, this methodology opens avenues for developing *scalable and inclusive ASR systems*. Future work will explore:

- Extending cross-lingual adaptation to enable truly *zero-shot TTS* for languages without any labeled audio,
- Integrating *synthetic speech quality estimation* for automated data filtering, and
- Combining TTS augmentation with *large-scale self-supervised pretraining* to maximize gains in extremely low-resource scenarios.

By unifying parameter-efficient flow-matching TTS with structured data augmentation, our framework demonstrates a **cost-effective, reproducible, and language-agnostic** pathway for advancing ASR in underrepresented languages worldwide.

7 Acknowledgment

My interest in automatic speech recognition (ASR) began during childhood trips to English-speaking countries, where I discovered the "cc" (closed caption) button on TV remotes. Pressing it revealed real-time subtitles—something I found fascinating. At the time, I imagined incredibly fast typists behind the scenes, only later learning that most captions were powered by ASR. Years later, when visiting countries like Vietnam and Thailand, I noticed that their TV remotes lacked this feature. That missing "cc" button stayed with me and ultimately inspired my decision to focus on Vietnamese ASR in this project.

I first met Prof. Xu at a Berkeley alumni event, where he introduced me to computing courses and encouraged me to explore research topics. Over time, he became a generous mentor—recommending papers, answering my questions, and guiding me toward deeper inquiry. Last fall, he showed me the paper "*F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching* [26]", which sparked a turning point in this project. I was amazed when I saw a model replicate my voice from just five seconds of audio. This experience led me to explore text-to-speech (TTS) models more deeply and to think about using them creatively rather than studying AI safety.

When I discovered the HuggingFace repository of multilingual TTS models, the connection became clear: if high-quality synthetic speech could be generated in almost any language, perhaps it could be used to improve ASR performance in low-resource settings. That insight became the foundation of this research.

Over six months, I worked independently to design experiments, train models, and evaluate results. Prof. Xu provided invaluable mentorship, from helping me refine research questions to providing access to an NVIDIA RTX 3090 GPU cluster. He taught me academic writing, experimental design, and the importance of clear result presentation. My mathematical modeling of WER and data volume was inspired by my participation in the RISE program at Boston University, where I explored data science research methodologies.

This work would not have been possible without the open-source community. The WeNet project provided an accessible framework for ASR training, and HuggingFace offered pre-trained TTS models that enabled rapid experimentation without requiring extensive hardware resources. These tools allowed me to focus on ideas rather than infrastructure.

Most importantly, this project deepened my appreciation for research as a creative process—combining curiosity, experimentation, and community knowledge. While I still have much to learn about the theory behind these tools, this experience has strengthened my desire to pursue computer science and contribute to open-source projects in the future.

Finally, I extend my deepest gratitude to Prof. Xu for his patient guidance, encouragement, and generosity. In many ways, this work represents my attempt to bring the missing "cc" button on TV remotes closer to reality for low-resource languages.

References

[1] H. Audio. "Open ASR Leaderboard." Accessed: May 21, 2024, Hugging Face. (2024), [Online]. Available: `https://huggingface.co/spaces/hf-audio/open_asr_leaderboard`.

[2] D. M. Eberhard, G. F. Simons, and C. D. Fennig, Eds., *Ethnologue: Languages of the World*, 28th. SIL International, 2025. [Online]. Available: `https://www.ethnologue.com`.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[4] A. Laptev, V. Chernykh, B. Nechaev, E. Ryumina, A. Silaev, and I. Medennikov, "You Do Not Need More Data: Improving End-to-End Speech Recognition by Text-To-Speech Data Augmentation," in *13th International Conference on Signal Processing and Communication Systems (CISP-BMEI)*, 2020.

[5] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, and M. Wieling, "Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

[6] A. Fazel, W. Yang, Y. Liu, *et al.*, "SynthASR: Unlocking Synthetic Data for Speech Recognition," in *Proceedings of INTERSPEECH*, 2021.

[7] C.-T. Do, S. Imai, R. Doddipatla, and T. Hain, "Improving Accented Speech Recognition using Data Augmentation based on Unsupervised Text-to-Speech Synthesis," in *Proceedings of the 32nd European Signal Processing Conference (EUSIPCO)*, 2024.

[8] G. Yang, F. Yu, Z. Ma, *et al.*, "Enhancing Low-Resource ASR through Versatile TTS: Bridging the Data Gap," 2024. [Online]. Available: https://arxiv.org/abs/2410.16726.

[9] K.-J. Kwon, J.-H. So, and S.-H. Lee, "Parameter-Efficient Fine-Tuning for Low-Resource Text-to-Speech via Cross-Lingual Continual Learning," in *Proceedings of INTERSPEECH*, 2025.

[10] R. Ardila, M. Branson, K. Davis, *et al.*, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020.

