Research Article

Let Me Read You a Story in Your Mother Tongue! Kids Story Reader in Sorani Kurdish (Central Kurdish)

Bala Farhad[1,2], Hossein Hassani[2]

1. University of Kurdistan Hewler, Erbil, Iraq; 2. University of Kurdistan Hewlêr (UKH), Erbil, Iraq

Text-to-speech (TTS) synthesis is the technique of generating synthetic speech from input text. Developing a TTS system for Sorani (Central) Kurdish is a challenge due to the lack of resources for the language. In this research, we assess the development of a storytelling TTS system in Sorani Kurdish for children aged five to ten by comparing two different TTS methodologies and technologies. We select proper children's storybooks to build a Sorani storytelling TTS system. We used two female narrators to narrate the stories, based on which we created the necessary datasets that the two chosen TTS frameworks use. The collected records are nearly seven hours long and are pre-processed, segmented, and aligned with the transcribed texts. The final dataset includes approximately five hours of speech consisting of 4149 speech segments and 34,523 words. We use Tacotron2 and Variational Inference with adversarial learning for end-to-end Textto-Speech (VITS) frameworks. We evaluated the results objectively and subjectively. The results indicate that the sound quality of the VITS-based model and its understandability outperforms the Tacotron2 model by a mean opinion score of 3.41 versus 1.91 for Tacotron2. We attribute that to two factors: the amount of training data and the training period.

Corresponding authors: Bala Farhad, bala.farhad@ukh.edu.krd; Hossein Hassani, hosseinh@ukh.edu.krd

1. Introduction

Text-to-speech (TTS), also referred to as speech synthesis, is a technology that delivers synthetic speech by converting input text into output speech. The speech generated through speech synthesizers is evaluated based on three main factors: the naturalness of the voice, the intelligibility, and the speech's

 doi.org/10.32388/NT5LBM

expressiveness (Kelechava, 2015). Speech synthesis aims to develop a machine to produce a natural-sounding voice that is highly intelligible in the desired accent, language, and voice. The two main parts of a TTS synthesis system are the natural language processing module and the digital signal processing module.

The generation of speech from a machine mechanically or electronically represents a new concept for people in terms of how a machine can generate speech (Lemmetty, 1999). Compared to the early attempts at building speech synthesizers, where the machines could only produce speech in a synthetic voice and only generate words or short sentences (Ning et al., 2019), speech synthesis technologies nowadays produce state-of-the-art speech in terms of naturalness and intelligibility. The evolution of speech synthesis technologies is due to the great advances in natural language processing technologies (Ning et al., 2019). However, this is not the case for less-resourced and less-studied languages such as Kurdish (Hassani, 2018).

In this research, we investigate the efficiency of a TTS for the Kurdish language while we consider its low-resourced status. It restricts its investigation to the efficiency of the system in reading children's stories in Sorani Kurdish.

The rest of the paper is organized as follows. Section 2 reviews the literature and the related work. Section 3 presents the method that the research follows. We provide the results and discuss the outcome in Section 4. Finally, Section 5 concludes the paper and provides some ideas about future work.

2. Related work

As far as the literature shows, the first attempt at building a Sorani TTS system dates back to 2009, (Barkhoda et al., 2009; Daneshfar et al., 2009; Bahrampour et al., 2009). Those works suggested that using diphone-based concatenative speech synthesis yielded the most natural-sounding speech compared to other concatenation units, such as allophones and phonemes. Concatenative synthesis, particularly using diphone units, was widely adopted due to its ease of implementation and the naturalness it achieved (Barkhoda et al., 2009; Hassani and Kareem, 2011).

However, despite recent attempts, Kurdish TTS did not improve much in the later years (Muhamad and Veisi, 2022). It means that for the more advanced approaches in TTS, we must look into other languages that have experienced more development in their TTS systems.

Neural network-based models, such as WaveNet (Oord et al., 2016), have shown remarkable performance in generating natural-sounding speech compared to traditional concatenative and statistical TTS models. WaveNet directly models linguistic features and generates waveforms probabilistically. However, the drawback of WaveNet is its slow waveform generation process, as it requires processing individual samples.

Arık et al. (2017) presents an optimized version of WaveNet, called Deep Voice, aiming to develop a complete end-to-end TTS system. Deep Voice demonstrated faster-than-real-time inference but required additional resources, such as audio-text transcription and a phoneme dictionary with duration and fundamental frequency information.

The Tacotron model (Wang et al., 2017), a deep neural network model that generates spectrograms from text, combined with waveform synthesis techniques, offers a complete end-to-end TTS synthesis system. Tacotron2 (Shen et al., 2018), an enhancement of Tacotron, achieved state-of-the-art speech synthesis, and subsequent studies further improved the naturalness of the synthesized speech using Tacotron and WaveNet models. Tacotron2 is also a sequence-to-sequence (seq2seq) model with an attention mechanism that has been proposed to map the input text to Mel spectrograms for speech synthesis. Seq2Seq neutral networks can transfuse an input sequence to an output sequence that may have a different length. Tacotron2 combines the front-end and back-end components of traditional speech synthesis frameworks into a unified framework. Two basic processes are usually required to generate the speech in a TTS system utilizing a seq2sqe model. First, a frequency representation of the text is generated (the Mel Spectogram), and then a waveform is generated from this representation. Usually, Tacotron2 is combined with a vocoder (a contraction of the term voice coder (Dolson, 1986)) to synthesize the Mel spectograms of the trained Tacotron2 model (Kwon et al., 2019; Zhao et al., 2023). Various open-source implementations of Tacotron2-WaveNet have emerged, enabling researchers and developers to experiment with TTS systems for different languages. These implementations, such as Tacotron2-WaveGlow, have achieved audio quality comparable to professionally recorded speech.

Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) (Kimetal., 2021) implements a conditional variational autoencoder (VAE) with adversarial learning to build an end-to-end TTS synthesis system, which utilizes the Generative Adversarial Network (GAN) to generate improved voice from the text. The general architecture of this model consists of a posterior encoder, a previous encoder, a decoder, a discriminator, and a stochastic duration predictor. The posterior encoder and discriminator are only used in the training phase, not in the inference phase. This technique utilizes

variational inference, which is then supplemented with normalizing flows and an adversarial training procedure. The end result is an increase in the expressive capability of the generative modeling. Using uncertainty modeling over latent variables and a stochastic duration predictor, this method expresses the natural one-to-many relationship by allowing a text input to be pronounced in various ways with varying pitches and rhythms. This is made possible by the combination of these two components.

While work on TTS for some languages is well-studied, the adoption of their approaches would not gain a similar quality output for the low-resourced languages. Therefore, the TTS studies on low-resourced languages could assist the current research more. For example, Latorre et al. (2019) demonstrates that in acceptable quality by increasing the number of narrators in the absence of large datasets. Furthermore, Studies on speech synthesis with limited data have shown varying results, with some indicating that a smaller dataset with high-quality recordings can achieve satisfactory results (Podsiadlo and Ungureanu, 2018). Also, Tu et al. (2019) shows the data of a few hours length can provide and acceptable quality.

The literature indicates that previous studies have employed both traditional (concatenative, unit selection, and statistical) and modern (deep learning, end-to-end) approaches for TTS synthesis. Modern approaches, particularly deep learning-based models, have shown promising results but often require large amounts of data. However, high-quality smaller datasets have also demonstrated satisfactory results. Given the lack of recent work on TTS synthesis for Sorani, this research applies an end-to-end approach using a fairly small but high-quality dataset.

Because we aimed at storytelling for children, we also studied the literature to find what parameters the specialists suggest to consider a book to be appropriate for children up to ten years old. According to various studies (Nodelman, 1988; Nodelman et al., 2017; Maryland Library Resource Center, 2023), books with illustrations, which are called picture books, are the best fit for our target age group. This type of book use illustrations to tell stories that children relate to and learn emotional intelligence, such as kindness, empathy, and forgiveness from life lessons, relationships, morals, and their culture. These books usually contain a few illustrations with small paragraphs in the third person and contain plenty of dialogue.

3. Method

This section explains the method we follow in conducting this research. It describes data selection, data collection, pre-processing of the collected data, the quality control of the created dataset in both text and

recording formats, the TTS models creation, the frameworks, and their environment configuration, and testing and evaluation approaches we use during the experiments.

3.1. Data Collection and Preparation

We collect short story books in Sorani for children between 5-10 years old. The collected stories are narrated by professional speakers in a storytelling-animated manner suitable for children. Then, recorded stories are validated manually to re-record the stories that don't pass the validation process. To prepare the transcription of the recorded stories, the collected stories, which we expect to be in PDF or image formats, are given to an optical character recognition (OCR) system and then manually reviewed to fix the possible errors of the OCR output. Figure 1 shows the process followed.

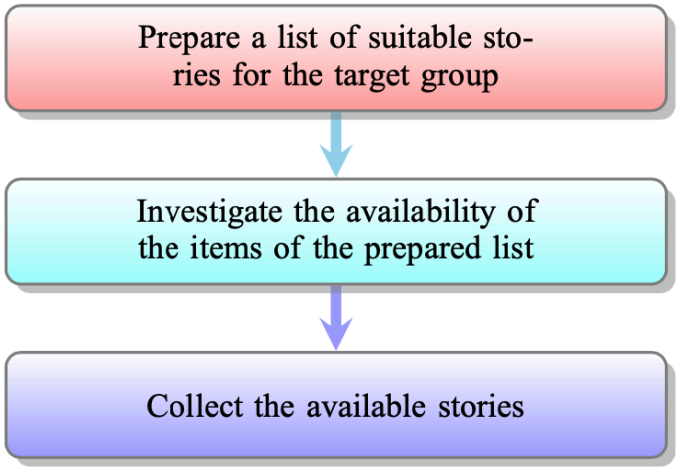


Figure 1. Data Selection Process

During the last step of Figure 1, the following points are also considered:

- The story needs must be in Sorani.
- The topic of the story must be suitable for the target group.
- The stories must be short comprising of four to five paragraphs.
- The stories must be accompanied with illustrations.

3.2. Audio Recordings

The collected data is then converted to an editable text file to be ready for data pre-processing.

doi.org/10.32388/NT5LBM

According to the reviewed literature, we employ professional female narrators to narrate and record the stories using high-quality devices. The essential factors and issues of this step are articulated as follows:

- Recordings
  - To have a studio-quality recording by using professional recording devices. The story transcripts should be read consistently and in a compatible manner with the usage of the targeted TTS environment. The legal and ethical issues regarding the publicity of the data should also be considered.
- Segmentation
  - To have a studio-quality recording by using professional recording devices. The story transcripts should be read consistently. The selection of appropriate segmentation criteria is not straightforward. We are interested in synthesizing shorter utterances in the sequence of sentences. Therefore, sentence-based segmentation appears to be appropriate for this case. On the other hand, it is preferable to have a balanced distribution of segment duration. However, according to the literature, separating sentences at the phrase-level frequency results in an asymmetrical length distribution and in a compatible manner with the usage of the targeted TTS environment. The legal and ethical issues regarding the publicity of the data should also be considered.
- Alignment
  - Aligning transcripts with recordings introduces yet another set of issues. A forced aligner is commonly utilized because manual alignment is either too expensive or too difficult. The precision of automatic aligners is restricted, which might lead to slightly altered alignments. That also might lead to transcripts containing missing or additional compared to the corresponding recordings; therefore, models trained on this kind of data frequently skip the first or final words of the input. Stability issues might arise as a result of either leading or trailing silence. Despite the benefits of automated aligners, we align the data manually because, as far as we know, currently Sorani Kurdish aligners are not available.

Figure 2 shows an overview of the tasks we follow in the audio recording phase.

3.3. Pre-processing the Data

Tacotron2 and VITS training are done in an end-to-end fashion; hence, the models require the data to be in a <text, audio> pair. The data preparation for these models is composed of two main tasks:

- Segmenting the audio recording to be between 0 and 10 seconds long.
- Aligning the segmented recordings with their corresponding text.

Before beginning with segmenting the audio recordings, we decide on a labeling approach to follow for each segmented recording file. We title each segmented recording to have four sections of information. The starting section starts with the letter B followed by a one-digit number indicating the number of the book the data was extracted from. The second section; contains the speaker identity that recorded this file. In a separate file, we kept the speaker information in a table and gave each speaker an ID. This ID is used in this section. In the third section, we add the number of the story in that book. Finally, in the fourth section, we have a three-digit number to indicate the number of the segmented file. Table 1 presents how the labels are structured.

Various tools are available to align text and audio for different languages. However, we couldn't find a tool that supports Sorani Kurdish Sorani, so we did the alignment manually.

