ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Adapting F5-TTS Model to Kurdish Sorani: Diffusion-based Speech Synthesis with a Specialized Dataset

Hamreen Ahmad, Aree Mohammed

Department of Computer Science, University of Sulaimani, Sulaymaniyah, Iraq

ABSTRACT

The Kurdish language is one of the low-resource languages in the field of speech synthesis. Most of the currently available Kurdish text-to-speech (TTS) systems lack both accuracy and naturalness. To address this issue, this study fine-tunes the F5-TTS model for the Kurdish (Sorani) language, an advanced model that had not previously been fine-tuned for this language. The process began with the creation of a high-quality, well-constructed, single-speaker dataset containing more than 10 h of recorded speech that was collected from news, interviews, and short videos. The dataset was very carefully curated to ensure balanced sample durations, accurate transcriptions, emotional diversity, and a wide range of topics and speaking styles. After ensuring that the training data was sufficient, clean, and prepared, the F5-TTS model was fine-tuned on this data. Both objective and subjective evaluations were conducted to verify the model's performance. For the objective evaluation, the model achieved a character error rate of 4.3% and a word error rate of 20.37%, indicating a high transcription accuracy in the generated audio. In the subjective evaluation, the mean opinion score reached 4.72, showing that the synthesized speech is very close to the original speaker's voice. These results demonstrate that diffusion-based models like F5-TTS can be effectively adapted to low-resource languages when supported by a well-designed dataset.

Index Terms: Speech Synthesis, Diffusion-Based Model, Low-Resource Language, F5-Text-to-speech, Kurdish Sorani

1. INTRODUCTION

Text-to-speech (TTS), also known as speech synthesis, is a technology that converts written words into spoken language. It plays an important role in modern technology. Most artificial intelligence chatbots, smart devices, and even organizations use such systems to provide fast, more realistic, and enjoyable service. For example, by converting written text into natural-sounding speech, this technology is mainly utilized in human-computer interaction systems to make them more user-friendly [1]–[3]. An advanced TTS system should not only read what has been written, but also be able to maintain emotion, rhythm, punctuation, and many other properties to produce speech that sounds like a real human.

However, achieving these capabilities is not equally accessible across all languages. For widely spoken languages such as English, Spanish, and Arabic, TTS technology has been improved significantly [4], while low-resource languages are neglected and outdated [5]. In the case of Kurdish, while recent studies [6]–[8] have made noticeable progress, they often show limitations in fluency and emotional expression, as indicated by their reported mean opinion score (MOS) of 4.1 [8] and 3.9 [7]. The two major reasons for this gap are: unavailability of a high-quality, domain-diverse dataset, and

Access this article online

DOI:10.21928/uhdjst.v9n2y2025.pp198-207

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Hamreen Ahmad, Department of Computer Science, University of Sulaimani, Sulaymaniyah, Iraq. Email: hamreen.ahmad@univsul.edu.iq

Received: 29-07-2025 Accepted: 21-09-2025 Published: 06-10-2025

the use of outdated TTS models that are not suitable for low-resource languages [5].

The Kurdish Sorani script is similar to that of Arabic, Persian, and other related languages. These languages share some phonetic consistency [9], [10]. Most words in Kurdish Sorani script are pronounced as they are written. However, this phonetic transparency does not eliminate challenges faced when building a TTS dataset. For example, during natural speech or when expressing emotions, speakers often skip or merge letters, alter pronunciation, or blend sounds [11]. In addition, the Kurdish language also lacks a standardized punctuation convention [11], [12], which makes it more difficult to map Kurdish script to speech. These issues introduce more complexity to the process of creating a clean and usable dataset for TTS models.

Given these inherent challenges in Kurdish TTS development, this work presents the first fine-tuned F5-TTS model for the Kurdish Sorani language. F5-TTS, which stands for A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching, is a recently developed open-source model that leverages flow-matching and diffusion-based techniques to generate natural and emotional speech without the need for extensive training data [13]. The process of fine-tuning this model began with the collection of 10.11 h of high-quality speech from professional TV recordings, which were then segmented into over 4,800 carefully curated samples covering diverse topics and speech styles. Each sample was manually reviewed for transcription accuracy and balanced duration.

The following are key contributions of this study:

- Introduces the first fine-tuned F5-TTS model for Kurdish Sorani, a low-resource language.
- Constructs a high-quality single-speaker Kurdish Sorani dataset with balanced segment durations, emotional variability, and transcription precision.
- Provides the first Kurdish TTS evaluation that combines both objective metrics (Character error rate [CER] and word error rate [WER]) and subjective MOS ratings, offering a more comprehensive assessment than prior studies.

The rest of this paper begins with a review of related studies on speech synthesis for low-resource languages and prior efforts in Kurdish TTS. Next, the dataset design is discussed, including how the speech data was collected, segmented, and transcribed. This is followed by an explanation of the F5-TTS model architecture and the fine-tuning process. Finally, the evaluation results are presented, and the implications of the findings are discussed, with concluding remarks and directions for future research.

2. RELATED WORK

TTS technology has seen fast advancements in recent years. However, for low-resource languages such as Kurdish, progress has been slower due to limited attention and support in the field of speech synthesis. This section provides an overview of traditional and neural speech synthesis methods, with a focus on approaches adapted to low-resource settings, and mentions previous efforts related to Kurdish TTS systems.

