Central Kurdish text-to-speech and its application in speech-to-text translation

Mohammad Mohammadamini, Meysam Shamsi, Marie Tahon

Central Kurdish text-to-speech and its application in speech-to-text translation

Mohammad Mohammadamini, Meysam Shamsi, Marie Tahon
LIUM, Le Mans University
{first.last}@univ-lemans.fr

Abstract

In this study, we show how to develop high-quality TTS models for low-resource scenarios that according to our extensive evaluation are competitive with the models trained on dedicated TTS data recorded in the studio. We develop three Text-to-Speech (TTS) models for Central Kurdish as a low-resource language using F5-TTS architecture. The models are trained on Central Kurdish TTS datasets in which two of them are curated from audiobooks during this study and the third one is evaluated for the first time. We also demonstrate the potential of TTS models for developing other speech technologies in low-resource languages by proposing a speech synthesis framework used in a speech-to-text translation application, achieving promising results on standard speech translation benchmarks.The dataset and models are publicly available under the CC BY-NC-ND 4.0 license.

Keywords: Speech synthesis, Speech translation, Low-resource, Kurdish language

1. Introduction

Recent advancements in speech synthesis have enabled the generation of speech that closely approximates natural human voices across a wide range of languages (Lux et al., 2023). Beyond their role as standalone applications, these systems tends to become important components of broader technologies including audio language models (Pu et al., 2025), automatic speech recognition (Li et al., 2025), deepfake detection (Wang et al., 2024; Casanova et al., 2023), and speech translation (Mizumoto et al., 2025). Indeed, they can augment the initial database with synthetic data, create speaker diversity, or generate data from a specific domain. Nevertheless, many low-resource languages remain underrepresented in these developments.

The development of Text-to-Speech(TTS) models for many languages have recently gained the interest of the community, as shown by the diversity of languages proposed in the last Blizzard challenges: English, French (2023), Hindi (2015) and recently for Bildts, a unique language variety from the Netherlands in an under resource scenario [1]. When come to low resource languages, TTS faces many different issues (Louw, 2023). The available data to train the TTS models is scarce, and is usually non dedicated. For instance, traditional speech synthesis relies on clean audio data (mainly audiobooks or studio recordings), while low resource data can come from radio/TV shows or in the wild recordings with non professional speakers which may lead to heterogeneous qualities (Guennec et al., 2023). The number of speakers is generally low, sometimes a unique speaker is present. This low diversity might prevent the training of multispeaker models.

The focus of the current study is on providing TTS datasets and models for Central Kurdish, which is a low-resource language. Kurdish is a macrolanguage comprising a dialect continuum (Northern Kurdish, Central Kurdish, Southern Kurdish, Hawrami, Zazaki, and Laki) (Sheyholislami, 2015) spoken by more than 35 million people across a broad region known as Kurdistan, spanning four Middle Eastern countries (Iran, Iraq, Turkey, and Syria). The current study is limited to Central Kurdish, which is spoken by approximately 8 million people and written in a modified Arabic script (Sheyholislami, 2021).

In recent years, few studies on Kurdish TTS have been released. In (Muhamad et al., 2024) 21 hours audio from one female speaker recorded and used in training Tacotron 2 model, but the TTS models and the dataset from this study are not public. In (Ahmad and Rashid, 2024) 13 hours of male audio recorded in studio which is used as one of the training datasets in our current research. The MMS models from Meta supports both Northern Kurdish and Central Kurdish varieties (Pratap et al., 2024) but has some limitations such as the absence of a detailed evaluation, the unavailability of public TTS datasets, and the lack of a voice cloning process, which is essential for increasing speaker diversity in auxiliary tasks such as speech translation (Section 5).

The limited number of publicly available TTS models and datasets in Kurdish language, as well as the growing use of TTS models as speech synthesizers in various applications, have motivated us to introduce two new TTS training datasets,

[1] https://blogs.helsinki.fi/ssw13-2025/the-blizzard-challenge-2025/

three models, and a setup for data augmentation in speech translation applications. To this end, we first demonstrate how publicly available web data can be leveraged to develop high-quality speech synthesis for a low-resource language. As part of this effort, we introduce two new Central Kurdish datasets derived from audiobooks: one featuring a female voice and another a male voice. To further examine the relative potential of such web-derived data compared with data collected on purpose for speech synthesis (generally read speech recorded by actors in studio conditions), we also develop a system leveraging a dataset called Giganet (Ahmad and Rashid, 2024). To the best of our knowledge, this dataset, recorded in a professional studio and publicly available, has not been systematically evaluated or used to develop a TTS model prior to our work.