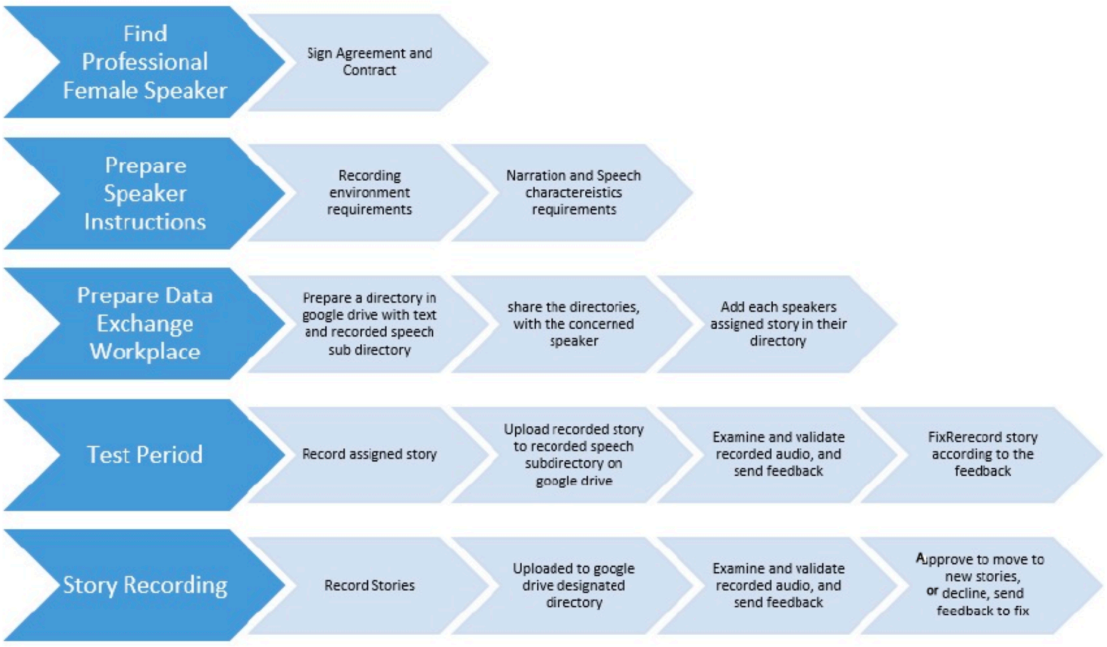


Figure 2. Story Audio Recording Task Flow Diagram

doi.org/10.32388/NT5LBM

Section 1	Section 2	Section 3	Section 4
Book number	Speaker ID	Story number	Segmented audio file number
Example: B1-S1-Story1-001			

Table 1. Labeling Approach Followed for Segmenting Audio Recording Files.

3.4. Preparing Dataset

We use two frameworks, Tacotron2, and the VITS model, using the Coqui-AI library (Gölge, 2022). These frameworks have different data preparation steps and different implementation steps.

For Tacotron2, We divide the recorded stories into 1 to 10-second segments, using Audacity (Team, 2022) and align them with corresponding transcribed text in a comma-separated-value (CSV) file.

For Festival, we split audio files into chunks of five to six words long and labeled them to create a tab-separated-value (TSV) file that contains the audio file labels and their corresponding transcribed text. Then, we construct a lexicon, transliterate them into their corresponding IPA format, and prepare a JavaScript Object Notation (JSON) file for the phones.

3.5. Testing and Evaluation

The quality of the model is next evaluated in terms of intelligibility and naturalness by two target audiences, children and adults, including the parents and instructors of the children. The sentences used during the evaluation process are not included in the dataset used for training. These testing procedures help in forming conclusions regarding the suggested method and provide insight for future research.

The collected stories are initially reviewed by the children's teachers to ensure they are appropriate for our target age range. Then, in order to evaluate the process of story recording, several sample recordings are analyzed and compared to the story recording criteria stated in the section on data collection. Based on the findings of the evaluation, the procedure has been fine-tuned and enhanced to produce recordings that meet our criteria and are of high quality.

Our model's performance and quality can be measured with an objective and subjective evaluation. There are two primary techniques for objective evaluations. First, the character error rate (CER) can be

determined using an automatic speech recognition (ASR) model. Second, the MCD values for the two models are compared. MCD is a measure of the difference in two sequences of MCD. In our experiment, we also relied on subjective evaluation by evaluating the Mean Opinion Score (MOS) to evaluate the quality of our model with two distinct targeted audiences.

For the subjective evaluation, several samples are played to the participants, who will then complete a questionnaire evaluating the naturalness and intelligibility of each sample. To accommodate the wide range of ages represented among the participants in our study, we devised two separate questionnaires, adhering to the MRS/ESOMAR standards for research involving children (MRS, 2014; MSR, 2018). Then, they inform a conclusion regarding what must be updated or changed to improve the model. All input is collected at the end, and our proposed method is evaluated based on this.

4. Experiments, Results, and Discussion

This section reports the experiments, presents the results, and discusses the outcome of related evaluations.

4.1. Data Collection

We considered available online guidelines and double-checked with kindergarten teachers through in-person interviews to set the criteria for selecting suitable books for children who fall in the age group of our research. The following is a list of the criteria that have been established for the textual corpus collection:

- The narrative must be written in Sorani dialect.
- The story's subject must be appropriate for our target age group.
- The stories must be concise, with no more than 4-5 paragraphs per piece of writing.
- A minimum of two illustrations is required for each story.

Finding online sources from which to collect stories that meet the criteria for the corpus was the first step in the collection process for the corpus. There were only a handful of Kurdish stories found. For instance, LanguageLizard (2021) offers a few English-Kurdish storybooks for children who are interested in learning a second language. However, they only distribute hard copies, and it is not available in this country, so acquiring them was incompatible with our experiment's timeline. During the time that we conducted this experiment, we were unable to discover any online resources that were open to the public

and could be used to collect stories. Consequently, we began our search for picture books in the various bookshops located throughout the city.

We were able to collect multiple books from various cities that met the requirements outlined previously. The books were examined by kindergarten teachers to verify that they were appropriate for our target age range. We collected eight books, but only six books were accepted, and the remaining two were not recommended by the kindergarten teachers since the stories were lengthy, contained more difficult vocabulary, and contained few illustrations. Figure 3 presents the coverage of the collected books used to create the corpus.

For the conversion we use the OCR trained for Sorani Kurdish by Idrees and Hassani (2021). The mentioned OCR has a few requirements which we had to apply to our data before passing it to the model. The requirements included the following points:

- Scanned images should be a single page and cropped to text only.
- Scanned documents should be ≧300 dots per inch (DPI).
- Preferably black text on a white background.
- Portable network graphics (PNG) image format is acceptable for the system.

This process follows the steps shown below in Figure 4.

The first step is to scan the stories collected. Before scanning, we configured the scanner to scan with 300 DPI in line with the requirements of the OCR model. Then, scanned images were manually converted to black and white, and then we cropped the images to text only and exported them in PNG file format since the OCR model showed better performance by appending multiple images into one text file. The stories were scanned and sectioned based on the book number the stories were collected from. Once the data were ready, the scanned images were OCR-ed in line with their book number, and the result was appended to the text files. Afterward, the resulting OCR-ed stories were manually reviewed to find the incorrectly OCR-ed text and replace it with the correct word from the scanned stories, for example, to replace by .

4.1.1. Recording and data prepartion

High-quality TTS synthesis systems need a high-quality dataset. Building a high-quality TTS system for Sorani is challenging because of the lack of resources in this language. However, based on our findings and other researchers' experiments on high-quality TTS for low-resource languages, we found that it's

possible to build a decent TTS system with a smaller dataset on the condition that the dataset is phonetically balanced and includes good-quality recordings. Consequently, we considered the synthesis system's requirements and our target age group of children between the ages of 5 and 10, as mentioned in the objectives. Listed below are the second set of requirements for collecting the speech corpus:

- The stories must be narrated by trained professionals.
- Speakers will narrate the stories in an animated manner appropriate for children's story reading.
- The recordings must be made in a quiet environment with a 22050 KHz sample rate in WAV format.

Following setting the standards for the speech collection, we started looking for professional speakers who were appropriate for our task and were able to maintain a balanced voice while speaking for extended periods of time. In the end, we decided to hire two female speakers. One of the speakers worked as a freelance digital creator and narrator in Sulaymaniyah, and the other worked as a novel narrator in one of the local radio stations in Erbil. The speaker in Sulaymaniyah agreed to record four hours of speech over the course of 25 days in exchange for signing a contract for 300 US dollars. The speaker in Erbil agreed to record three hours of speech over the course of 40 days in exchange for signing a contract for 100 US dollars. Initially, we were only able to find a suitable speaker in Sulaymaniyah, but the cost of recording the speech was prohibitively expensive, and there were no other speakers who were suitable for the task

Figure 3. List of Collected Books for the Corpus

doi.org/10.32388/NT5LBM

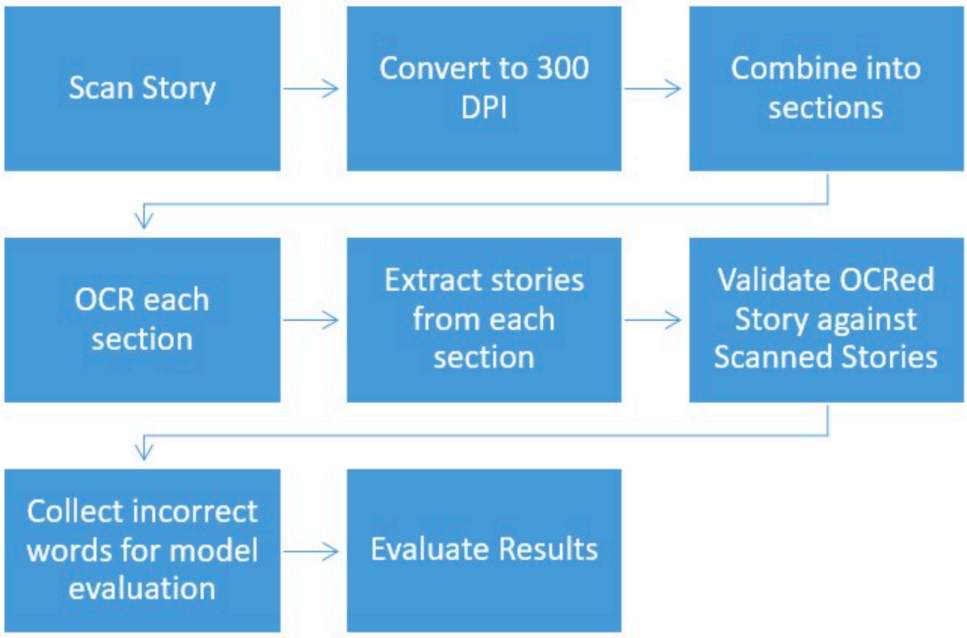


Figure 4. Story OCR-ing Process Flow Diagram

and willing to perform it. After hiring that speaker, we were able to find an Erbil-based speaker willing to complete the task at a lower cost. Because our experiment required the collection of a minimum 6-hour dataset, a second speaker was hired to record the remaining corpus.

To ensure the quality of the recordings, the speakers received instructions about how to narrate the stories, the recording environment and settings required, and how to submit each story. The speakers were asked to record in a quiet environment in their studio and record the stories with a sample rate of 22050 kHz and export the files in a WAV (16 PCM) file format based on the requirements of the training frameworks (Tacotron2 and VITS). Furthermore, both speakers went through a test period of 3-5 days to ensure the collected data met the specifications required for high-quality speech.

In this test period, each speaker was given a couple of stories to record and then shared with us to validate the recordings and send feedback to the speaker to improve the recorded speech. In the beginning, the speakers would read the story too fast or not in an animated manner, and we also received recordings with background noise or speaker breathing sounds. The location of the noise was identified in each case and relayed to the speakers in order to remove it, and the speakers were asked to re-record the story if it didn't match our criteria.

After the testing period concluded, the collected stories were divided, and each section was assigned to a different speaker to record within the time frame specified in the signed contract. During that period, every day, the recorded speech was uploaded to a shared drive folder on the cloud for validation. If the validator accepted the recordings, the speakers moved on to record new stories. If the recorded stories included background noise, incorrectly pronounced words, or didn't meet the instructions, the validator rejected them and sent them back to the speakers with feedback e for fixing partially or re-recording.

Finally, after finishing all the recording hours for each speaker based on the agreement we had, the speakers were asked to collect the data on a hard drive and share it with us. We did this to make sure the quality of the files was not affected by being downloaded from the shared folder. The recording was carried out in a dedicated space specifically for that purpose. The microphone used was a Rode NT1-A model, which has a reputation for producing high-quality recordings of human speech. It was equipped with a pop shield to mitigate the impact of exhaled air on the microphone.

In total, 209 files containing a total of 34530 words (after transcript revisions) were given as individual files for each narrative. The first speaker produced 115 recordings and the second 94, which together made seven ours of speech.

The speech was not delivered in a single file since sequence models are better at aligning shorter utterances than larger ones. Having utterances that range in duration is typically thought of as a goal since it has the potential to enrich the prosodic coverage of the corpus. The statistics regarding the duration of

	Seconds	**Minutes**
Minimum Duration of Segments	41.12	0.68
Maximum Duration of Segments	724	12.1
Average Duration of Segments	120	1.99
Total Duration	25053	417.5

Table 2. Story Duration Statistics Before Pre-Processing

doi.org/10.32388/NT5LBM

the number of recordings contained in this corpus before processing and segmentation are presented in Table 2.

All clips were segmented manually based on the length of the sentence, which is a maximum of 10 seconds. The segmented recordings are aligned with the text manually, and there is a quality check phase to ensure the text and the spoken words are accurately aligned. The following steps are followed to align the data.

The audio recordings are played using Audacity. We checked the sample rate of the recording to ensure it was 220500 KHz to ensure that they were single-channel audio. Otherwise, we used a built-in function in Audacity to split stereo to mono and remove the second channel, leaving single-channel audio with 220500 KHz for the sample rate. Then we listened to the recordings and added labels to the selected segment, then added the corresponding text of that segment to its corresponding label. Once the story was segmented, we used the ”export to multiples” function to export each segment with its label as the file name.

4.2. Data Pre-Processing

The collected data is pre-processed to ensure data consistency. The input to end-to-end speech synthesis is a <text, audio> pair consisting of two components. The first component is a folder with all the audio files for the utterances segmented between 0 and 10 seconds long. The second is a spreadsheet that uses the utterance segment file name and its transcript to identify the utterances. The details of pre-processing activities are given below.