[11] A. Conneau, M. Ma, S. Khanuja, *et al.*, "FLEURS: Few-Shot Learning Evaluation of Universal Representations of Speech," in *Proceedings of IEEE Spoken Language Technology Workshop*, 2022.

[12] Y. Yang, Z. Song, J. Zhuo, *et al.*, "GigaSpeech 2: An Evolving, Large-Scale and Multi-domain ASR Corpus for Low-Resource Languages with Automated Crawling, Transcription and Refinement," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[13] Y. Zhang, W. Han, J. Qin, *et al.*, "Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages," 2023. [Online]. Available: https://arxiv.org/abs/2303.01037.

[14] F. M. Ramirez, L. Chkhetiani, A. Ehrenberg, *et al.*, "Anatomy of Industrial Scale Multilingual ASR," *CoRR*, 2024. [Online]. Available: https://arxiv.org/abs/2404.09841.

[15] N. R. Koluguri, M. Sekoyan, G. Zelenfroynd, *et al.*, "Granary: Speech Recognition and Translation Dataset in 25 European Languages," 2025. [Online]. Available: https://arxiv.org/abs/2505.13404.

[16] X. Li, S. Takamichi, T. Saeki, W. Chen, S. Shiota, and S. Watanabe, "YODAS: Youtube-oriented dataset for audio and speech," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023.

[17] J. Zhuo, Y. Yang, Y. Shao, *et al.*, "VietASR: Achieving Industry-level Vietnamese ASR with 50-hour labeled data and Large-Scale Speech Pretraining," 2025. [Online]. Available: https://arxiv.org/abs/2505.21527.

[18] R. Zevallos, "Text-To-Speech Data Augmentation for Low Resource Speech Recognition," 2022. [Online]. Available: https://arxiv.org/abs/2204.00291.

[19] A. Van Den Oord, S. Dieleman, H. Zen, *et al.*, "WaveNet: A Generative Model for Raw Audio," in *Proceedings of 9th ISCA Workshop on Speech Synthesis Workshop*, 2016.

[20] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of INTERSPEECH*, 2017.

[21] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural TTS synthesis by Conditioning Wavenet on Mel Spectrogram Predictions," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[22] Y. Ren, Y. Ruan, X. Tan, *et al.*, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[23] J. Kim, S. Kim, and J. Kong, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[24] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," in *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[25] P. Anastassiou, J. Chen, J. Chen, *et al.*, "Seed-TTS: A Family of High-Quality Versatile Speech Generation Models," 2024. [Online]. Available: `https://arxiv.org/abs/2406.02430`.

[26] Y. Chen, Z. Niu, Z. Ma, *et al.*, "F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching," 2024. [Online]. Available: `https://arxiv.org/abs/2410.06885`.

[27] S. Chen, C. Wang, Y. Wu, *et al.*, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, 2025.

[28] Y. Liang, W. Liu, C. Qiang, *et al.*, "Towards Flow-Matching-based TTS without Classifier-Free Guidance," 2025. [Online]. Available: `https://arxiv.org/abs/2504.20334`.

[29] *F5-Spanish: F5-TTS Spanish Language Model.* [Online]. Available: `https://huggingface.co/jpgallegoar/F5-Spanish`.

[30] *EraX-Smile-UnixSex-F5: Giving F5-TTS a Unisex Vietnamese Twist.* [Online]. Available: `https://huggingface.co/erax-ai/EraX-Smile-UnixSex-F5`.

[31] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, "Fast WordPiece Tokenization," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021.

[32] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proceedings of INTERSPEECH*, 2020.

[33] B. Zhang, D. Wu, C. Yang, *et al.*, "Wenet: Production first and production ready end-to-end speech recognition toolkit," in *Proceedings of INTERSPEECH*, 2021.

[34] *F5-TTS GitHub page.* [Online]. Available: `https://github.com/SWivid/F5-TTS`.

[35] *F5-TTS HuggingFace page.* [Online]. Available: `https://huggingface.co/SWivid/F5-TTS`.

[36] *Bud500: A Comprehensive Vietnamese ASR Dataset (HuggingFace Mirror)*. [Online]. Available: https://github.com/apluka34/Bud500.

[37] "LSVSC: Large-Scale Vietnamese Speech Corpus (HuggingFace Mirror)." (2024), [Online]. Available: https://huggingface.co/datasets/doof-ferb/LSVSC.

[38] "VLSP 2020 - VINAI 100 Hours Speech Corpus (HuggingFace Mirror)." (2024), [Online]. Available: https://huggingface.co/datasets/doof-ferb/vlsp2020_vinai_100h.

[39] *Voxforge Spanish Corpus (HuggingFace Mirror)*. [Online]. Available: https://huggingface.co/datasets/ciempiess/voxforge_spanish.