While subjective evaluation is the standard approach for assessing speech synthesis quality and intelligibility, it is often impractical for low-resource languages due to the difficulty of recruiting native speakers. To address this challenge, we perform an objective evaluation of the synthetic speech quality using multiple metrics in different aspects (Section 4). To validate this approach, we also conduct a comprehensive subjective evaluation of the audio quality with Mean Opinion Score (MOS) test.

Beyond evaluation, we further demonstrate the value of such TTS systems by generating synthetic data for training speech translation models. Through comprehensive experiments, we show that in scenarios where no dedicated natural speech exists, speech synthesis systems can produce diverse, high-quality data to enable the rapid and cost-effective development of other speech technologies for low-resource languages.

The current study presents diverse contributions: (i) the development of TTS systems for Central Kurdish including datasets and models, (ii) a comprehensive objective and subjective evaluation of TTS quality for different application purposes, and (iii) the development of a speech translation model for Central Kurdish trained solely on synthetic data.

2. Central Kurdish TTS datasets

Recording read speech in controlled acoustic environments is a common way of collecting data dedicated for TTS. However, creating such resources is expensive and sometimes results in limitations, such as lack of spontaneity and limited prosodic variability (Tahon et al., 2017). In this paper, we show how to reuse the available resources to extract TTS training datasets for Central Kurdish as a low-resource language. We provide two datasets, one male and one female voice, each containing around 11 hours of audiobook podcast recordings. The audio-books format was 44khz resampled to 24khz mono channel wav format. The two datasets are coming from different bibliographic domain (Table 1). The data preparation follows these steps:

- **Segmentation:** The long audiobook podcasts, each around 30 minutes in length, are segmented automatically using an energy-based voice activity detection. The segments between 1 and 15 seconds are retained.
- **Automatic transcription:** The segmented utterances are transcribed automatically using a fine-tuned version of Seamless for Central Kurdish developed in (Mohammadamini et al., 2025b). The ASR system performance on Asosoft benchmark (Veisi et al., 2022) which is a clean read speech, is at the level of 8% WER which shows the high quality of generated transcriptions and a minimum revision is required.
- **Manual revision and validation:** The Central Kurdish transcriptions are manually revised to correct possible errors and to remove problematic samples, such as noisy segments or distorted speech.

At the end of this process, we obtained two new datasets with the specifications listed in Table 1. The third dataset –Gigant denoted as “studio-M”– has been recorded in studio conditions by a male professional dubber reading sentences from different domains (Ahmad and Rashid, 2024).

Table 1: TTS training datasets

Dataset	Duration	#Utts	Domain
audiobook-M	10h51mins	6044	biography
audiobook-F	10h54mins	5572	biography
studio-M	13h35mins	6055	multi

3. TTS Models

In low resource contexts, the available data for model training is often heterogeneous and non dedicated, originating from radio or TV broadcasts, online repositories, or spontaneous recordings, resulting in varied acoustic conditions and inconsistent transcription quality. By choosing the training data from audiobooks, we tried to minimize these variabilities. However, since the datasets were not purposely recorded for TTS training, we do not have complete control over them.

Given these challenges, our design priorities were data efficiency, speaker flexibility, and robustness to noisy or non-standardized data. We therefore selected the multilingual F5-TTS (Chen et al., 2025) among other multilingual alternatives such

as VITS (Kim et al., 2021) or XTTS (Casanova et al., 2024). This choice is motivated by previous experiments in the framework of the Blizzard 2023 challenge (Mas et al., 2025). F5-TTS is a fully non-autoregressive text-to-speech system based on flow matching with diffusion transformer. Besides the input text for synthesis, the model takes as input a prompt that includes an audio sample from the target speaker and its transcription. It outputs mel-spectrograms which are given to Vocos (Siuzdak, 2023) vocoder.

This design ensures stable and efficient training while requiring fewer resources. The model takes as input a prompt composed of an audio sample from the target speaker and its corresponding transcription, together with the text to be synthesized. Since, one application of the developed models is using them to generate synthetic data for training speech translation models, this configuration enables speaker cloning and enhances prosodic diversity in the generated speech which are required for creating varied and natural-sounding dataset for speech translation.

In our experiments, we fine-tuned the English checkpoint of F5-TTS-base[2]. In general, F5-TTS does not rely on any grapheme-to-phoneme conversion which makes it particularly suited for low-resource languages lacking phonetic dictionaries or phonetizer. However its input vocabulary miss many Kurdish characters. Since Central Kurdish language is almost a phonemic language and have very high precision Grapheme-to-Phoneme (G2P) tools, two preprocessing steps are applied:

- **Normalization:** A normalization is done in all the pipelines for training and inference. The normalization includes Unicode, punctuation, number unification, and number-to-word conversion implemented in Asosoft library [3] (Mahmudi et al., 2019).

- **G2P Conversion:** A G2P conversion is applied on the normalized input text to make it compatible with the vocabulary of the F5-TTS baseline model (Mahmudi and Veisi, 2021).