2.1. Traditional and Neural Speech Synthesis Methods

In the past, concatenative speech synthesis methods were the most widely used. They could achieve a noticeable amount of naturalness [14]. Concatenative data-driven approaches even made their way into speech recognition and some musical synthesis applications. Although some statistical techniques like the hidden Markov model provided some advancements in the field [15], [16], they were still lacking naturalness and accuracy [17].

Recent advancements in neural network-based models have led to improvements in TTS models [18]. End-to-end architectures such as Tacotron and Tacotron 2 [19] were able to produce more realistic and natural speech, but they require large amounts of high-quality training data. This leads to more difficulty when using it with low-resource languages like Kurdish.

2.2. TTS for Low-Resource Languages

Due to limited data availability, fine-tuning TTS models for low-resource languages presents several difficulties [20]. A common approach to address these challenges is transfer learning [21]. Transfer learning involves utilizing a pre-trained model and extending it to a new language. For example, a recent study [22] compared multiple cross-lingual transfer techniques for some low-resource languages, including Bulgarian, Georgian, Kazakh, Swahili, Urdu, and Uzbek. The source languages used in their research were English, Finnish, Hindi, Japanese, and Russian. This variability helped identify the source language that produced the best quality results. Achieving CERs from 6.70% to 61.92% and predicted MOS scores between 2.24 and 3.02, the study demonstrated that phonological features offer better generalization and are more effective than conventional phone mapping in cross-lingual TTS. However, their approach relied on around 10 h of source-language data and 10 min of target data.

Another technique that supports low-resource languages is multilingual training [23]–[25]. Training a model on multiple languages simultaneously can improve performance on low-resource languages by sharing linguistic properties among them.

Un addition, zero-shot learning techniques, where a model performs a task without having been explicitly trained on that task or language, can also help models to generalize to unseen languages [26], [27] without requiring a large amount of training data.

2.3. Related Kurdish TTS Work

At present, only a limited number of neural network-based TTS models exist for the Kurdish language. Some organizations have privately developed datasets and models for the language to meet internal needs. For example, Rudaw TV has developed a TTS model for reading news on their website. On the other hand, some recent studies have made noticeable contributions in the field:

- Muhamad and Veisi [6]: Utilizing transfer learning within an end-to-end architecture, they used a pretrained HiFi-GAN vocoder, a model that converts intermediate acoustic features into audible waveforms, together with Tacotron 2 on LJ-Speech dataset [28] to develop a Kurdish TTS system. They used English character embedding from a pretrained English model while feeding Kurdish characters as input. The dataset used in this study consisted of approximately 10 h of Kurdish speech recordings and corresponding transcriptions. Results showed a MOS of 4.1, which indicates a high level of naturalness of the developed TTS system compared to TTS systems in high-resource languages.
- Ahmad and Rashid [7]: They introduced an end-to-end TTS system specifically for Kurdish Sorani dialect. Their approach uses a pretrained variational autoencoder, a generative model that learns compact latent representations of data and reconstructs them back into realistic outputs, for audio waveform reconstruction, combined with adversarial training, a technique where a generator and discriminator compete to improve realism, to enhance the quality of the synthesized speech. In addition, they incorporated a stochastic duration predictor to enhance the naturalness of the output audio. The method facilitates real-time generation of Kurdish speech audio with variations in pitch and rhythm. Evaluation on a custom dataset showed a MOS of 3.94, showing better performance compared to one-stage and two-stage models using subjective human evaluation.
- Abdullah *et al.* [8]: Instead of using pretrained English models, this study improves an existing Kurdish TTS architecture based on Tacotron by training a WaveGlow vocoder from scratch. The training data consist of a 21-h native Kurdish speech dataset, particularly for Central Kurdish (Sorani) dialect. In addition to optimizing WaveGlow architecture, the study also introduces advanced prosody modeling techniques to improve the rhythm, stress, and intonation of synthesized speech. The adapted model achieved a MOS of 4.91, which sets a new benchmark for Kurdish speech synthesis.

While these studies have shown significant progress in the field of speech synthesis, they either depend on English-based architectures or need a large amount of training data. In contrast, our study focuses on fine-tuning F5-TTS, a more modern, up-to-date, and efficient architecture for the Kurdish Sorani language using a carefully prepared dataset.

3. DATASET DESIGN

Building a speech dataset from scratch is a gradual and careful process in which every single detail matters [20], [29]. You must care about every single detail. In the case of TTS, every single word matters because the quality and the diversity of the training data will undoubtedly impact the performance of the resulting model.

3.1. Dataset Collection

As the goal of this work was to fine-tune the F5-TTS model with a single-speaker dataset, a consistent and high-quality audio source was needed. After multiple options were evaluated, the recordings of Shaho Amin, a well-known newsreader from Rudaw TV, were selected for the following reasons:

- Public Availability: Rudaw TV's archives are publicly available and can be easily collected from online sources.
- Clear Pronunciation: The speaker's voice is understandable, clear, and well expressed.
- Loudness: Audio levels are consistent, and additional normalization is not required.
- Recording quality: As a major media channel, Rudaw TV uses advanced recording equipment, resulting in clean and high-fidelity audio.
- Familiarity: The speaker's voice is widely recognized, which allows listeners to better judge the similarity between synthesized and real speech.