4.2.1. Text Pre-Processing

Text pre-processing began with the preparation of the stories for the OCR model. In our experiment, we used an OCR model that requires input data to be formatted in a specific way. The data had to be in JPG format and black and white with a DPI greater than three hundred. Following the scanning and conversion of the images to fulfill the prerequisites of the OCR model, the images were arranged into six books before being sent off to be OCR-ed. The output of the OCR model was five text files, one for each of the six books from which the stories were collected.

For the OCR process, each story was scanned as depicted in Figure 6. In addition to removing all illustrations from the image, the scanned image was converted to black and white as presented in Figure

6.

4.2.2. Speech Pre-Processing

After the recording was complete, the speaker went through the entire recording to make the following modifications:

- Including brief pauses at the beginning and end of sentences. That is required to provide context for each recorded utterance.
- To execute ”Dynamic Range compression” on the loudness (intensity) of each utterance. The signal is multiplied by a dynamic gain to maintain the signal within a predetermined limit, and this is utilized to make the intensity as uniform as feasible. We decided to choose 12 db, although it is possible to re-export the output with other limits.
- To decrease the length of the pauses in speech that were excessively long. No exact length was agreed upon, but the speaker was provided with comments on how to reduce large pauses in a way that maintains acceptable variability in pause length without jeopardizing the precision of the alignment.

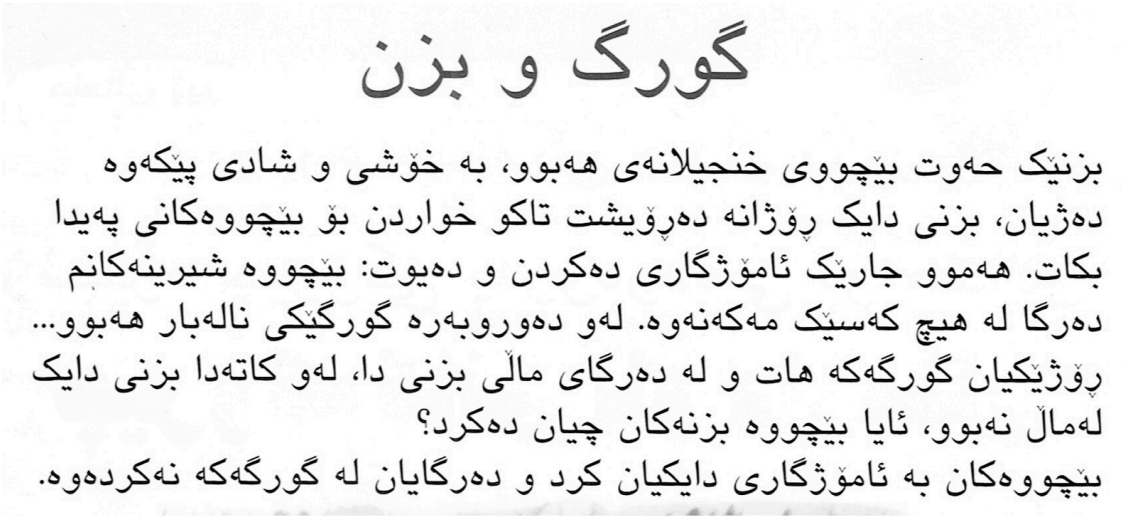
چیرۆکی (١)

گورگ و بزن

بزنێک حەوت بێچووی خنجیلانەی هەبوو، بە خۆشی و شادی پێکەوە دەژیان، بزنی دایک ڕۆژانە دەڕۆیشت تاکو خواردن بۆ بێچووەکانی پەیدا بکات. هەموو جارێک ئامۆژگاری دەکردن و دەیوت: بێچووە شیرینەکانم دەرگا لە هیچ کەسێک مەکەنەوە. لەو دەوروبەرە گورگێکی نالەبار هەبوو... ڕۆژێکیان گورگەکە هات و لە دەرگایماڵی بزنی دا، لەو کاتەدا بزنی دایک لەماڵ نەبوو، ئایا بێچووە بزنەکان چیان دەکرد؟
بێچووەکان بە ئامۆژگاری دایکیان کرد و دەرگایان لە گورگەکە نەکردەوە.

Figure 5. Original Scanned Story

doi.org/10.32388/NT5LBM



گورگ و بزن

بزنێک حەوت بێچووی خنجیلانەی هەبوو، بە خۆشی و شادی پێکەوە
دەژیان، بزنی دایک ڕۆژانە دەڕۆیشت تاکو خواردن بۆ بێچووەکانی پەیدا
بکات. هەموو جارێک ئامۆژگاری دەکردن و دەیوت: بێچووە شیرینەکانم
دەرگا لە هیچ کەسێک مەکەنەوە. لەو دەوروبەرە گورگێکی ناڵەبار هەبوو...
ڕۆژێکیان گورگەکە هات و لە دەرگایماڵی بزنی دا، لەو کاتەدا بزنی دایک
لەماڵ نەبوو، ئایا بێچووە بزنەکان چیان دەکرد؟
بێچووەکان بە ئامۆژگاری دایکیان کرد و دەرگایان لە گورگەکە نەکردەوە.

Figure 6. Edited Scanned Story

doi.org/10.32388/NT5LBM

doi.org/10.32388/NT5LBM

qeios.com

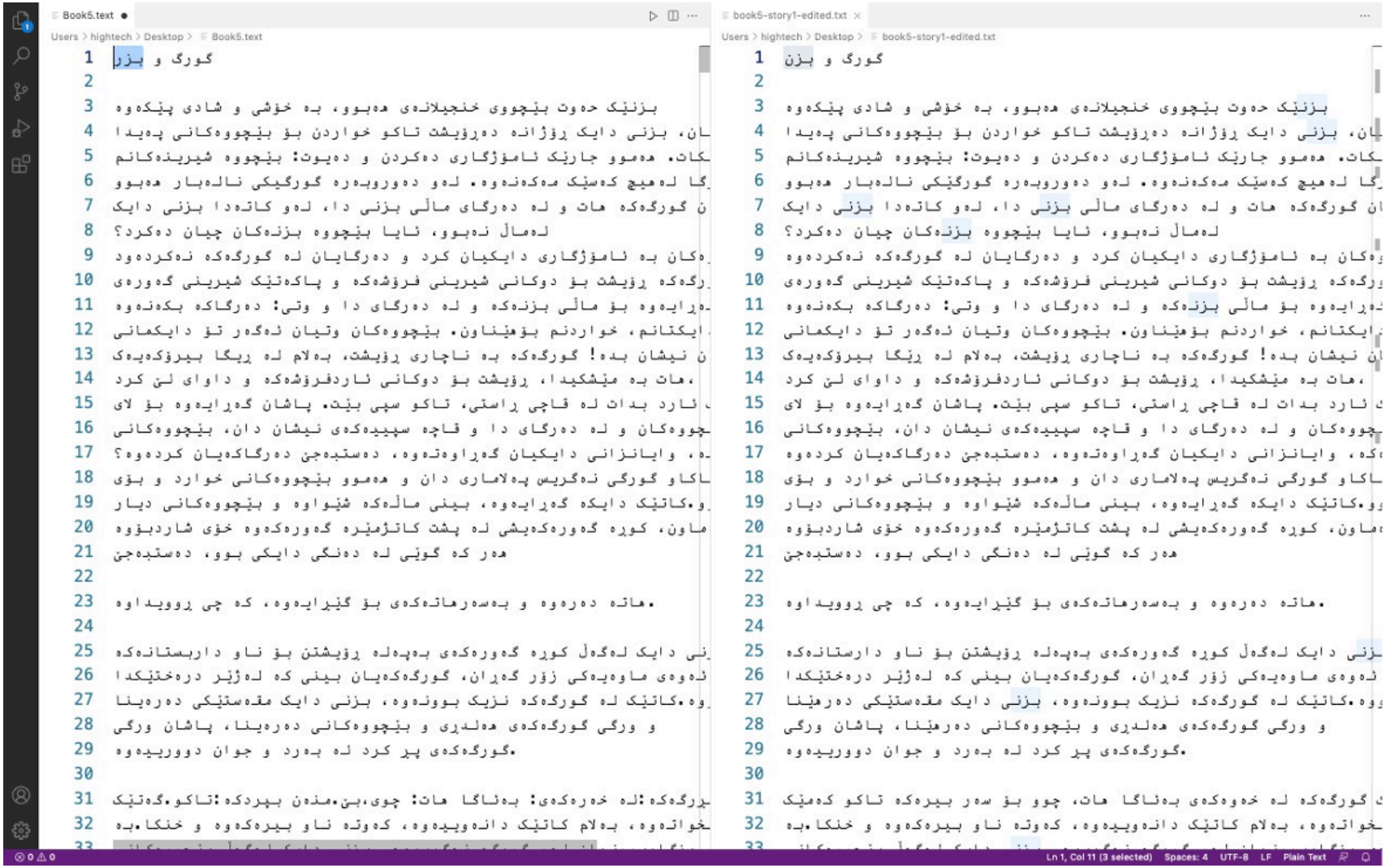


Figure 7. Reviewing OCR-ed Story

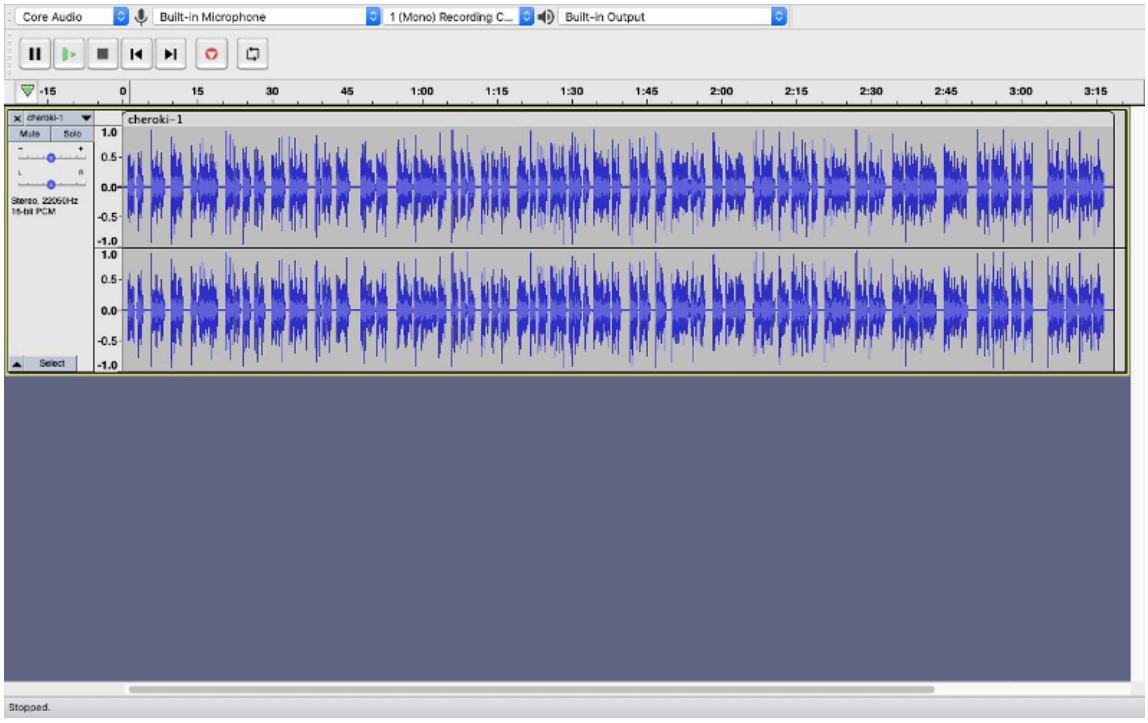


Figure 8. Raw Recorded Stereo Channeled Story

In the literature, segmentation, and alignment are used interchangeably to describe the general procedures of labeling a speech corpus and aligning the utterances with their corresponding transcripts. In this paper, segmentation refers to the annotation of a speech corpus with a sequence of labels derived from the story labels of this corpus, which are generated automatically based on the number of segments. To further refine the segmentation process, it is necessary to identify the beginning and ending points of sentences that are between 0 and 10 seconds long. Microsoft Excel functions are used to create the segment labels from the story labels, but the segmentation of the corpus is done manually in this study.

Alignment is the process of determining the exact timestamps of the borders of the utterances. This can be performed automatically (using speech recognition technologies or an HMM model) or manually by a person whose task it is to conduct segmentation and then alignment manually, which is known to be a time-consuming process. Because there was no available technology for aligning Kurdish (Sorani) transcripts, we had to perform the alignment manually.

doi.org/10.32388/NT5LBM

The Audacity software was utilized throughout the process of segmentation and alignment that we carried out. The recorded stories were first imported into Audacity as raw data. Below is an example of a story that was imported. Figure 8 demonstrates that the audio recording has stereo channels.

In Audacity, there is a built-in feature that allows you to split a stereo track into a mono track. Two mono-channeled audio tracks are created from the stereo audio track, as shown in Figure 9. In order to obtain a single audio track with a mono channel, we must first remove the track that is located at the very bottom.

The utterances are extracted from the audio track. The process of segmentation begins with the selection of an utterance from the track, which is followed by the creation of a label for that utterance within the track. The duration of the track's timeline is presented in the form of hours, minutes, seconds, and milliseconds within the timeline toolbar that is displayed on top of the track. Because we need to divide the track into sections ranging from 0 to 10 seconds for our experiment, the toolbar has been zoomed in so that the track can be viewed in terms of the seconds. Enlarging the timeline's toolbar to seconds also aids in displaying the utterance's waveform clearly. This is depicted in Figure 10

The labels are added to the utterances by first selecting the segment from the track and then pressing "CTRL+B" in order to create the label for the selected segment. See Figure 11 for an example of how the label is added to a segment.