As the number of speakers is clearly not enough to train a Central Kurdish multi-speaker TTS model, we choose to train three mono-speaker models. The fine-tuning of F5-TTS was performed independently for three datasets, leveraging 1 GPU (RTX8000, 48GB) in 2 days. For each dataset 500 samples separated as the test set and the rest of the data is reserved for training the model.

[2] https://huggingface.co/SWivid/F5-TTS/tree/main/F5TTS_Base

[3] https://pypi.org/project/asosoft/

4. TTS Evaluation

So far, there is no standard metric to precisely determine the exact epoch at which we should stop training a TTS model (Peiró-Lilja et al., 2022). This identification is generally realized with an empiric number of training steps. In our experiments, we design a protocol to automatically evaluate the model during the training stage. Once the training finished, this protocol will be examined through a perceptive evaluation process.

4.1. Objective evaluation

To evaluate the TTS systems automatically, we focus on three key aspects of synthetic speech quality: *Naturalness*, *Signal Quality*, and *Intelligibility* following our previous work (Mas et al., 2025).

1. Naturalness measures how closely the synthetic speech resembles human speech, regardless of the listener's familiarity with the language. We evaluate naturalness from two complementary perspectives:

- Estimating the overall perceived naturalness using a Mean Opinion Score (MOS) predictor.

- Exploiting a deepfake classifier model to assess the likelihood of the audio being bona fide versus fake.

The predicted MOS (MOS_p) is obtained using `Squim` package from `Torchaudio`, a non-intrusive model that estimates perceived naturalness without requiring a matching reference. This model has been trained on synthetic speech sentences in English generated with various TTS and Voice Conversion systems, the BVCC dataset (Huang et al., 2022). This returns a score ranging from 1 (poor naturalness) to 5 (highly natural).

To complement the MOS_p, we extracted the *pseudo bona-fide probability* (*DF*) from the deepfake classifier *AASIST*[4] (Jung et al., 2022). This score ranges from 0 (fake) to 1 (bona-fide). Under the assumption that more natural speech is harder to classify as fake, a higher pseudo bona-fide probability is interpreted as an indicator of improved naturalness. This metrics is motivated by prior work (Miniconi et al., 2025).

2. Signal Quality evaluates the integrity of the synthetic signal, considering factors such as noise, artifacts, and reverberation. We adopted three non-intrusive metrics, originally developed for assessing telecommunication speech quality:

- *PESQ* (Perceptual Evaluation of Speech Quality) (Rix et al., 2001) estimates human perceived audio quality by accounting for dis-

[4] https://git-lium.univ-lemans.fr/asini/df_rank_ssq

tortions such as noise, clipping, and interference, without requiring clean reference signals. PESQ produces scores typically within $[-0.5, 4.5]$.

- *WSNR* (Weighted Signal-to-Noise Ratio) (Kim and Stern, 2008) provides an estimate of the effective signal-to-noise ratio.

- *SRMR* (Speech-to-Reverberation Modulation Energy Ratio)[5] measures the ratio of modulation energy between the direct speech and its reverberant components, using a spectral modulation approach. Higher SRMR values indicate better speech quality.

3. Intelligibility assesses how accurately the synthesized speech conveys the semantic and textual content of the input text. To measure intelligibility, we used SeamlessM4T V2 Large models fine-tuned for Central Kurdish prsented at (Mohammadamini et al., 2025b). According to the reported results, this is the SOTA ASR model for Kurdish language by giving a 8 and 20 *WER* on Asosoft (Veisi et al., 2022) and Fleurs benchmarks (Conneau et al., 2023) respectively.

We developed `tts4all_eval` [6] a Python package to calculate all evaluation metrics described above.

The fine-tuning process and performance evolution of different quality metrics with respect to training steps are illustrated in the Figure 1, which summarizes the convergence behavior. The *DF* measure increases consistently across all systems, indicating that the generated speech signals become progressively more similar to natural speech as perceived by a deepfake classifier. The ASR performance, measured in terms of *CER*, shows a steady improvement in intelligibility during the first 100k iterations, after which it stabilizes and converges for all systems. Regarding signal quality, represented by the *PESQ* metric, distinct behaviors are observed among systems. While the Studio-M system maintains a consistently high and stable signal quality, the Audiobook systems benefit from the early stages of training, which lead to improved *PESQ* scores. However, the Audiobook-M system exhibits a slight degradation in later iterations, which we believe results from a trade-off between different aspects of quality.