Although the final dataset contains 10.11 h of audio, more than 15 h of video were initially collected. During preprocessing,

low-quality and noisy sections were removed. The two main sources for the recordings were Rudaw TV's website and YouTube Channel. For further processing, all collected content was then converted into WAV format using a lossless conversion process to preserve audio quality.

The collected recordings covered a variety of topics, including news broadcasts, interviews, shorts, and other random videos. This diversity was intended to increase the model's generalization by providing emotions, intonation patterns, and different speech rates in the dataset. After it was observed that the speaker's vocal tone was evolving over time, recordings were selected from different time periods (2018–2024). This helped enhance the dataset by adding more variability to the data.

3.2. Data Segmentation

After data collection, the dataset passed several preparation stages, with segmentation and transcription being the most time-consuming steps. In data segmentation, audio files were divided into manageable clips while ensuring the following:

- Proper logical sentence boundaries
- Natural starting and ending tone
- A balanced distribution of clip durations.

To achieve a balanced dataset, segments were distributed approximately equally across different duration ranges (e.g., 2–4 s, 4–6 s, 6–8 s, etc.), as shown in Fig. 1. This prevented the model from overfitting to specific speech length ranges. During this process, Ocenaudio software was utilized, as it allows the manual selection and export of clean segments.

3.3. Transcription Process

After all segments were cleaned, finalized, and prepared, the transcription process was initiated. Initially, segments were transcribed quickly to match what was heard in the recordings. In later stages, each segment was carefully reviewed multiple times word by word to eliminate the remaining spelling mistakes.

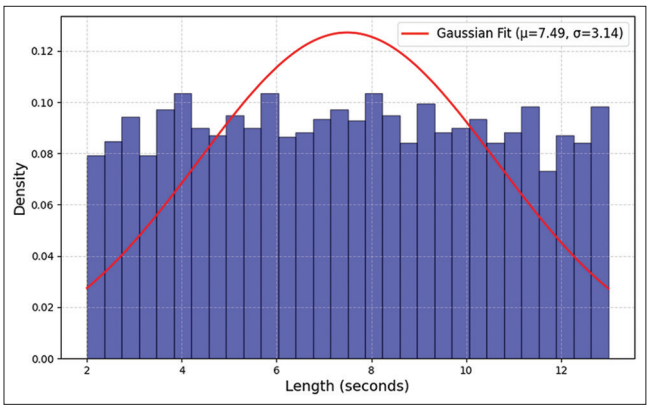

Fig. 1. Distribution of dataset samples.

Since punctuation plays an important role in expressing the tone and rhythm of synthesized speech, it was added where appropriate. Adding punctuation in the transcription helped the model to produce more realistic and expressive audio. Moreover, it allowed better control over speaking style when synthesizing new text. Most common punctuation use cases were:

- Adding a full stop (.) at the end of complete sentences to indicate closure
- Using a comma (,) to include short pauses or transitions between phrases
- Using a colon (:) when introducing reported speech
- Using question mark (?) for interrogative sentences
- Using an exclamation mark (!) for astonishing expressions.

While punctuation was added for expressiveness, each word itself was transcribed exactly as it was recited in the recordings. As Kurdish language is a non-phonemic writing system like Arabic and Persian, several challenges unique to the language were encountered despite efforts to avoid transcription mistakes:

- Some letters are written in a way but pronounced differently. For example, the letter (س) is pronounced as (ص) in words like شەست, موسولمان, سەت; similarly, (ت) may be pronounced as (ط) in تازە, سەتا, پێنجسەت.
- In fast speech, some letters are dropped or merged. For example, in چاوت تێژبێ, only one (ت) is clearly pronounced.
- Sometimes, when a letter comes twice repeatedly, it is hard to distinguish whether it is recited as one letter or two. For example, words like چاککردنەوە, رێککەوتنی contain consecutive (ک) sounds, but it is often unclear whether one or both are audibly pronounced.
- The English/ŋ/sound appears in some Kurdish words such as سنگ, مانگ, while there is not a dedicated letter for the sound in script.

3.4. Dataset Statistics

The final version of the dataset was 10.11 h long in 4,856 samples. Each sample was stored in WAV format. The dataset vocabulary had 2,567 unique tokens. Detailed statistics are provided in Table 1.

Fig. 1 demonstrates how samples are distributed evenly based on duration, which helps the model generalize across different speech lengths.

4. F5-TTS MODEL FINE-TUNING

Fine-tuning TTS models for low-resource languages like Kurdish Sorani can be challenging, primarily due to the lack of training data and limited computational resources. This section describes the steps taken to fine-tune the F5-TTS model for Kurdish Sorani using a specialized speech dataset. It includes an overview of the F5-TTS architecture, the data preparation and training environment, and a description of the fine-tuning process. An overview of the full training pipeline is presented in Fig. 2.

4.1. Overview of F5-TTS Architecture

Recent advancements in TTS technologies have improved the naturalness and flexibility of synthesized speech. Previous TTS systems were primarily based on auto-regressive (AR) models, which generate speech one token at a time, with each step depending on the last. Although models such as variational inference with adversarial learning for end-to-end text-to-speech and Tacotron2 [19] achieved high-quality results, they still suffer from inference latency and errors during sequential generation.