This process is repeated on the remaining audio track and segmented until the end of the file is reached,

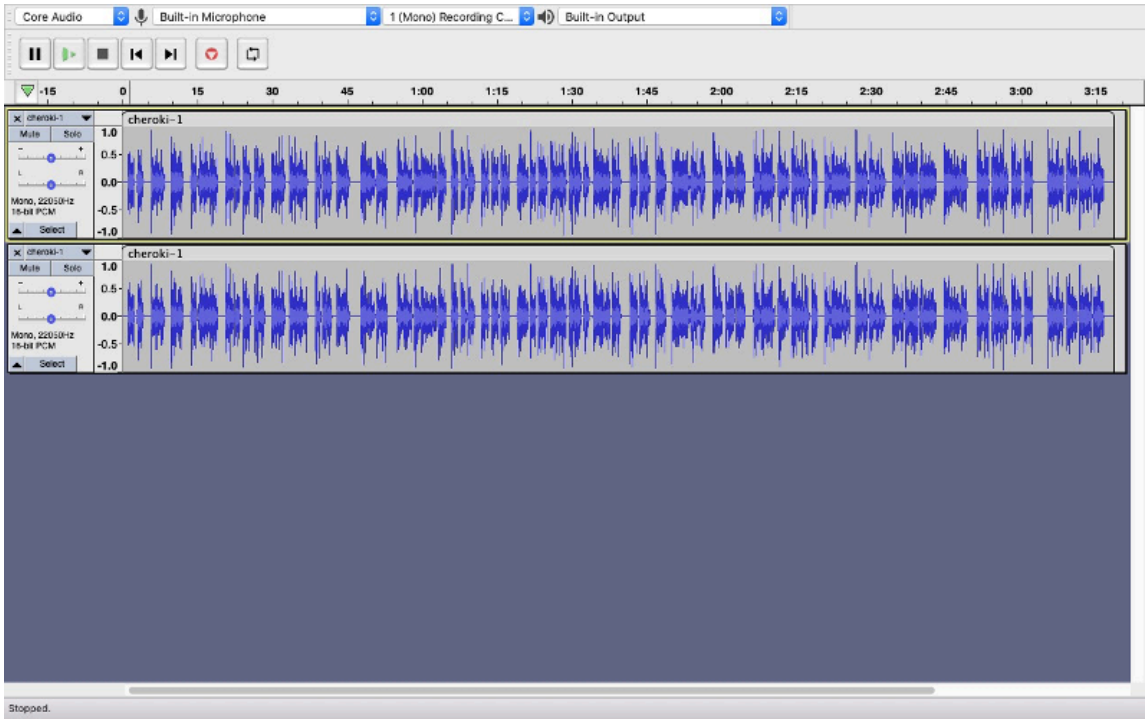


Figure 9. Raw Recorded Story with Separated Stereo and Mono Channel

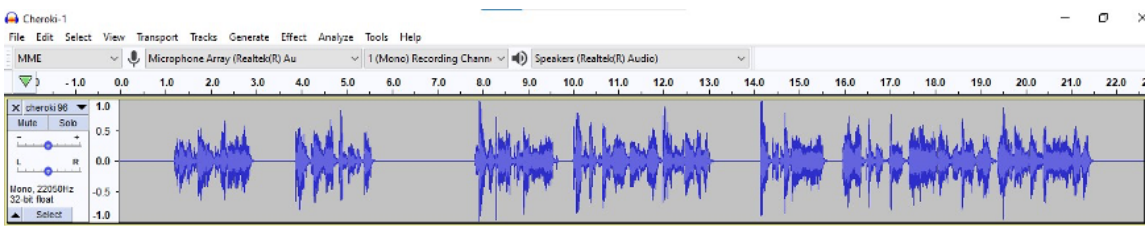


Figure 10. Audacity Timeline Toolbar in Seconds

doi.org/10.32388/NT5LBM

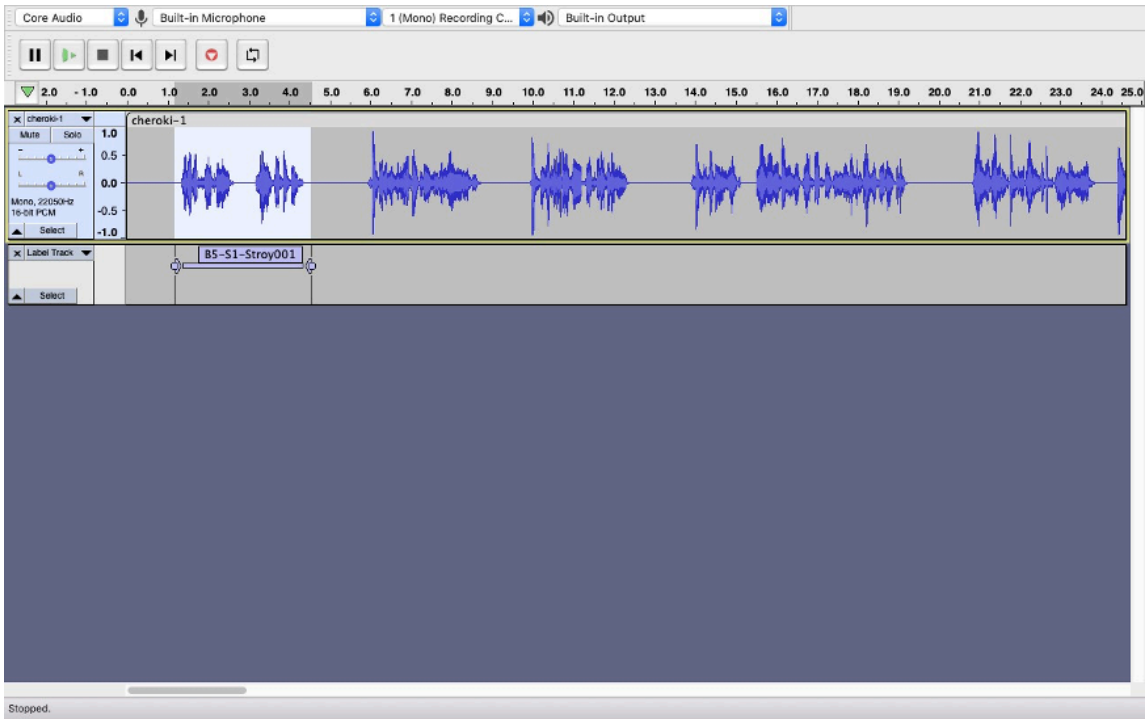


Figure 11. Creating Label for a Selected Utterance on Audacity

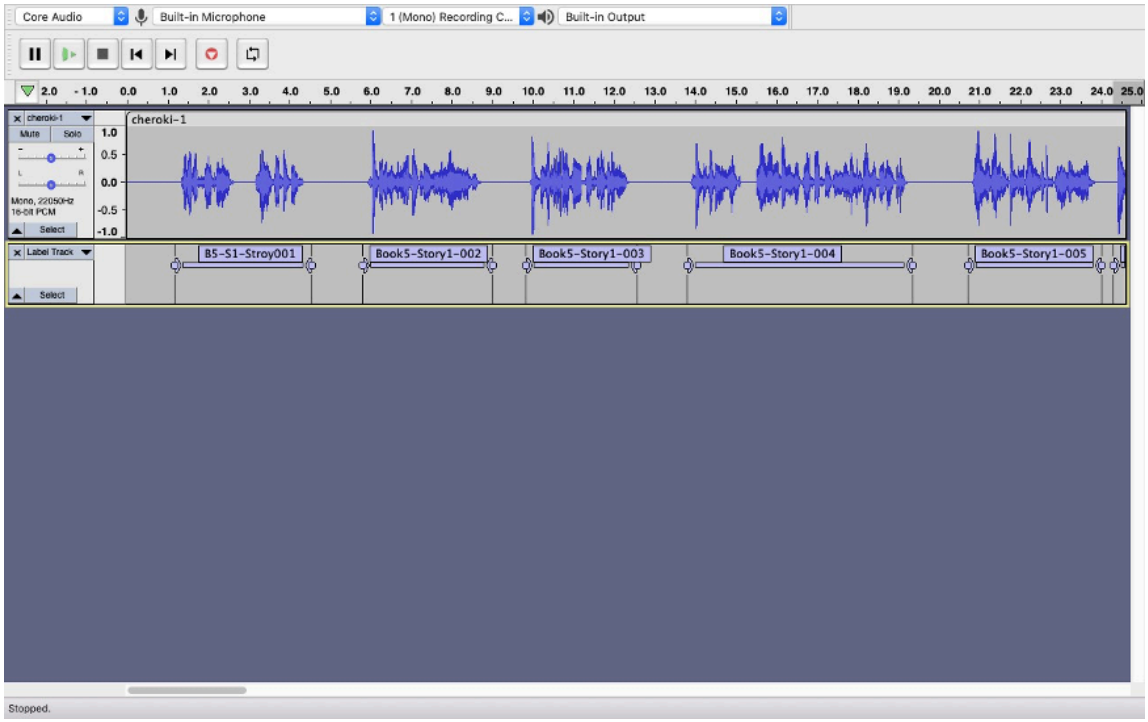


Figure 12. An Example of a Labeled Story

doi.org/10.32388/NT5LBM

	Seconds	Minutes
Minimum Duration of Segments	0.66	0.011
Maximum Duration of Segments	10	0.22
Average Duration of Segments	4.45	0.074
Total Duration of Segments	18459	308
Count of All Segments	4142	
Count of All Stories	209	

Table 3. Pre-Processed Story Duration Statistics

as illustrated in Figure 12.

All of the labels are collected into a single spreadsheet, ”Transcriptions.xlsx”. This sheet is used to keep track of the labels for each segmented utterance and the transcripts that correspond to those utterances. After the audio file is segmented and labeled, the next step is to listen to each utterance and add the utterance transcripts corresponding to those labels. Figure 13 shows how the transcripts are added to the labels.

4.2.3. Collected Data

Table 3 illustrates the duration of the utterances after processing, cleaning, and segmentation of the recorded stories.

During the data pre-processing phase, segments of unnecessary silence and noise are removed. This reduces the collected dataset size from 417 minutes to 308 minutes of recordings. Figure 14 depicts a comparison of the collected data before and after pre-processing.

Removing the pauses and lengthy silences at the beginning and end of the recordings, as well as the noisy portions of the recordings, reduced the dataset by 110 minutes because the stories were recorded in a storytelling manner. Figure 15 gives the amount of data removed during the data pre-processing phase in comparison to the data collected originally.

Nearly 35K words and nearly 5 hours' worth of speech can be found in 4142 segments from the total of 209 stories. Each segment is approximately 5 seconds long and contains 8 words on average. Figure 16