[5] https://lightning.ai/docs/torchmetrics/stable/audio/speech_reverberation_modulation_energy_ratio.html

[6] https://git-lium.univ-lemans.fr/jsalt2025/wp1/tts4all_eval

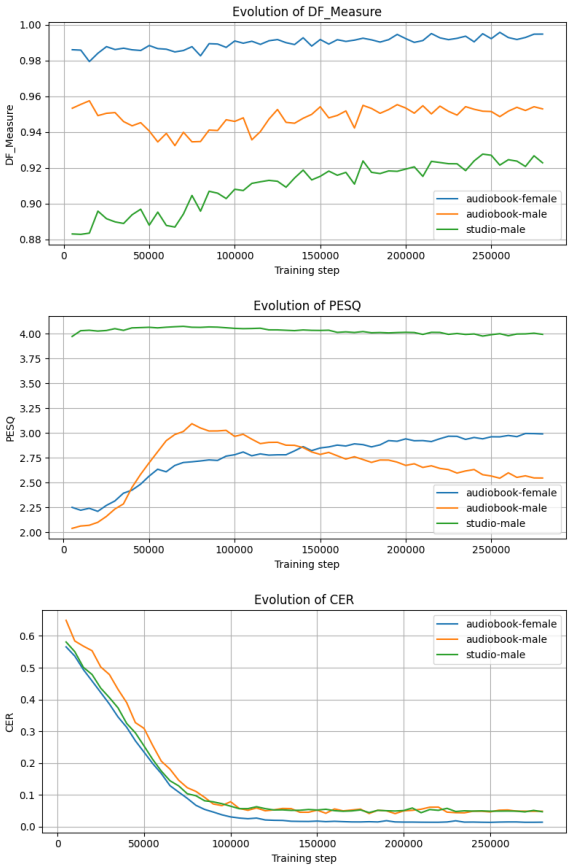

Figure 1: Learning curves for different aspects of synthetic speech quality. While the improvement in intelligibility (reduction of *CER*) is evident for all systems, the improvements in *DF* measure (naturalness) and *PESQ* (Signal Quality) for the audiobook-M voice are not optimal at the final checkpoints.

4.2. Subjective evaluation

Subjective evaluation remains a crucial component in assessing the effectiveness of TTS systems, particularly for languages like Central Kurdish where the nuances of pronunciation, idiomatic expressions, and code-switching patterns are underrepresented in existing datasets. To capture these perceptual aspects, we designed an evaluation across multiple linguistic domains that reflect the diversity and complexity of the language, allowing us to measure how well synthesized speech preserves naturalness, expressiveness, and intelligibility in different scenarios.

The subjective evaluation is conducted across several categories, which are listed bellow:

- **Code switching:** A sentence that has at least one English loanword. The English word is written in Kurdish script according to accepted orthographic rules.

- **Prosodic variability:** A sentence that expresses an emotion such as fear, sadness,

or happiness, or a sentence that shows disappointment or lamentation.

- **Expressions and idioms:** A sentence that contains a Kurdish expression or idiom.
- **Semantically unpredictable (sus):** A sentence that is syntactically correct but semantically nonsensical.
- **News:** A neutral sentence adopted from news agencies.
- **In-domain test:** Test sentences separated from the TTS dataset.
- **Natural test:** Speech utterances separated as test set from the TTS dataset.

For each category, we selected sentences which reveals distinctive characteristics of the Central Kurdish language such as code-switching, or target application vocabulary or expressions. In each category, 20 synthesized sentences and 20 genuine speech sentences are included, resulting in 140 sentences per system for evaluation. We decided to run a Mean Opinion Score (MOS) evaluation.

Participants were asked:

"How close is the following voice to real human speech?"

and were instructed to rate them on a MOS scale from 1 to 5. Each listener was asked to evaluate 40 sentences.

From the 3,101 submitted answers by 88 participants, the perceived MOS was calculated for both synthetic and natural speech samples. The average MOS scores are presented in Table 2. The first column shows the average MOS assigned to genuine samples by the evaluators. The evaluators were not informed about the presence of real speech samples. The second column shows the average MOS for synthesized samples across different domains and categories. Among the three systems, those trained on data with higher MOS achieved higher MOS scores in the TTS system. The results indicate marginally better performance for the female audiobook dataset compared to the multi-domain dataset recorded in the studio.

Table 2: Average and std of MOS per system

speaker	natural	tts_mos	std_mos
audiobook-F	4.42	4.08	0.98
audiobook-M	4.34	4.01	1.06
studio-M	4.40	4.06	1.04

The results in Table 3 show the performance of three TTS systems trained with different datasets, audiobook-M, audiobook-F, and studio-M, across multiple application categories. In general, predicted MOS_p values remain consistently high ($\simeq$4.3–4.4) and closely match subjective MOS ratings. In all categories, the *CER* is less than 8% and remains consistent across the three speakers.

The audiobook-F system achieves slightly higher MOS in code-switching conditions (4.29). The studio-M system, however, outperforms both audiobook-trained voices in acoustic fidelity (*PESQ*=4.0 for in-domain).