TABLE 1: Training dataset statistics

Metric	Value
Total duration	10.11 h
Number of samples	4,856
Shortest clip	2.0 s
Longest clip	13.0 s
Audio format	WAV
Vocabulary (Unique words)	2,567

F5-TTS is a fully non-autoregressive (NAR) TTS system, meaning it generates all audio tokens in parallel rather than sequentially, that produces high-quality speech using a flow matching framework, a generative modeling technique that learns to map noise into realistic data distributions [30], combined with a diffusion transformer (DiT) backbone, a transformer-based architecture tailored for diffusion models [31]. NAR approaches, particularly those based on diffusion-based models, which iteratively denoise random noise to produce realistic data [32], and flow matching techniques, allow for faster and more robust synthesis.

Unlike traditional TTS systems that rely on phoneme alignment, explicit duration models, or complex text encoders, F5-TTS presents a simplified architecture. It can directly learn how to speak from text characters and noisy speech. The model architecture can be described in five main phases, as shown below:

- Input Processing: The input text is turned into a sequence of characters, then padded with filler tokens to match the length of target speech frames. This enables the model to learn implicit alignment during training, which leads to a more natural rhythm.
- Feature Refinement: Input text features are refined using ConvNeXt blocks, a modern convolutional neural network design that improves feature extraction efficiency and alignment with speech frames.
- Denoising Process: A DiT gradually denoises a sampled noisy speech signal, conditioned on the refined text representation. It turns the noise into clear speech based on input text.

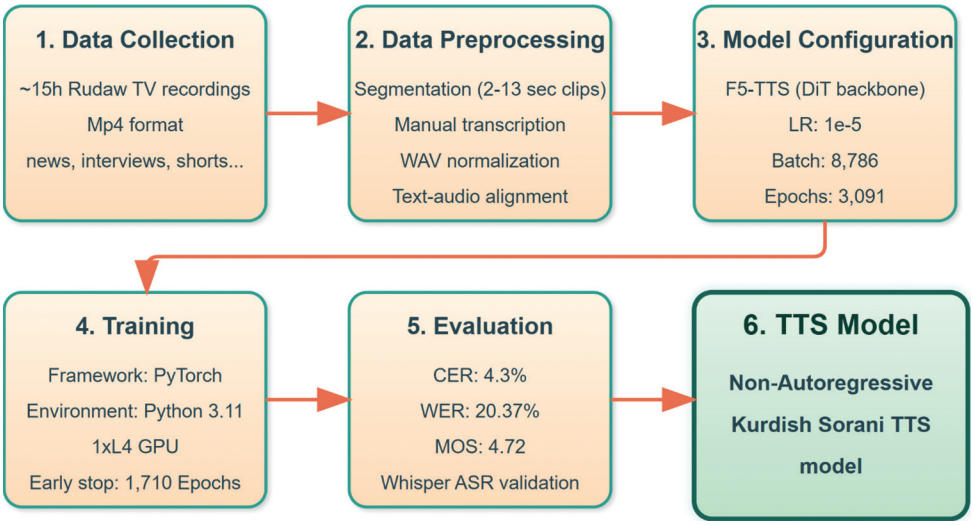

Fig. 2. Proposed model workflow.

- Training Objective: Training optimizes a flow-matching loss combined with a text-guided speech infilling task.
- Inference Strategy: Since diffusion models are typically slow due to step-by-step denoising, F5-TTS employs a sway sampling technique, an accelerated sampling method that reduces generation time while maintaining naturalness.

An overview of F5-TTS training (left) and inference (right) flow is illustrated in Fig. 3.

4.2. Data Pipeline and Training Environment

The dataset used for fine-tuning comprised 4,856 audio-text samples, totaling 10.11 h of Kurdish Sorani audio. Since the audio samples varied in duration, dynamic batching, a technique that groups sequences of similar lengths together to optimize training efficiency [33], was employed based on the number of spectrogram frames, with each batch being configured to contain approximately 8,786 frames to maximize GPU utilization [34]. As F5-TTS is a resource-intensive model, the fine-tuning process was conducted on Google Colab in a well-optimized environment. Hardware specifications and configuration parameters of the training process are provided in Tables 2 and 3.

4.3. Fine-tuning Process

After the environment was set up on Google Colab and 2 TB of Google Drive storage was allocated for saving checkpoints, the fine-tuning process began on April 18th, 2025, and continued for approximately 1 week without major difficulties.

The model started to produce reasonable speech within the first few thousand steps, and the quality gradually improved over time. While the fine-tuning was originally set to 3,091 epochs, it was terminated after 1,710 epochs (around 700,000 steps) because live output evaluations conducted every 5,000 steps revealed a close resemblance between the synthesized speech samples and the original recordings, indicating that further training was unnecessary.

As shown in Fig. 4, early in the training, the loss value hardly went under 0.4. However, by the final steps, the loss had decreased to 0.24272, which demonstrates a real