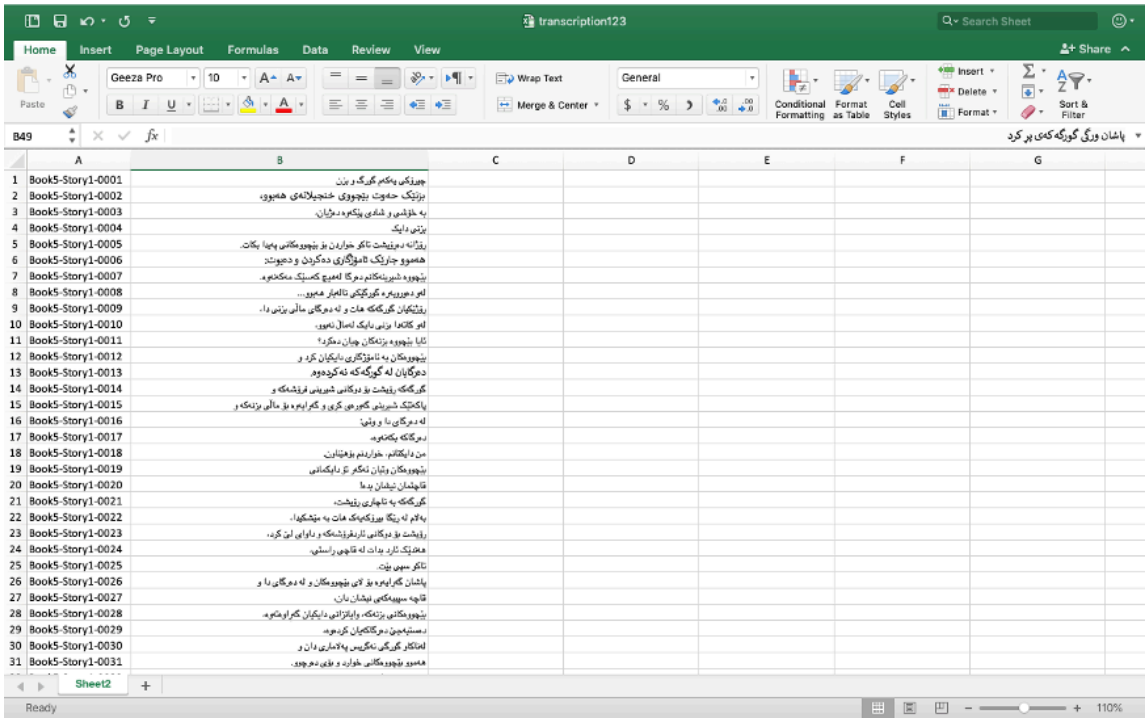


Figure 13. Transcription Sheet to Map Labels and Utterance Transcripts

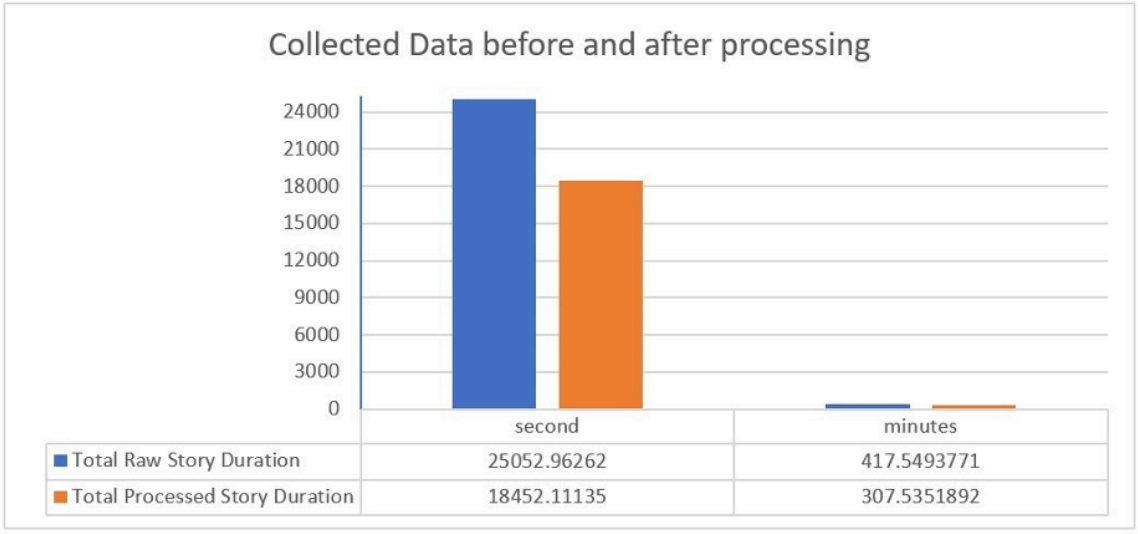


Figure 14. Collected Data Before and After Pre-Processing

doi.org/10.32388/NT5LBM

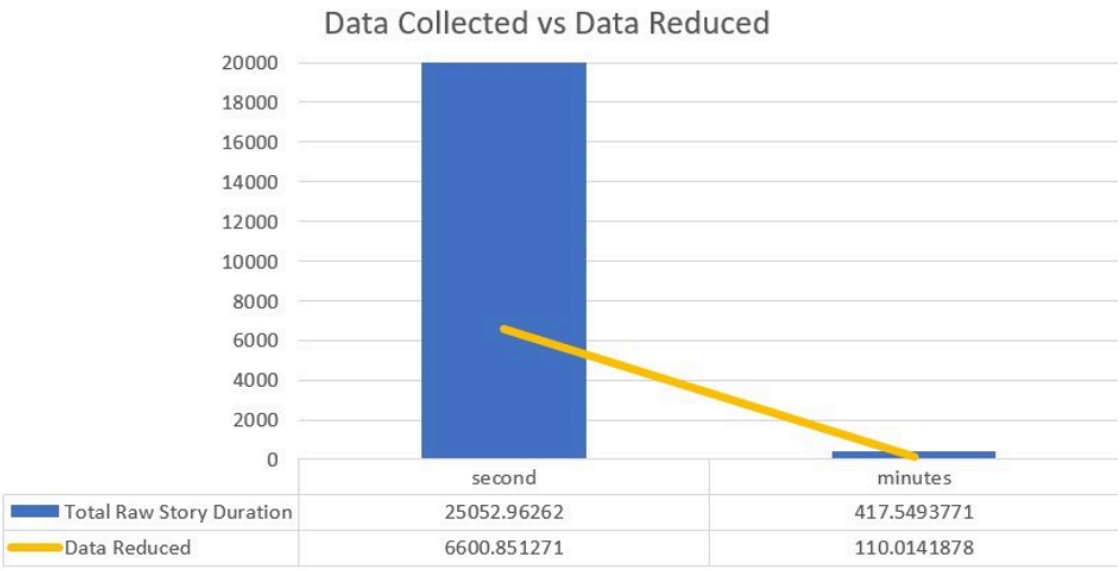


Figure 15. Data Collected vs Data Reduced

	SUM	AVERAGE	MIN	MAX
Duration (Seconds)	18452	5	0.66	10
Words	34523	8	1	25
Segments	4142	22.75274725	5	151
Stories	209	41.8	7	87

Table 4. Dataset Statistics

shows that the number of words increases as the length of the segments increases. The left vertical axis of the graph represents the number of segments per story, while the right vertical axis represents the number of words per segment. Table 4 provides a summary of the size of the collected dataset in terms of duration in seconds, number of words, number of segments, and number of stories.

4.3. Experiments

A forced alignment tool like Aeneas (readbeyond, 2020) and Montreal Forced Aligner (MontrealCorpusTools, 2021) is utilized for this task since doing it is manually often time-consuming and difficult to do.

4.3.1. Training

Training end-to-end models are time-consuming and computationally intensive. To prepare for the training environment, we had to fulfill certain conditions. The acquired machine had the following specifications:

- Ryzen 9 3950x 16-Core 32-Thread CPU
- RAM 32GB 3200 MHz
- NVMe Memory 512 SSD and 1TB HDD memory
- PCU Gigabyte 850W
- NVIDIA 1080ti GTX GPU with 11 GB memory

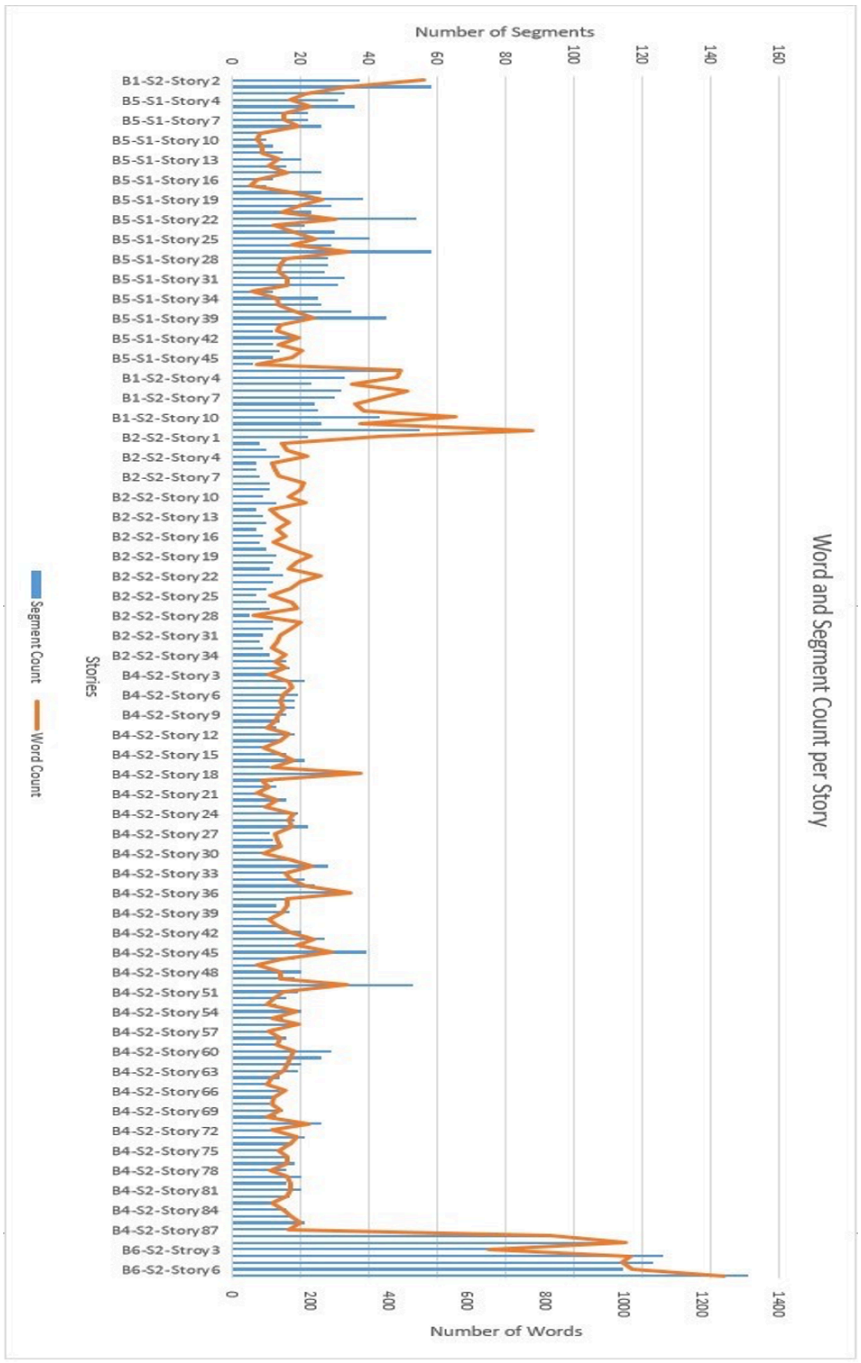


Figure 16. Word and Segment Count per Story

doi.org/10.32388/NT5LBM

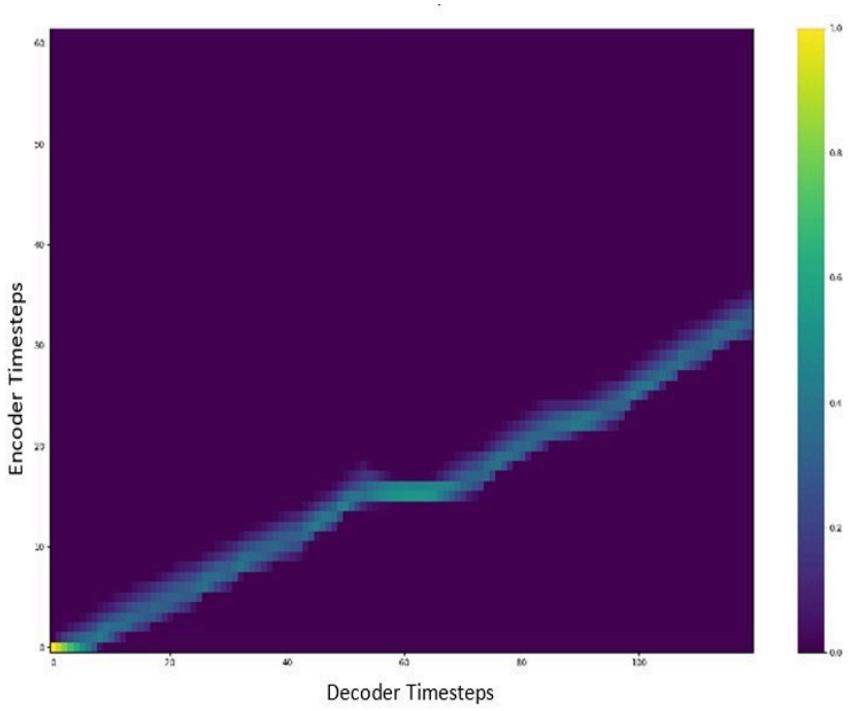


Figure 17. Predicted alignment of Tacotron2 model at 30000 steps

We use the Coqui-ai library to implement Tacotron2 and VITS frameworks. Coqui-ai TTS is a library for generating advanced TTS models. Gölge (2022) is based on the most recent research and was created to provide the optimal balance between training simplicity, speed, and quality. Coqui-ai TTS provides pre-trained models and is now utilized in over 20 languages for product development and research projects. The Software environment was as follows:

- Ubuntu 20
- Python3.8
- CUDA Toolkit 10.1
- cuDNN library compatible with the CUDA version installed
- eSpeakNG
- Git for cloning the project
- torch, torchvision, torchaudio
- TensorBoard

4.4. Training Results

All the segmented audio files with a sampling rate of 22050 kHz, the scripts we use for inference from the trained models, the audio files generated in training and inference, as well as the training logs of the models for the TensorBoard reports, the trained models and pre-trained models were put into related folders.

The model parameters define and represent a model in Machine Learning or Deep Learning. Hyperparameters regulate the learning process and decide the model parameter values a learning algorithm ultimately learns. Unless stated differently, the setting of hyperparameters is identical for both frameworks. The Tacotron2 model is trained with a batch size of 16, 3000 epochs, and a learning rate of 0.01, while the VITS model is trained using a batch size of eight, 4000 epochs, and a learning rate of 0.00001.

The Tacotron2 model was trained for a total of 125,000 steps over the course of 17 days and 14 hours. Figures 17, 18, and 19 illustrate the progression of the predicted alignments produced by the trained model. The total training loss for Tacotron2 was decreased from 3.4735 to 0.93240. The subjective evaluation during training revealed that the Tacotron2 model did not perform as expected. Likewise, the synthesized speech was somewhat natural-sounding but not as understandable as expected.