The subjective MOS reveals a difference between natural and in-domain conditions for three systems, this distinction is not captured by objective measures such as *DF* or MOS_p, highlighting the limitations of current objective metrics. Among the objective metrics, *DF* shows the highest correlation with subjective MOS, with a Spearman rank correlation coefficient of 0.42 ($p < 0.001$).

Across systems, all models show difficulty in handling expressiveness and prosodic variability, where subjective MOS decreases. The synthetic speech in prosodic-variability scenario remain the most challenging case. This can be attributed to the lack of contextual information for both synthesis and evaluation steps, as listeners are often unable to accurately judge the quality or appropriateness of a synthetic signal without a predefined context that frames the intended emotion, emphasis, or prosodic pattern.

5. Speech translation

In this section, we explore the use of TTS models for generating synthetic data to train Central Kurdish to English speech translation models as an auxiliary task.

5.1. End-to-end Speech-to-Text Translation (E2E S2TT)

E2E S2TT is the direct translation of speech from a source language into text in a target language (Barrault et al., 2025). Training E2E S2TT models requires a large amount of parallel data consisting of source language audio and corresponding target language translations. However, many low-resource languages lack such datasets. In this section, we apply our TTS models to extend the available bitext data (i.e. parallel Kurdish to English text pairs) by synthesizing speech from the source language text. Finally, we use the synthetic speech and English translations to train E2E S2TT models. In our experiments, we employ three TTS models and a pool of sentences randomly selected as references. The references are selected randomly from Central Kurdish part of Common Voice (Ardila et al., 2020).

Table 3: Averages and Confidence Intervals (95%) for metrics by training datasets and categories.

TTS	Category	DF ↑	MOS_p ↑	PESQ ↑	WSNR ↑	SRMR ↑	CER ↓	MOS ↑
audiobook-M	natural test	0.98±0.03	4.36±0.03	3.27±0.12	39.15±7.13	1.68±0.43	0.02±0.01	4.34±0.14
	in-domain	1.00±0.00	4.35±0.05	2.41±0.17	70.03±13.09	1.91±0.58	0.02±0.01	4.20±0.16
	code-switching	0.97±0.06	4.39±0.02	2.56±0.21	95.80±6.07	3.68±0.46	0.07±0.03	4.04±0.16
	sus	0.96±0.05	4.40±0.01	2.54±0.24	79.77±10.73	3.09±0.51	0.04±0.02	3.97±0.17
	news	1.00±0.00	4.38±0.01	2.39±0.14	36.04±3.20	0.69±0.08	0.07±0.07	3.97±0.18
	prosodic-variablity	0.81±0.14	4.37±0.03	2.89±0.29	94.30±7.35	4.01±0.50	0.07±0.05	3.88±0.17
	expression	0.94±0.07	4.39±0.02	2.52±0.26	91.40±8.43	3.63±0.43	0.04±0.03	3.66±0.19
audiobook-F	natural test	1.00±0.00	4.37±0.03	3.74±0.16	50.59±13.51	1.92±0.47	0.01±0.01	4.42±0.14
	in-domain	1.00±0.00	4.39±0.01	3.04±0.22	41.85±7.52	1.88±0.49	0.01±0.01	4.26±0.14
	code-switching	1.00±0.00	4.35±0.02	3.30±0.19	89.46±8.34	3.65±0.47	0.03±0.01	4.29±0.14
	sus	1.00±0.00	4.37±0.01	3.25±0.24	69.61±11.94	3.17±0.48	0.02±0.01	3.91±0.16
	news	1.00±0.00	4.40±0.01	2.88±0.11	28.89±2.27	0.77±0.14	0.08±0.07	4.09±0.14
	prosodic-variablity	0.93±0.08	4.32±0.03	3.32±0.23	90.44±9.49	4.46±0.76	0.07±0.06	4.01±0.15
	expression	0.98±0.03	4.33±0.07	3.18±0.26	89.45±10.27	3.63±0.45	0.04±0.04	3.61±0.19
studio-M	natural test	0.98±0.04	4.42±0.02	4.16±0.06	15.41±1.22	2.32±1.05	0.09±0.08	4.40±0.14
	in-domain	0.97±0.05	4.43±0.01	4.00±0.07	36.37±10.57	1.87±0.78	0.08±0.08	4.20±0.17
	code-switching	0.93±0.08	4.44±0.01	4.05±0.10	52.37±11.47	4.17±0.47	0.04±0.03	4.11±0.18
	sus	0.96±0.05	4.43±0.01	3.96±0.13	48.36±9.23	3.46±0.43	0.02±0.01	3.96±0.19
	news	1.00±0.00	4.44±0.00	4.08±0.07	23.73±2.02	0.81±0.12	0.07±0.07	3.88±0.16
	prosodic-variablity	0.74±0.16	4.38±0.05	3.78±0.17	67.81±12.91	4.61±0.59	0.06±0.05	3.95±0.16
	expression	0.87±0.13	4.43±0.03	4.02±0.07	54.59±9.80	3.98±0.44	0.03±0.02	3.89±0.17