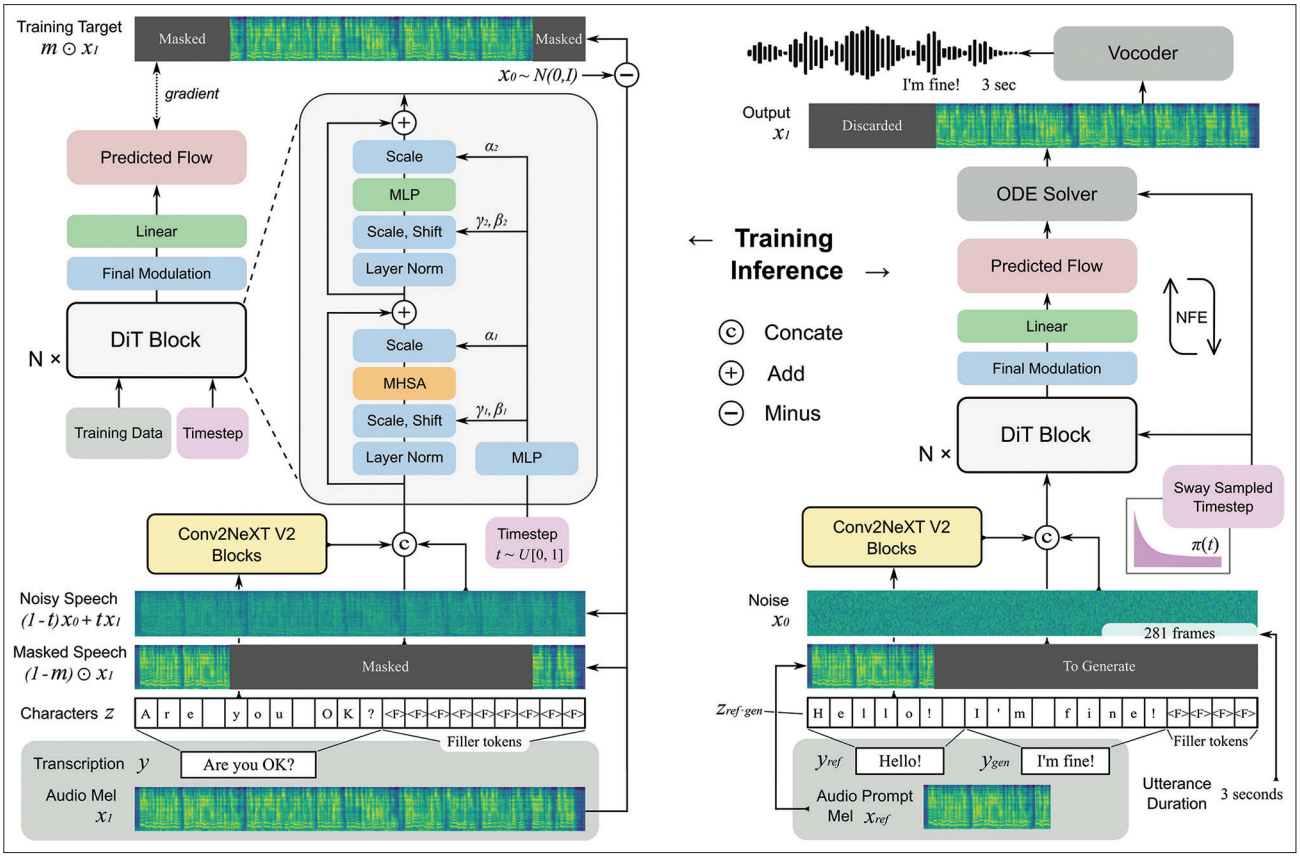

Fig. 3. F5-text-to-speech model architecture [13].

improvement in the model's convergence. This trend is also shown in Table 4, which presents the minimum loss values recorded across different fine-tuning intervals. In parallel, the learning rate schedule, demonstrated in Fig. 5, shows a gradual reduction in learning rate, contributing to a more stable training process. During the training, important metrics such as the loss curve, learning rate schedule, and other training statistics were monitored and recorded using the Weights and Biases (wandb) platform [35]. The decision to end the training early in the process was influenced by these tracking tools.

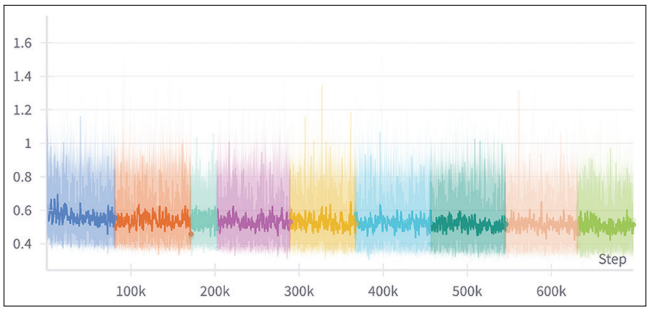

Fig. 4. Training loss value over steps.

TABLE 2: Training environment specifications

Component	Specification
GPU	NVIDIA L4 (22.5 GB VRAM)
CPU	6 physical cores, 12 logical threads
Operating system	Linux (Kernel 6.1.123+)
Python version	3.11.12
Training framework	PyTorch (with Weights and Biases (wandb) logging)

TABLE 3: Training configuration parameters

Parameter	Value
Batch size per GPU	8,786 frames
Batch size type	Frame-based
GPUs used	1
Gradient accumulation steps	1
Maximum samples	64
Learning rate	1e–5
Maximum gradient norm	1
Number of warmup updates	242
Number of epochs	3,091
Logging	Weights and Biases (wandb)
Checkpoint saving	Every 5,000 updates; save last at 2,000 updates

5. RESULTS DISCUSSION AND EVALUATION

The loss curve shown in Fig. 4 proves that the model has achieved excellent convergence and effective learning throughout the training process. To further assess the performance of the fine-tuned model, three evaluation metrics were used: CER, WER, and MOS. CER and WER are objective metrics that measure transcription accuracy at the character and word level, while MOS is a 1–5 scale for subjective rating of naturalness.