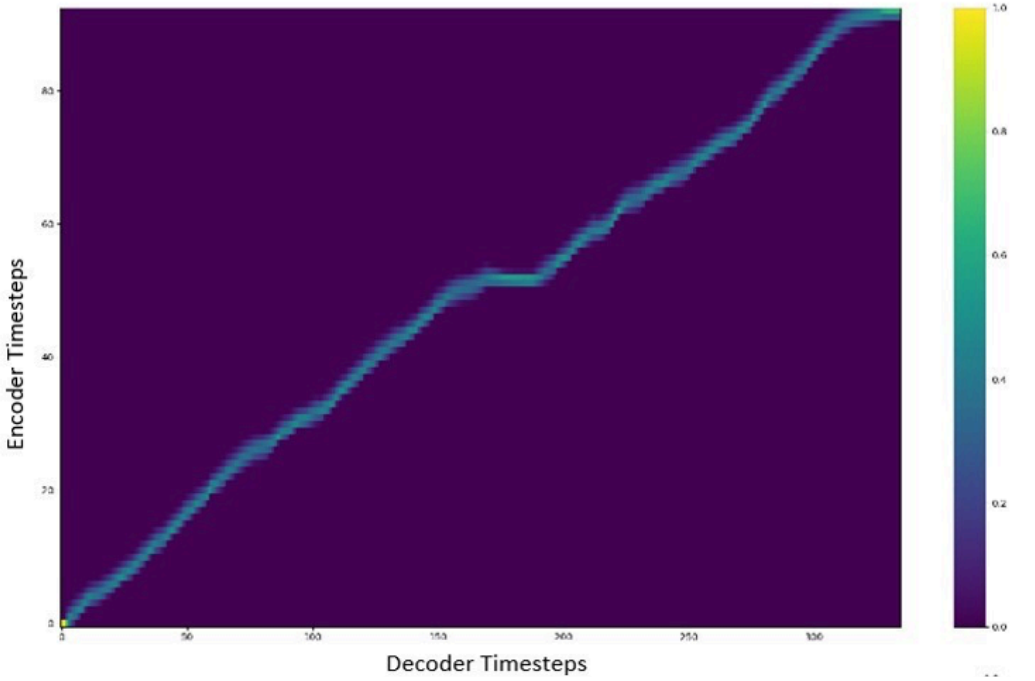


Figure 18. Predicted alignment of Tacotron2 model at 60000 steps

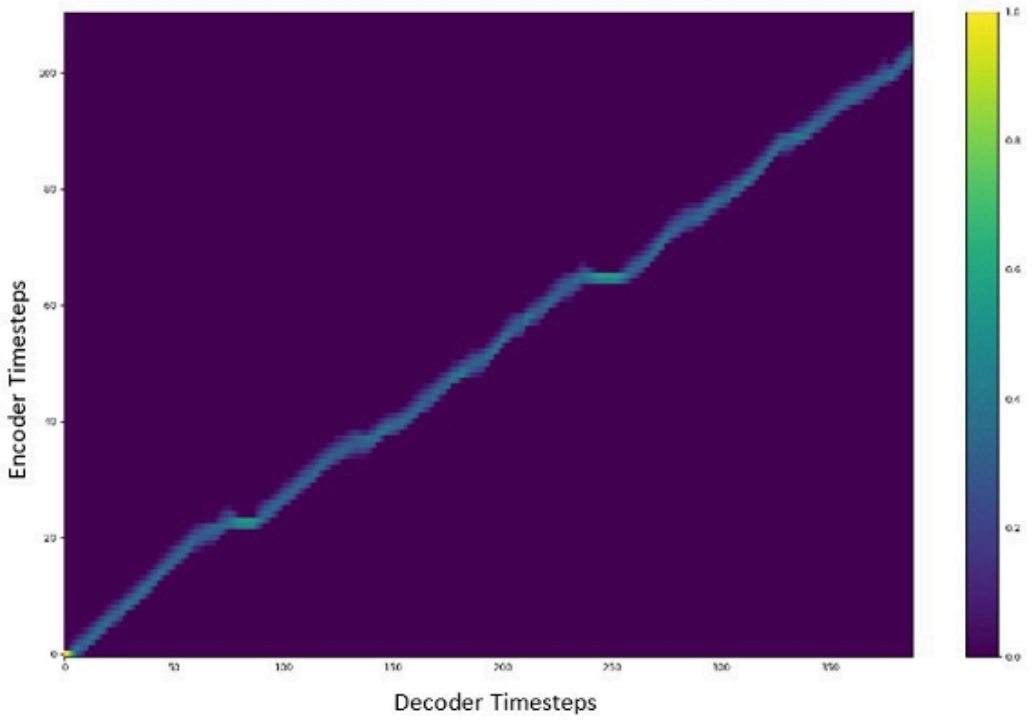


Figure 19. Predicted alignment of Tacotron2 model at 125000 steps

doi.org/10.32388/NT5LBM

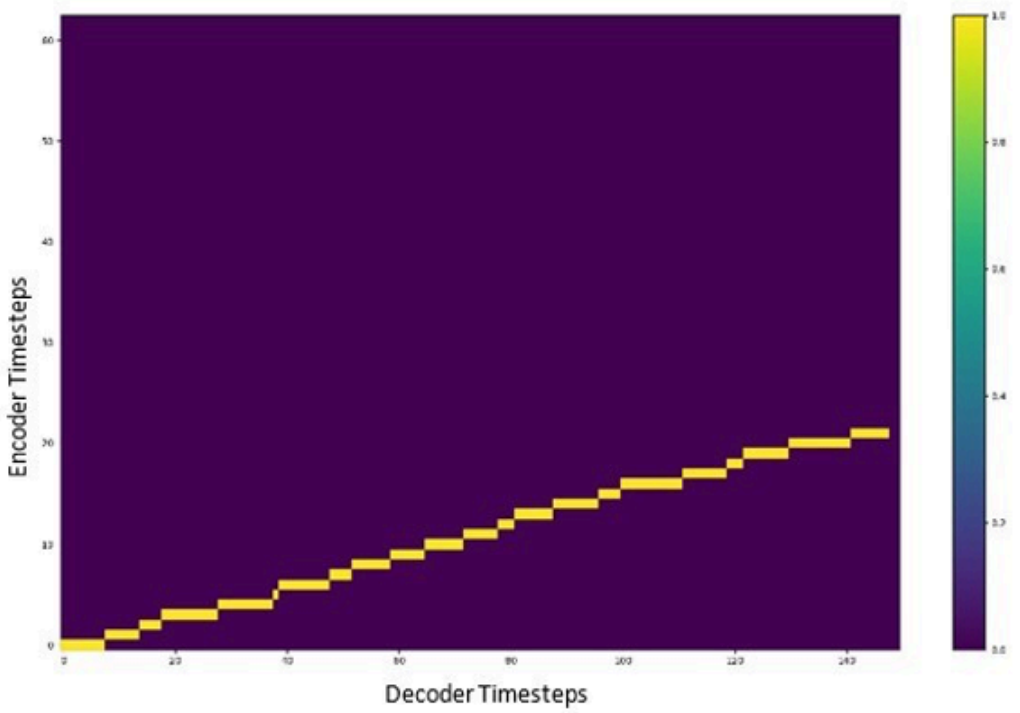


Figure 20. Predicted alignment of VITS model at 15000 steps

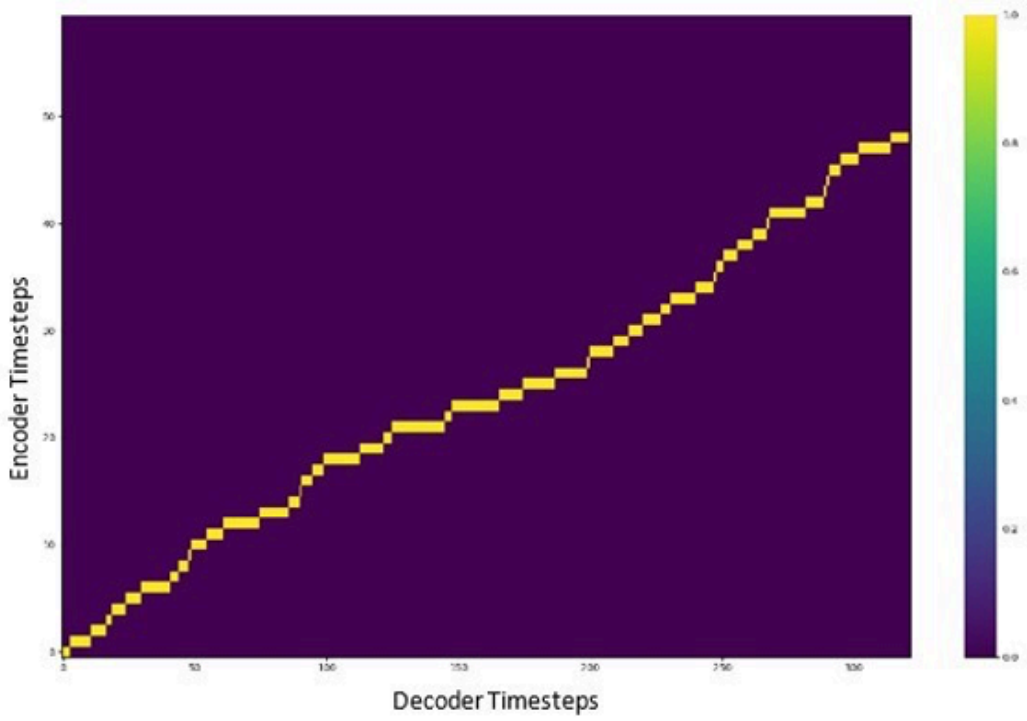


Figure 21. Predicted alignment of VITS model at 40000 steps

doi.org/10.32388/NT5LBM

The VITS model was trained for a total of nearly 69,000 steps over the course of 7 days and 8 hours. Figures 20, 21, and 22 illustrate the progression of the predicted alignments produced by the trained model. The total training loss for the VITS model was lowered from 3.6795 to 0.2425. The subjective evaluation conducted during training revealed that the VITS model produced speech that was generally understandable and sounded natural in the majority of samples but seemed to speak at a faster rate.

4.5. Testing and Evaluation

This section presents the testing and evaluation results.

4.5.1. Objective Testing

For the objective evaluation, the MCD values of the two models were measured and compared. Due to the lack of a valid ASR model for Kurdish (Sorani) that could be used in the evaluation procedure, we were unable to measure the character error rate. To measure the MCD for both models, Python libraries were utilized. The results of our objective evaluations indicate that the VITS model outperformed the Tacotron2 model by nearly two points. Table 5 displays the average calculated MCD values for the two trained models.

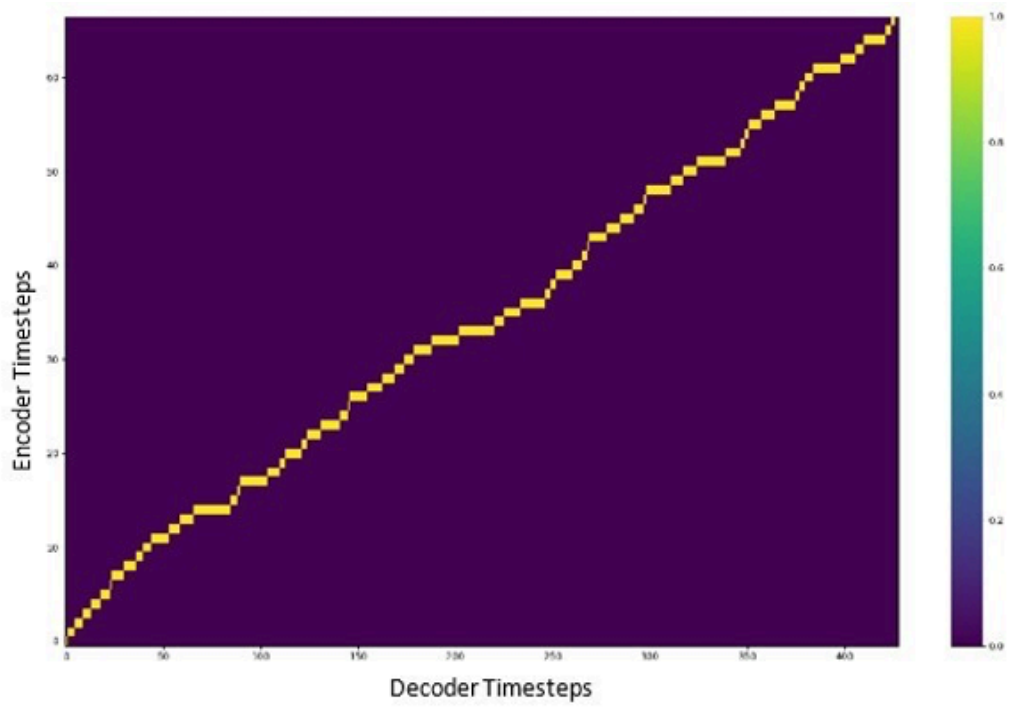


Figure 22. Predicted alignment of VITS model at 69000 steps

doi.org/10.32388/NT5LBM

	MCD
Tacotron2 TTS	9.89
VITS TTS	7,91

Table 5. Average Calculated Mel-Cepstral Distortion Value for the Tacotron2 and VITS Trained Models

4.5.2. Subjective Testing

Subjective evaluation of a TTS system relies on human perception, which is often measured by subjective listening tests (Wang et al., 2017; Shen et al., 2018; Arık et al., 2017; Gibiansky et al., 2017). In order to assess the quality and precision of the Tacotron2 and VITS models in our experiment, we use the MOS test.

A subjective evaluation is done to determine which of the proposed models performed best. The purpose is to compare the quality of the two TTS systems to determine how each proposed system performs and which model gives a better result. Because our research focuses on children, we need to evaluate the model from their point of view. In addition, we need the teachers and parents of the children to evaluate the model to determine its quality and accuracy. Then, we draw a conclusion by comparing the evaluation results of the two groups.

For each participant age group, we created a survey with age-appropriate language and question formats. We had a total of 40 participants; 20 children and 20 adults (teachers and parents of the children). For the purpose of designing the children's survey, we consulted with four kindergarten teachers to inquire about appropriate vocabulary and a question format that would be simple and straightforward for children to comprehend. We discovered that children respond better to short questions with a limited number of response options and questions that include images or illustrations. Taking into account what we learned from speaking with kindergarten teachers and incorporating MRS/ESOMAR guidelines, we established a set of criteria for designing the survey that are listed below:

- Questions should be concise.
- Questions should use simple language appropriate for children.
- Reducing the number of response options would make it easier for choldren to select an answer.