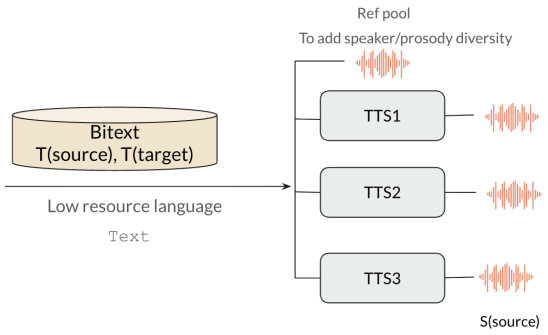

Figure 2: Speech synthesis pipeline for S2TT

5.2. Looking for data synthesize scenario

Leveraging synthetic speech as training data lags behind real speech mainly due to limited speakers, expressivity and acoustic diversity, which leads to strong convergence on the training set but poorer generalization to real conditions. To make synthetic training data closer to real-world scenarios, we explored several setups. To identify the best setup, we simplified the problem by removing the linguistic variability required in speech translation. Specifically, we used the FLEURS test set as both the training and evaluation data (Conneau et al., 2023). Although this scenario is not realistic, having identical linguistic content in training and test is expected to yield the highest possible performance. In this way, we remove the performance degradation due to linguistic variability or translation quality, and any remaining performance loss can be attributed to speech/acoustic factors.

A series of speech translation experiments are reported in Table 4. First, we fine-tune a Whisper v3 large (Liu et al., 2024) model on the real FLEURS CKB→EN test set, achieving a BLEU score of 92.42, which we treat as the upper bound. We then evaluate configurations combining one to three TTS models and prompt references drawn from a pool. The reference pool includes 5000 utterances chosen randomly from Common Voice 18. As intuition suggests, using three TTS models with randomly selected prompts produces richer training data and achieved a BLEU score of 79.95 which is better than using one TTS or one fixed reference from the same speaker. Accordingly, in our speech translation experiments we randomly selected both the TTS model and the prompt to generate the training data.

When we use only one TTS model for data generation, we expect to find a correlation between the quality of our TTS model and the performance of S2TT model. The results obtained by three TTS models show this correlation. As it is shown in Table 2, the TTS model trained on the audiobook-M achieves a lower MOS (4,01) and in Table 4 achieving 73,92 BLEU score, the synthetic data gives lower performance in S2TT application in comparison to two other TTS models.

5.3. Speech translation experiments

Using the setup shown in Figure 2, we generated approximately 500 hours of synthetic speech from a bitext corpus. The total number of synthesized samples is 340k. The first source of the bitext data used for synthesis is Kuvost, the Kurdish translation of Common Voice English (Mohammadamini et al., 2025a), which accounts for 200k of the synthesized samples. The second source is a parallel dataset introduced in (Mohammadamini et al., 2025b), compiled from parallel books, certified translators, philo-

Table 4: BLEU scores given by Whisper model, fine-tuned on real and synthetic data obtained with various TTS setups. The real and synthesized data are Fleurs test set. The Random prompt means a randomly chosen reference while the Fixed one comes from the same speaker as the TTS speaker.

Training Material				BLEU
TTS			Prompt	
studio-M	audiobook-F	audiobook-M		
Yes	Yes	Yes	Random	**79.95**
Yes	No	No	Random	76.87
No	Yes	No	Random	74.64
No	No	Yes	Random	73.92
Yes	No	No	Random	76.87
Yes	No	No	Fixed	76.89
Fleurs genuine speech				92.42

sophical encyclopedias, and other sources. The remaining 140k synthesized samples come from this second resource. During the synthesis process, one of the three TTS models is selected at random, and the reference is also chosen randomly from among the 5000 files in the Central Kurdish Common Voice speech dataset.

With the synthetic audio for Central Kurdish and the English translation, we fine-tune Whisper v3 large model. The fine-tuned model is evaluated on the CKB→ENG part of Fleurs benchmark. The second evaluation benchmark used in this paper is called Asosoft benchmark. This benchmark originally proposed for multi-domain ASR which includes 100 unique sentences from different domains and we extended for speech translation by adding the English translations (Veisi et al., 2022). Table 5 shows the results for zero-shot case and fine-tuned Whisper model. The fine-tuned model achieve a 27.23 BLEU score on Asosoft benchmark and 18.50 BLEU score on Fleurs benchmark.