For objective metrics, the synthesized speech was transcribed using a pretrained Whisper Kurdish Sorani Automatic Speech Recognition (ASR) model ("PawanKrd/asr-large-ckb"), and the transcriptions were then compared to the original text. In contrast, the MOS metric was obtained based on human listener ratings, as the naturalness of speech may not be fully captured by objective metrics alone.

The results of CER and WER for three different samples are presented in Table 5. In both metrics, lower average scores indicate better synthesis quality. The average WER was 20.37%, and the average CER was 4.3%. These results demonstrate the model's strong transcription accuracy, particularly at the character level. The relatively high WER value of 20.37% may be due to limitations in the ASR model for Kurdish Sorani rather than actual quality issues, as demonstrated by the low CER and high MOS scores.

TABLE 4: Training loss value over time

From	To	Runtime	loss (Min)	lr
April 26, 2025 09:09	April 27, 2025, 02:59	17 h 50 m 16 s	0.24272	4.47E-06
April 25, 2025 02:54	April 26, 2025 02:46	23 h 52 m 27 s	0.25062	4.99E-06
April 23, 2025 21:46	April 24, 2025 21:34	23 h 47 m 59 s	0.25407	5.67E-06
April 22, 2025 20:36	April 23, 2025 20:30	23 h 54 m 15 s	0.27873	6.38E-06
April 21, 2025 19:49	April 22, 2025 19:14	23 h 25 m 6 s	0.28213	7.08E-06
April 20, 2025 17:58	April 21, 2025 17:49	23 h 51 m 46 s	0.28673	7.71E-06
April, 20 2025 07:47	April 20, 2025 16:52	9 h 4 m 31 s	0.30502	8.40E-06
April 19, 2025 06:42	April 20, 2025 06:08	23 h 26 m 19 s	0.31336	8.64E-06
April 18, 2025 06:50	April 19, 2025 04:49	21 h 58 m 36 s	0.3218	9.36E-06

TABLE 5: Word error rate (WER) and character error rate (CER) for synthesized Kurdish speech samples

Sample	Original text	Generated automatic speech recognition for the synthesized speech	WER (%)	CER (%)
1	هاونیشتیمانیەک بە ڕوداوی گوت زۆرم برسی بوو، بەرامبەر صەتڵێ شەبەت و وەجبەیە خورما دانیشتبووم هەل لەگەڵ مامۆستا گوتی ئەڵاهو ئەگبەر پڕم دایە ڕاست تا تێر نەبووم وازم نەهێنا	هاونیشتیمانییەک بە ڕوودوای گوت زۆرم برسی بوو بەرامبەر سەتڵێ شەبەت و جەبەی خورما دانیشتبووم هەر لەگەڵ مامۆستا گوتی ئەڵاو کەو ڕەپ دایە ڕاست تا تێر نەبووم وازم نەئێنا	35.71	7.10
2	لە نیویۆرک، ئەم پیاوە ئەو مۆزەی کڕی بە 6 ملیۆن و 200 هەزار دۆلار	لە نیویۆرکی ئەم پیاوە ئەو مۆزەی کڕی بە 6 ملیۆن و 200 هەزار دۆلار	14.29	3.08
3	صەدان هەزار کەس کارێکی کوردی وەک یەکێک لە ناوازەترین کارەکانی ساڵ و گێڕانەوەی چیرۆکی کۆرۆوای بە کوردی ئەبینن.	سەدان هەزار کەس کارێکی کوردی وەک یەکێک لە ناوازەترین کارەکانی ساڵ و گێڕانەوەی چیرۆکی کۆرۆوای بە کوردی ئەبینن	11.11	2.73
Average			20.37	4.30

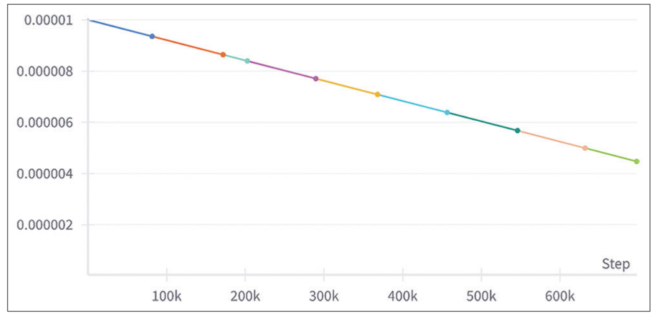

Fig. 5. Learning rate schedule during training.

To complement these objective measures with human perception assessments, a subjective evaluation was conducted using the MOS metric. Synthesized speech samples were shared online with more than 50 different native Kurdish Sorani speakers of different ages and genders. Each listener was asked to rate the naturalness of the same set of synthesized speech clips on a five-point scale ranging from 1 (Bad) to 5 (Excellent). The average score achieved in this metric was 4.72 out of 5, implying that the generated speech was perceived as highly natural. While this result is promising, it is important to remember that the number of listeners was limited, and no statistical significance testing was applied. Future evaluations with larger participant groups and formal statistical analysis would lead to more reliable results.

At the same time, these findings were obtained using clean single-speaker data in controlled settings. Several factors that may impact system performance in real-world scenarios, including background noise, overlapping speech, as well as speaker and dialect variability. Handling these challenges may require additional techniques and extended datasets, which can be explored in future research directions.