- Illustrations or audio should be used to make the survey interesting for children.

We designed the survey using SurveyMonkey, an online tool for creating interactive surveys. The survey consists of ten questions for both groups. The questions are divided into two groups, with the first six questions assessing the quality and intelligibility of the speech and the final four assessing the naturalness of the speech and evaluating which model performs better. Each model receives an equal

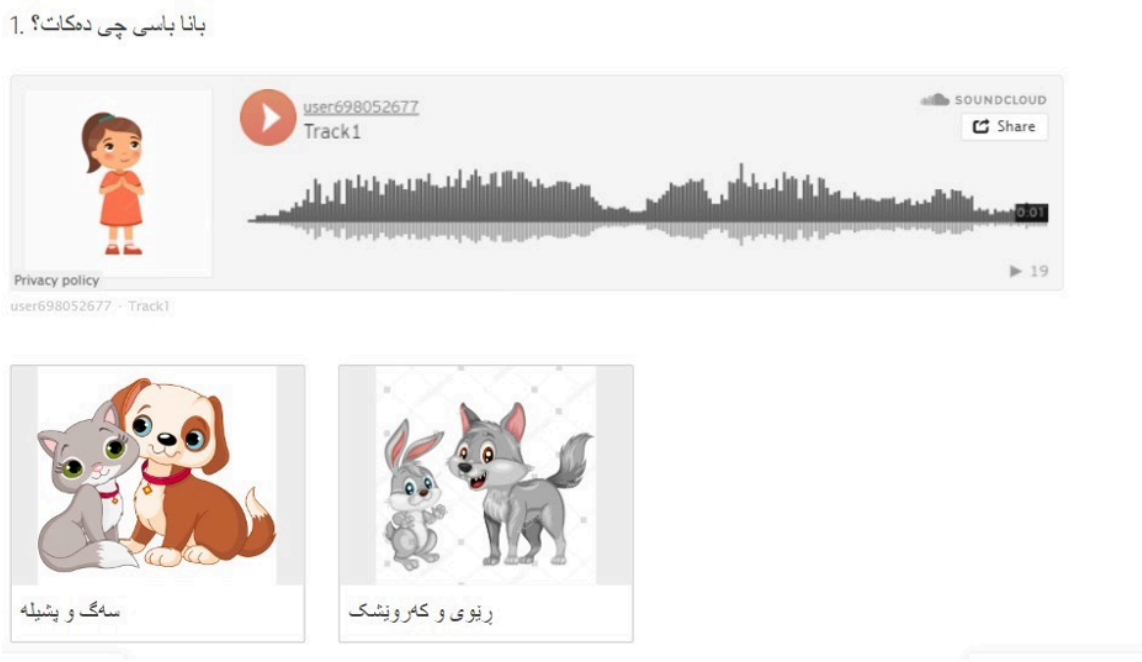


Figure 23. A Sample of the First Set of Questions for the Children's Survey

	Tacotron2 - MOS	VITS- MOS
Children's Group	1.32	3.25
Parents' and Teachers' Group	2.52	3.58
Average Result of Both Groups	1.92	5.41

Table 6. MOS Scores

doi.org/10.32388/NT5LBM

number of questions. Each model is randomly assigned three questions from the first set of questions and two questions from the second set. For the survey, we used a SurveyMonkey component called Image Choice that enables the addition of images to the response options. These recordings were made available on SoundCloud so that we could include them in our survey questions. A sample of each question type for the children's survey is presented in Figure 23 and Figure 24, respectively.

The first six questions for the second group of participants were designed to assess how comprehensible the model is on a five-point scale. Then final four questions were intended to determine which sample sounds better compared to the other. A sample of each question type for the parents' and teacher's survey are presented in Figure 25 and Figure 26, respectively.

After collecting survey responses and analyzing the data, we discovered that the VITS model outperforms the Tacotron2 model in terms of naturalness and intelligence. However, we discovered that children had more difficulty comprehending sentences than single-word audio files. Figure 25 Figure 25 presents the survey results for the children and parents group, respectively, and the MOS score results are presented in Table 6.

The purpose of the evaluation session was to play the audio files for each participant group and discuss the naturalness and intelligibility of the trained models after answering the survey. During the test, the children replayed the audio files twice or three times in order to comprehend the speech. We determined that this was a result of the naturalness of the voice. Children are sensitive to the sound of the speaker, and since the speech generated by the VITS model, despite being intelligible and somewhat natural-sounding, is not at a level where it can be easily comprehended by children.

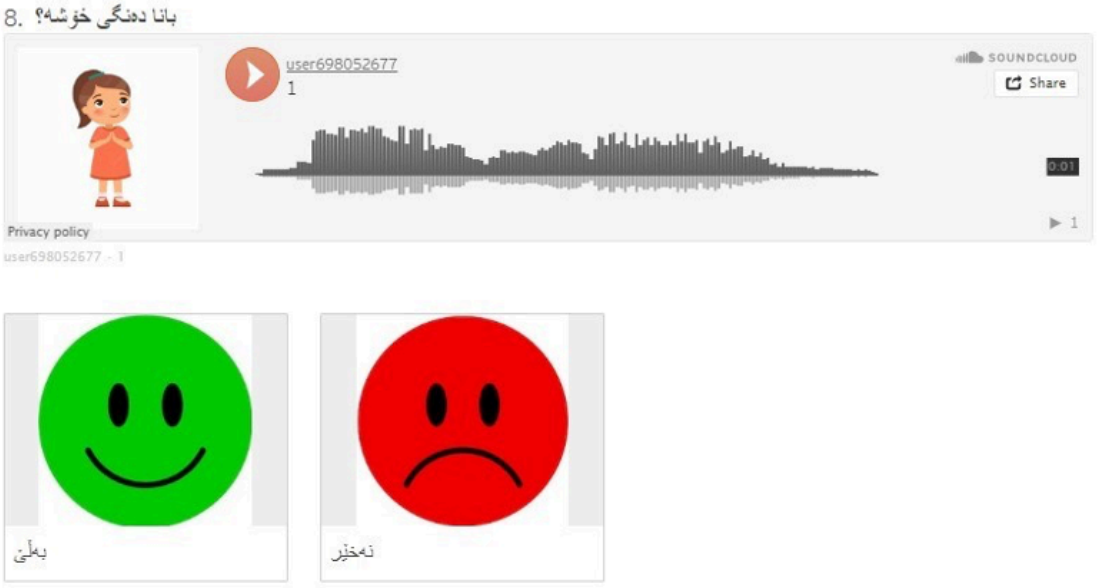


Figure 24. A Sample of the Second Set of Questions for the Children's Survey

doi.org/10.32388/NT5LBM

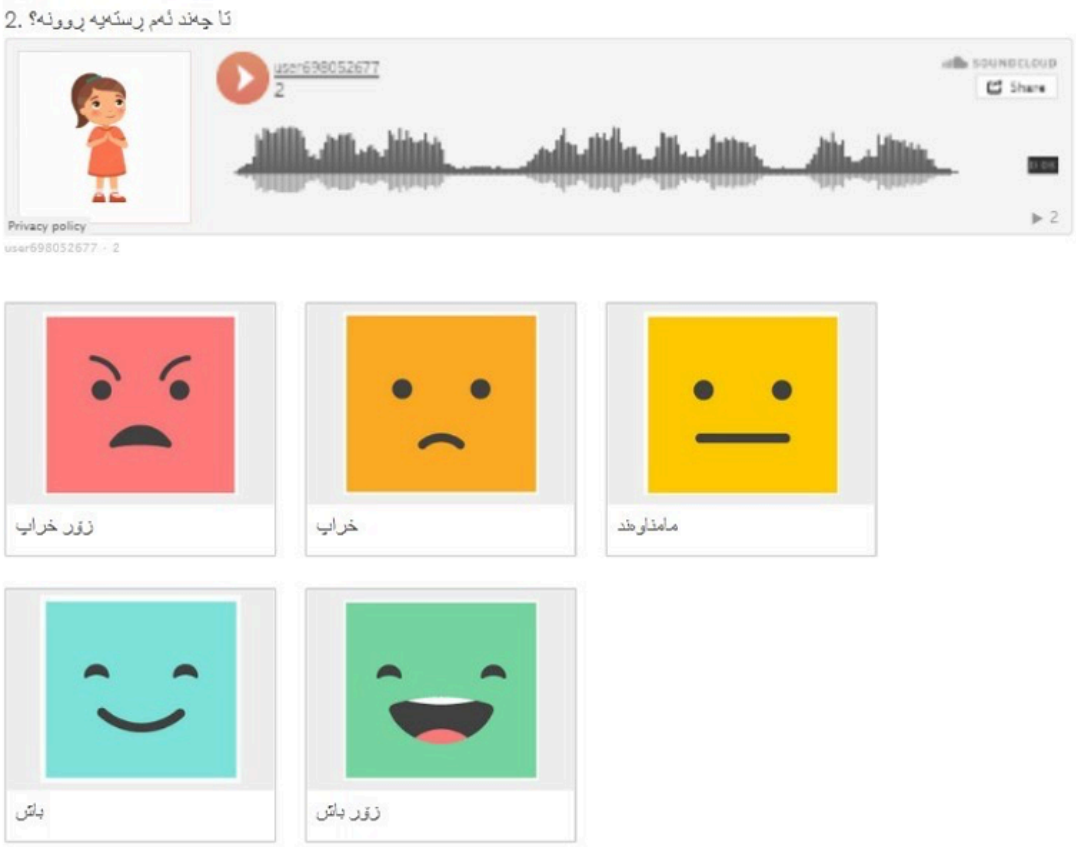


Figure 25. A Sample of the First Set of Questions for the Parents' and Teachers' Survey

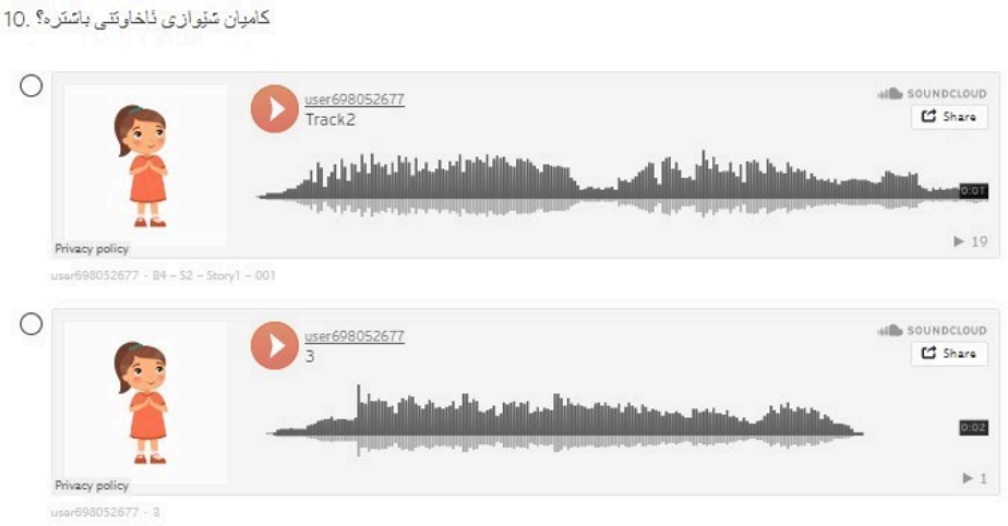


Figure 26. A Sample of the Second Set of Questions for the Parents' and Teachers' Survey