Table 5: S2TT BLEU score under two evaluation protocols

Model	Train Data	Asosoft	FLEURS
Zero-Shot	—	2.31	1.51
Synthesized	340k (514h)	27.23	18.50

As discussed in the Section 4, one reason behind choosing F5-TTS model is accepting prompts as reference from different speakers. We hypothesized that this characteristic of F5-TTS model can bring a speaker diversity which is required to design generalizable speech translation models.

In order to show the speaker diversity in the synthetic speech, we plot the ECAPA-TDNN (Desplanques et al., 2020) speaker embeddings of 20 synthetic utterances generated by the three systems and five different prompts from 5 different speakers (Figure 3). The Speechbrain toolkit is used as speaker embedding extractor (Ravanelli et al., 2021). For a given reference and a given system all embeddings are plotted in the same cluster sharing the same speaker characteristics. When the same reference is given as prompt, we expect that all systems give close speaker embeddings. The T-SNE visualization shows that in some cases utterances generated by a system tends to be more similar than utterances generated with same prompt. Therefore, we can entail that the training speaker characteristics still exist in the synthetic utterances. This visualization and the results presented in Table 4 shows that the speaker information captured by prompting is limited and it is required to target the integration of speaker characteristics with the synthesized speech more explicitly.

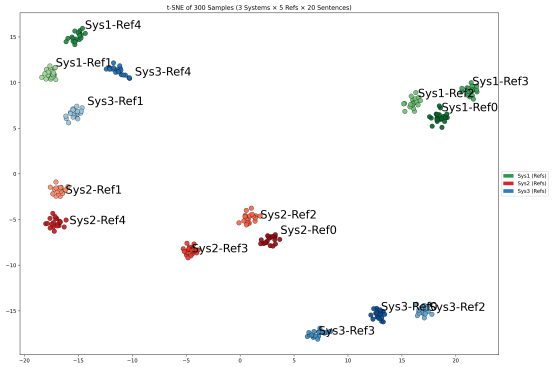

Figure 3: T-SNE speaker embeddings for 20 utterances generated by 3 TTS models and 5 references prompts

6. Conclusion

In this paper we introduced a general methodology of rapid development of TTS models for low-resource languages from available resources. In our experiments we worked on Central Kurdish which is a low-resource language which might favored over other languages due to phonemic consistency. Two new TTS datasets curated and an existing one evaluated. We performed an extensive subjective and objective evaluation of the developed systems. We demonstrate that audiobook data reach competitive results in comparison with high quality studio data. Also we proposed an approach on using these models for low-resource speech translation by finding the best scenario of data generation.

The reported results demonstrate the strong potential of the developed TTS models for auxiliary tasks such as speech translation. Future work will focus on enhancing synthetic dataset diversity to improve model robustness and generalization. This includes broadening linguistic diversity through automatic translation, enriching prosodic diversity via varied speaking styles and emotions, and increasing speaker diversity. The availability of multiple TTS systems with voice cloning and prompt-based

prosody control enables systematic synthesis of diverse speech data. Building on this, we aim to define diversity metrics that guide the automatic generation of task-optimized datasets, supporting more robust and adaptable multilingual TTS and speech translation systems.

7. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research, the innovation programme under the Marie Sklodowska-Curie grant agreement No 101007666, and received funding from the DGA/AID RAPID COMMUTE project. The research reported here was conducted at the 2025 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, hosted at Brno University of Technology (Czeck Republic) and sponsored by Johns Hopkins University. The experiments are done using the LIUM computational infrastructure at Le Mans University. We sincerely thank Malaka Mustafa Soltani, the writer of the female audiobook, for kindly granting us permission to use the content of her audiobook. We also thank Shahin Shahlai, the narrator of the female audiobook, and Fateh, the narrator of the male audiobook, for allowing us to use their voices. We further appreciate the volunteer participants who took part in the perceptual evaluation.

8. Copyrights

References

Hawraz A. Ahmad and Tarik A. Rashid. 2024. Gigant-ktts dataset: Towards building an extensive gigant dataset for kurdish text-to-speech systems. *Data in Brief*, 55:110753.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Mike Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4211–4215.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, et. all., and SEAMLESS Communication Team. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637(8046):587–593.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. In *Interspeech 2024*, pages 4978–4982.

Edresson Casanova, Christopher Dossman Shulby, Andrey Korolev, Amaury Cândido Júnior, André S. Soares, Sandra M. Aluísio, and Moacir Antonelli Ponti. 2023. Asr data augmentation in low-resource settings using cross-lingual multi-speaker tts and cross-lingual voice conversion. In *Proceedings of Interspeech 2023*, pages 1643–1647.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, pages 3830–3834.