In comparison to older TTS approaches for low-resource languages such as transfer learning, the results of our diffusion-based model show significant improvements. Table 6 provides a comparison of various methods applied to low-resource languages, including different architectures and learning strategies. Notably, while prior Kurdish TTS studies have reported their MOS scores, none have evaluated their systems using objective evaluation metrics, such as CER or WER. This work is the first to apply these objective measures to Kurdish Sorani TTS for a more precise evaluation.

Beyond the technical achievements, this work has significant cultural and linguistic importance. Developing high-quality TTS systems for underrepresented Kurdish Sorani can

TABLE 6: Comparison of text-to-speech systems for low-resource languages

Study	Target language	Model architecture	Training data	Learning type	Character error rate (%)	Mean opinion score (1–5)
Do *et al*. [22]	Bulgarian (bg) Georgian (ka) Kazakh (kk)	FastSpeech 2+HiFi-GAN	Source: ~10 h/ speaker Target: ~10 min/speaker	Transfer learning	6.70 32.05 18.83	3.02 2.43 2.37
Abdullah *et al*. [8]	Kurdish Sorani (ku)	Tacotron 2+HiFi-GAN	10 h	Transfer learning	N/A	4.1
Ahmad and Rashid [7]	Kurdish Sorani (ku)	Variational autoencoder+Adversarial	Custom dataset	Stochastic duration predictor	N/A	3.9
Proposed model	Kurdish Sorani (ku)	F5-TTS (diffusion-based)	10.11 h	Diffusion-based	4.30	4.72

support digital inclusion, education, and even cultural preservation. It ensures the language is not left behind in the advancements of TTS technology. The fine-tuned model becomes a core module in many applications such as audiobooks, e-learning materials, and automated dubbing tools. In addition, building computational resources for Kurdish helps document and standardize the language, which faces dialectal diversity and limited digital resources. For this reason, the benefits of this research go beyond speech synthesis only; it can serve both practical applications and the preservation of Kurdish language identity.

6. CONCLUSION AND FUTURE WORKS

In this study, the first F5-TTS model was successfully fine-tuned for Kurdish Sorani, a low-resource language that has received less attention in the field of speech synthesis. The process began by creating a 10.11-h speech dataset. The work focused mostly on balancing segments, topic variety, emotions, transcription accuracy, punctuation, and other details that optimized the dataset quality. Using this high-quality dataset, the open-source F5-TTS model was fine-tuned in a well-optimized environment and configuration settings. Evaluations showed that the system learned to synthesize natural, emotional, and human-like Kurdish speech.

Compared to previous Kurdish TTS works, this study indicates that modern diffusion-based models like F5-TTS can improve the naturalness and reliability of speech synthesis for low-resource languages without requiring a massive amount of training data or using a pre-trained English-based architecture. The results show that a carefully constructed dataset can overcome the typical limitations of low-resource TTS systems.

While the fine-tuned model provides good results, there are still some areas for further improvement. One limitation is the use of a single-speaker dataset, which may reduce generalizability to different voices or dialects. Using a multi-speaker dataset that includes various emotions and accents would enable additional training to improve this fine-tuned model. This can improve the model's generalization for different speaking styles. In addition, real-world deployment introduces challenges such as background noise, overlapping speech, and speaker variability, which may reduce accuracy. Addressing these conditions in future work will improve the system's robustness and support its integration into practical Kurdish applications such as audiobooks, dubbing, and other voice-based systems.

REFERENCES

[1] L. Mohasi and D. Mashao. "Text-to-Speech Technology in Human-Computer Interaction". In: 5th *Conference on Human Computer Interaction in Southern Africa, South Africa (CHISA 2006, ACM SIGHI)*, pp. 79-84, 2006.

[2] V. R. Reddy and K. S. Rao. "Better Human Computer Interaction by Enhancing the Quality of Text-to-Speech Synthesis". In: *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, IEEE, United States, pp. 1-6, 2012.

[3] R. Zhen, W. Song, Q. He, J. Cao, L. Shi and J. Luo. "Human-computer interaction system: A survey of talking-head generation". *Electronics*, vol. 12, no. 1, p. 218, 2023.

[4] T. Xie, Y. Rong, P. Zhang, W. Wang and L. Liu. "*Towards Controllable Speech Synthesis in the Era of Large Language Models: A Systematic Survey*". [arXiv Preprint]; 2025.

[5] T. Reitmaier, E. Wallington, D. K. Raju, O. Klejch, J. Pearson, M. Jones, P. Bell, S and Robinson. "Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1-17, 2022.

[6] S. Muhamad and H. Veisi. "End-to-end kurdish speech synthesis based on transfer learning". *Passer Journal of Basic and Applied Sciences*, vol. 4, no. 2, pp. 150-160, 2022.

[7] H. A. Ahmad and T. A. Rashid, "Central Kurdish text-to-speech synthesis with novel end-to-end transformer training". *Algorithms*, vol. 17, no. 7, p. 292, 2024.

[8] A. A. Abdullah, S. S. Muhamad and H. Veisi. "*Enhancing Kurdish*

Text-to-Speech with Native Corpus Training: A High-Quality Waveglow Vocoder Approach". [ArXiv Preprint]; 2024.

[9] M. K. Mahmood, A. Q. H. Rash, M. A. Q. H. Rash, S. K. Mahmood and H. Güler. "Kurdish and Persian: Dialects or separate languages?" *International Journal of Social Science and Human Research*, vol. 6, p. 2216, 2023.