doi.org/10.32388/NT5LBM

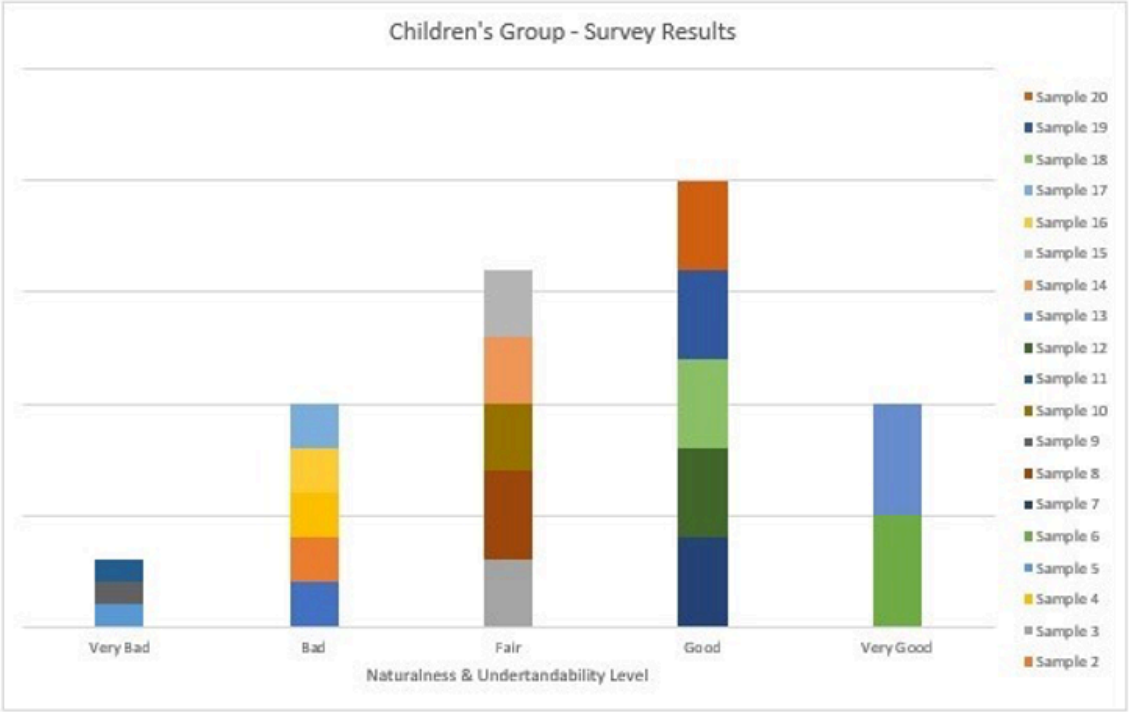


Figure 27. Children's Group - Survey Results

doi.org/10.32388/NT5LBM

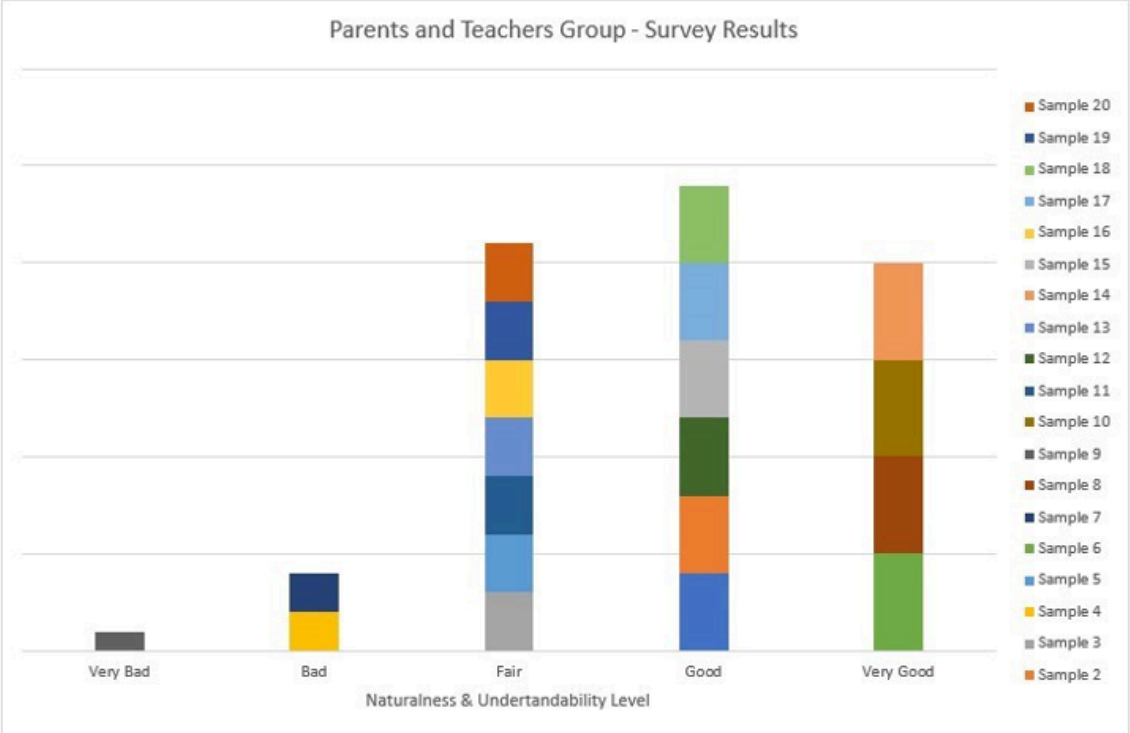


Figure 28. Parents' and Teachers' Group - Survey Results

4.6. Discussions

After training the models, we evaluate their training outcomes and performance during training. In general, the speech generated by the VITS model is more understandable and natural-sounding than that generated by the Tacotron2 model. We compared the two models based on training time, evaluation loss, and alignment prediction during training.

A good alignment is typically represented by a straight slope between the Y-axis encoder steps and the corresponding X-axis decoder steps (X-axis). From Figures 17, 18, and 19, we can observe that the slop of the predicted alignment for the Tacotron2 model improves slightly over the course of the training, although the slop appears to have been smoothed out and is not deemed a good alignment slop. On the other hand, if we examine the predicted alignment between the two models, we find that there is a significant difference between them. Compared to the VITS model, the alignment slope of the Tacotron2 model is smoother and has a smaller slope. This suggests that Tacotron2 has not yet predicted an accurate alignment. This indicates that additional training samples are required for the Tacotron2 model to produce more accurate results. For instance, to better align the samples and produce understandable

doi.org/10.32388/NT5LBM

speech, it would be helpful if the speaker maintained a consistent speaking rate and pitch throughout the recordings. Aside from the fact that the VITS model is aligned more accurately, the color and shape of the slop both indicate that the VITS model produces samples that sound clearer than those produced by the Tacotron2 model.

From another perspective, the manual inspection of the generated sample at various stages of Tacotron2 model training indicates that the model converges slowly with each iteration. We observed that the model produced slightly better speech for declarative sentences, which we related to the narrator's maintaining a consistent speaking rate and pitch. Passing interrogative and exclamatory sentences to the model generally resulted in speech that was difficult to understand. Since there is no information regarding the rate at which a sentence is delivered or the amount of time the model needs to generate frames, this must be derived from the training data, and the model must independently choose when to stop speech generation. This suggests that the model may better match the data if the speech is more homogeneous and has less variance.

In our experiment, subjective and objective testing produced results that were consistent with one another. The MCD test demonstrates that the VITS model is superior to the Tacotron2 model because its MCD score is nearly two points lower. A low MCD could indicate that the speaker speaks more clearly or more naturally, with fewer chances of mispronunciation. During the listening MOS test conducted with the children, we observed that they frequently replayed the audio samples before comprehending the speech, and we observed that this was related to the naturalness of the generated speech. From a different perspective, adult participants in the listening MOS test repeated the samples less frequently than children. This observation indicates that the speech generated by the models must be highly intelligible and sound as natural as possible if it is to be easily understood by children. To summarize the results of our training experiments, we discovered that it is advantageous to train the VITS model on Sorani when the training dataset is small. We also discovered that it is possible to train a Sorani model using a pre-trained English model that can synthesize imperfect but understandable speech with less than six hours of aligned data. More research is needed to determine the reason behind this, which is outside the scope of this thesis.

5. Conclusion

The purpose of this study was to identify the most effective approach to producing high-quality synthetic speech using limited training data for Sorani Kurdish. We developed a new text-to-speech

dataset, which includes roughly seven hours of speech from female speakers. Our dataset contains 209 stories collected from six children's books. A total of 4149 utterances were collected that have a duration that ranges from 0.1-10 seconds, which is approximately 1-25 words. We trained two models using the collected data and evaluated our system in the same context. Our dataset represents the first attempt at building a Sorani Kurdish <text, audio> pair dataset that could be used for deep learning.

We conducted our experiments using two text-to-speech frameworks: Tacotron2 and VITS framework. We observed that, in the absence of a large dataset, a high-quality dataset is advantageous for training textto-speech models. We discovered that for the model to converge more quickly, the speaker's speaking style and rate must be balanced. We also discovered that training a model for a new language requires either a large dataset or a dataset consisting mainly of sentences that have a balanced speaking rate, which helps in improving the intelligibility of the trained model. In contrast, we discovered that we can train a model for a new language with a small dataset by utilizing a model that has already been trained on the VITS framework.

By conducting these experiments, we were able to determine which aspects of the procedure present the greatest challenges. Due to the lack of linguistic tools for aligning text and audio datasets for Sorani Kurdish, segmenting and aligning the dataset was one of the most labor-intensive processes in our experiment.

The two models were subsequently evaluated objectively using the MCD test and subjectively using the MOS test. We conducted the subjective test using 20 samples from both models with two groups of participants: first, a group of children, followed by a group of teachers and parents. To facilitate the children's participation in the subjective examination, we designed an interactive survey. During this evaluation, the children's focus was on whether or not the samples were easy to comprehend, while the second group concentrated on both the models' understandability and naturalness. The results of the subjective MOS test revealed that the VITS model performed better than the Tacotorn2 model, with a score of 3.52 and 3.58 on the MOS scale for the first and second groups of participants, respectively. In conclusion, with a small amount of training data, the VITS synthesis system outperforms Tacotron2 in terms of naturalness and intelligibility on both subjective and objective tests.

In the future, we intend to improve the quality of the model by expanding the dataset to include 20 to 25 hours of speech, the bare minimum needed to train Tacotron2. In addition, to improve the intelligibility of the trained model, we aim to increase the number of declarative sentences that employ a normal speaking style while maintaining a consistent speaking rate throughout the recordings.

From another perspective, the model's reliance on data from a single speaker may be of importance given the limited quantity of data available for Sorani Kurdish. We can strive to build an architecture capable of learning from data collected from multiple speakers. Ideally, they would complement one another in terms of improving the accuracy of the model while being able to provide different voices for each speaker.

Finally, the effort and time necessary for aligning and segmenting the data collection continues to be the major bottleneck. If we can create a technology to automatically align text and its related speech for the Sorani Kurdish, then there will be a big improvement in reducing the human labor-intensive tasks. This tool can also help make the alignment more accurate by reducing the number of mistakes made by humans.

Acknowledgements

We acknowledge the efforts of our narrators, who decided to stay anonymous. We appreciate very much their high-quality contribution and their patience with us during the recording and auditing of the produced records.

Where to find the dataset?

The data are publicly available for non-commercial use under the CC BY-NC-SA 4.0 license 5 at https://github.com/KurdishBLARK/.

References

- Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., and Shoeybi, M. (2017). Deep voice: Real-time neural text-to-speech. In Doina Precup et al., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 195–204. PMLR, 06–11 Aug.
- Bahrampour, A., Barkhoda, W., and Azami, B. Z. (2009). Implementation of three text to speech systems for Kurdish language. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 14th Iberoamerican Conference on Pattern Recognition, CIARP 2009, Guadalajara, Jalisco, Mexico, November 15-18, 2009. Proceedings 14*, pages 321–328. Springer.

- Barkhoda, W., ZahirAzami, B., Bahrampour, A., and Shahryari, O.-K. (2009). A comparison between allophone, syllable, and diphone based TTS systems for Kurdish language. In *2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 557–562.
- Daneshfar, F., Barkhoda, W., and Azami, B. Z. (2009). Implementation of a text-to-speech system for kurdish language. In *2009 Fourth International Conference on Digital Telecommunications*, pages 117–120. IEEE.
- Dolson, M. (1986). The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27.
- Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J., and Zhou, Y. (2017). Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems*, 30.
- Gölge, E. (2022). Coqui-ai/TTS: a deep learning toolkit for Text-to-Speech, battle-tested in research and production.
- Hassani, H. and Kareem, R. (2011). Kurdish text to speech (ktts). In *Tenth International Workshop on Internationalisation of Products and Systems*, pages 79–89. Kuching Malaysia.
- Hassani, H. (2018). Blark for multi-dialect languages: towards the kurdish blark. *Language Resources and Evaluation*, 52(2):625–644.
- Idrees, S. and Hassani, H. (2021). Exploiting script similarities to compensate for the large amount of data in training tesseract lstm: Towards kurdish ocr. *Applied Sciences*, 11(20).
- Kelechava, B. (2015). Text-to-speech technology (speech synthesis), Dec.
- Kim, J., Kong, J., and Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.
- Kwon, O., Jang, I., Ahn, C., and Kang, H.-G. (2019). Emotional speech synthesis based on style embedded tacotron2 framework. In *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pages 1–4. IEEE.
- LanguageLizard. (2021). Award-winning kurdish-english bilingual children's books, audio books and dual language picture books. https://www.languagelizard.com/Kurdish-Bilingual-Children-s-Books-s/2733.htm.
- Latorre, J., Lachowicz, J., Lorenzo-Trueba, J., Merritt, T., Drugman, T., Ronanki, S., and Klimkov, V. (2019). Effect of data reduction on sequence-to-sequence neural tts. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7075–7079. IEEE.
- Lemmetty, S. (1999). Review of speech synthesis technology.
- Maryland Library Resource Center. (2023). Guide to Picture Books. https://www.slrc.info/resources/guides/books-reading/guide-to-picture-books/. Accessed on:

26.09.2023.

- MontrealCorpusTools. (2021). Montreal-forced-aligner.
- MRS. (2014). MRS Guidelines for Online Research, Sep.
- MSR. (2018). ESOMAR/GRBN Guideline on Research and Data Analytics with Children, Young People, and Other Vulnerable Individuals.
- Muhamad, S. and Veisi, H. (2022). End-to-End Kurdish Speech Synthesis Based on Transfer Learning. *Passer Journal of Basic and Applied Sciences*, 4(2):150–160.
- Ning, Y., He, S., Wu, Z., Xing, C., and Zhang, L.-J. (2019). A review of deep learning based speech synthesis. *Applied Sciences*, 9(19).
- Nodelman, P., Hamer, N., and Reimer, M. (2017). *More words about pictures: Current Research on Picturebooks and Visual/Verbal Texts for Young People*. Routledge-Taylor & Francis.
- Nodelman, P. (1988). *Words about pictures: The Narrative Art of Children's Picture Books*. University of Georgia Press.
- Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio.
- Podsiadlo, M. and Ungureanu, V. (2018). Experiments with training corpora for statistical text-to-speech systems. readbeyond. (2020). aeneas.
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.
- Team, A. (2022). Audacity.
- Tu, T., Chen, Y.-J., Yeh, C.-c., and Lee, H.-Y. (2019). End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508*.
- Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis.
- Zhao, S., Yuan, Q., Duan, Y., and Chen, Z. (2023). An End-to-End Multi-Module Audio Deepfake Generation System for ADD Challenge 2023. *arXiv preprint arXiv:2307.00729*.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.