David Guennec, Lily Wadoux, Aghilas Sini, Nelly Barbot, and Damien Lolive. 2023. Voice cloning: Training speaker selection with limited multi-speaker corpus. In *12th ISCA Speech Synthesis Workshop, SSW 2023, Grenoble, France, August 26-28, 2023*, pages 170–176. ISCA.

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2022. The VoiceMOS Challenge 2022. In *Interspeech 2022*, pages 4536–4540.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022 -*

[7] https://huggingface.co/datasets/aranemini/tts4all-ckb-dataset

[8] https://huggingface.co/aranemini/ckb-f5-tts

2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6367–6371.

Chanwoo Kim and Richard M. Stern. 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Interspeech 2008*, pages 2598–2601.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

Zhaolin Li, Yining Liu, Danni Liu, Tuan Nam Nguyen, Enes Yavuz Ugan, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. 2025. KIT's low-resource speech translation systems for IWSLT2025: System enhancement with synthetic data and model regularization. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 212–221, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Johannes Abraham Louw. 2023. Cross-lingual transfer using phonological features for resource-scarce text-to-speech. In *12th ISCA Speech Synthesis Workshop (SSW2023)*, pages 55–61.

Florian Lux, Julia Koch, Sarina Meyer, Thomas Bott, Nadja Schauffler, Pavel Denisov, Antje Schweitzer, and Ngoc Thang Vu. 2023. The ims toucan system for the blizzard challenge 2023. In *18th Blizzard Challenge Workshop*, pages 40–45.

Aso Mahmudi and Hadi Veisi. 2021. Automated grapheme-to-phoneme conversion for central kurdish based on optimality theory. *Computer Speech & Language*, 70:101222.

Aso Mahmudi, Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2019. Automated kurdish text normalization.

Pauline Mas, Natacha Miniconi, Kevin Vythelingum, Meysam Shamsi, Aghilas Sini, Lu Zuo, Elias Okat, and Marie Tahon. 2025. VO2Lium : Voxygen and LIUM contribution for Blizzard 2025. In *The Blizzard Challenge 2025*, pages 18–23.

Natacha Miniconi, Meysam Shamsi, Aghilas Sini, and Anthony Larcher. 2025. Using a deepfake classifier to rank speech synthesis quality. In *The Speech Synthesis Workshop - ISCA*.

Tomoya Mizumoto, Atsushi Kojima, Yusuke Fujita, Lianbo Liu, and Yui Sudo. 2025. Is Synthetic Data Truly Effective for Training Speech Language Models? In *Interspeech 2025*, pages 1808–1812.

Mohammad Mohammadamini, Daban Jaff, Sara Jamal, Ibrahim Ahmed, Hawkar Omar, Darya Sabr, Marie Tahon, and Antoine Laurent. 2025a. Kuvost: A large-scale human-annotated English to Central Kurdish speech translation dataset driven from English common voice. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 106–109, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Mohammad Mohammadamini, Aghilas Sini, Marie Tahon, and Antoine Laurent. 2025b. Scaling pseudo-labeling data for end-to-end low-resource speech translation (the case of Kurdish language). In *Interspeech 2025*, pages 898–902.

Sabat Salih Muhamad, Hadi Veisi, Aso Mahmudi, Abdulhady Abas Abdullah, and Farhad Rahimi. 2024. Kurdish end-to-end speech synthesis using deep neural networks. *Natural Language Processing Journal*, 8:100096.

Alex Peiró-Lilja, Guillermo Cámbara, Mireia Farrús, and Jordi Luque. 2022. Naturalness and Intelligibility Monitoring for Text-to-Speech Evaluation. *Proc. Speech Prosody*, pages 445–449.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Yu Pu, Xiaoqian Liu, Guangyu Zhang, Zheng Yan, Wei-Qiang Zhang, and Xie Chen. 2025. Empowering Large Language Models for End-to-End Speech Translation Leveraging Synthetic Data. In *Interspeech 2025*, pages 26–30.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori,

and Yoshua Bengio. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

Jaffer Sheyholislami. 2015. *The Kurds: History, Religion, Language, Politics*, chapter Language Varieties of the Kurds. Austrian Federal Ministry of the Interior.

Jaffer Sheyholislami. 2021. *The Cambridge History of the Kurds*, chapter The History and Development of Literary Central Kurdish. Cambridge University Press.

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

Marie Tahon, Gwénolé Lecorvé, Damien Lolive, and Raheel Qader. 2017. Perception of expressivity in TTS: linguistics, phonetics or prosody? In *Statistical Language and Speech Processing (SLSP)*, volume 10583, page 262.

Hadi Veisi, Hawre Hosseini, Mohammad MohammadAmini, Wirya Fathy, and Aso Mahmudi. 2022. Jira: a central kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon. *Lang. Resour. Eval.*, 56(3):917–941.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi H. Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. 2024. Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8.