[10] G. Tavadze. "Spreading of the Kurdish language dialects and writing systems used in the Middle East". *Bulletin of the Georgian National Academy of Sciences*, vol. 13, no. 1, pp. 170-174, 2019.

[11] K. S. Esmaili. "*Challenges in Kurdish Text Processing*". [arXiv Preprint]; 2012.

[12] E. M. Qadir and H. H. Padar. "Punctuation in English and Kurdish: A contrastive study". *Koya University Journal of Humanities and Social Sciences*, vol. 5, no. 1, pp. 41-61, 2022.

[13] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu and X. Chen. "*F5-tts: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching*". [Preprint]; 2024.

[14] A. J. Hunt and A. W. Black. "Unit selection in a concatenative speech synthesis system using a large speech database". In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. IEEE, Atlanta, GA, USA, pp. 373-376, 1996.

[15] A. Falaschi, M. Giustiniani and M. Verola. "A hidden Markov model approach to speech synthesis". In: *First European Conference on Speech Communication and Technology, Presented at the EUROSPEECH*. pp. 2187-2190, 1989.

[16] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi and K. Oura. "Speech synthesis based on hidden Markov models". *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234-1252, 2013.

[17] H. Zen, K. Tokuda and A. W. Black. "Statistical parametric speech synthesis". *Speech Communication*, vol. 51, no. 11, pp. 1039-1064, 2009.

[18] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, Q. Le, Y. Agiomyrgiannakis, R. Clark and R. A. Saurous. "*Tacotron: Towards End-to-End Speech Synthesis*". [ArXiv Preprint]; 2017.

[19] J. Shen, R. Pang, R. J. Weiss, M. Schuste, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis and Y. Wu. "Natural tts Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions". In: *Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, United States, pp. 4779-4783, 2018.

[20] H. Ali, S. Subramani, R. Varahamurthy, N. Adupa, L. Bollinani and H. Malik. "*Collecting, Curating, and Annotating Good Quality Speech Deepfake Dataset for Famous Figures: Process and Challenges*". Cornell University, United States, 2025.

[21] K. Azizah and W. Jatmiko. "Transfer learning, style control, and speaker reconstruction loss for zero-shot multilingual multi-speaker text-to-speech on low-resource languages". *IEEE Access*, vol. 10, pp. 5895-5911, 2022.

[22] P. Do, M. Coler, J. Dijkstra and E. Klabbers. "*Strategies in Transfer Learning for Low-Resource Speech Synthesis: Phone Mapping, Features Input, and Source Language Selection*". [arXiv Preprint]; 2023.

[23] R. Huang, C. Zhang, Y. Wang, D. Yang, J. Tian, Z. Ye, L. Liu, Z. Wang, Z. Jiang, X. Chang, J. Shi, C. Weng, Z. Zhao and D. Yu. "Make-a-Voice: Revisiting Voice Large Language Models as Scalable Multilingual and Multitask Learners". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Vol. 1. [Long Papers], pp. 10929-10942, 2024.

[24] T. Saeki, G. Wang, N. Morioka, I. Elias, K. Kastner, F. Biadsy, A. Rosenberg, B. Ramabhadran, H. Zen, F. Beaufays and H. Shemtov. "Extending multilingual speech synthesis to 100+ languages without transcribed data". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, United States, pp. 11546-11550, 2024.

[25] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. J. Skerry-Ryan, Y. Jia, A. Rosenberg and B. Ramabhadran. "*Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning*". [arXiv Preprint]; 2019.

[26] T. Saeki, S. Maiti, X. Li, S. Watanabe, S. Takamichi and H. Saruwatari. "*Learning to Speak from Text: Zero-Shot Multilingual Text-to-Speech with Unsupervised Text Pretraining*". [arXiv Preprint]; 2023.

[27] Y. Xian, B. Schiele and Z. Akata. "*Zero-Shot Learning -- the Good, the Bad and the Ugly*". [arXiv Preprint]; 2020.

[28] "*The LJ Speech Dataset*". Available from: https://www.kaggle.com/datasets/mathurinache/the-lj-speech-dataset [Last accessed on 2025 Apr 28].

[29] A. Katumba, S. Kagumire, J. Nakatumba-Nabende, J. Quinn and S. Murindanyi. "A curated crowdsourced dataset of Luganda and Swahili speech for text-to-speech synthesis". *Data Brief*, vol. 62, p. 111915, 2025.

[30] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel and M. Le. "*Flow Matching for Generative Modeling*". [arXiv Preprint]; 2023.

[31] W. Peebles and S. Xie. "*Scalable Diffusion Models with Transformers*". [arXiv Preprint]; 2023.

[32] J. Ho, A. Jain and P. Abbeel. "*Denoising Diffusion Probabilistic Models*". [arXiv Preprint]; 2020.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and L. Polosukhin. "*Attention is All You Need*". [arXiv Preprint]; 2023.

[34] Z. Ge, L. Kaushik, M. Omote, and S. Kumar. "Speed up Training with Variable Length Inputs by Efficient Batching Strategies". In: *Interspeech*. pp. 156-160, 2021.

[35] "*Weights and Biases: The AI Developer Platform*". Weights and Biases. Available from: https://wandb.ai/site [Last accessed on 2025 May 04].