Open Peer Review on Qeios

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

Bala Farhad[1], Hossein Hassani[2]

1 University of Kurdistan Hewler
2 University of Kurdistan Hewlêr (UKH)

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

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 Text-to-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.

Bala Farhad, and **Hossein Hassani**

University of Kurdistan Hewlêr

Kurdistan Region - Iraq

{bala.farhad, 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 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 because of the necessity t for individual sample processing.

Arık et al. (2017) present 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 Tacotronmodel (Wanget 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 maps the input text to Mel spectrograms for speech synthesis. Seq2Seq neutral networks can transfuse an input sequence to an output sequence with the possibility of having a different length. Tacotron2 combines 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 (by the Mel Spectogram), and second, a generated waveform from this representation. Usually, Tacotron2 is combined with a vocoder [vcoder is a contraction of the term voice coder (Dolson, 1986)] to synthesize the Mel spectrograms 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, and According to Kim et al. (2021), the general architecture of this model consists of “a posterior encoder, a prior encoder, decoder, discriminator, and stochastic duration predictor.” The posterior encoder and discriminator are only used in the training phase, not the inference phase. This technique utilizes variational inference, which is then supplemented with normalizing flows and an adversarial training procedure. The result is an increase in the expressive capability of the generative modeling. By using uncertainty modeling over latent variables and stochastic duration prediction, this method expresses a natural one-to-many relationship by allowing a text input to be pronounced in various ways with varying pitches and rhythms. That 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 an 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 this work 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 that is five to ten. 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.

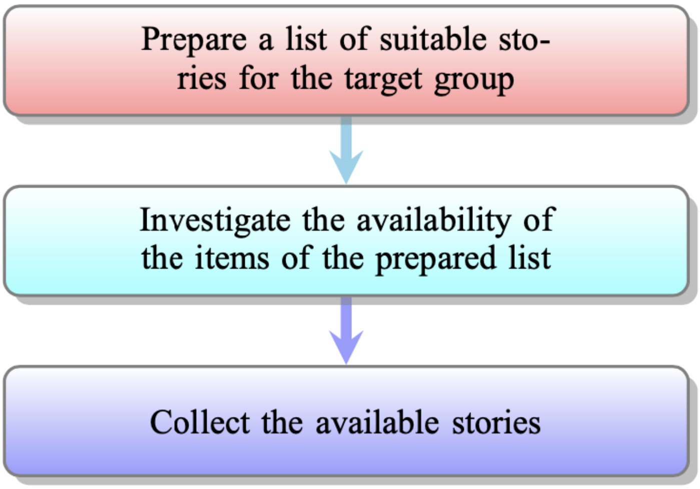

Prepare a list of suitable sto-
ries for the target group
Investigate the availability of
the items of the prepared list
Collect the available stories

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.

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

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

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

3.2. Audio Recordings

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 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 in the audio recording phase.

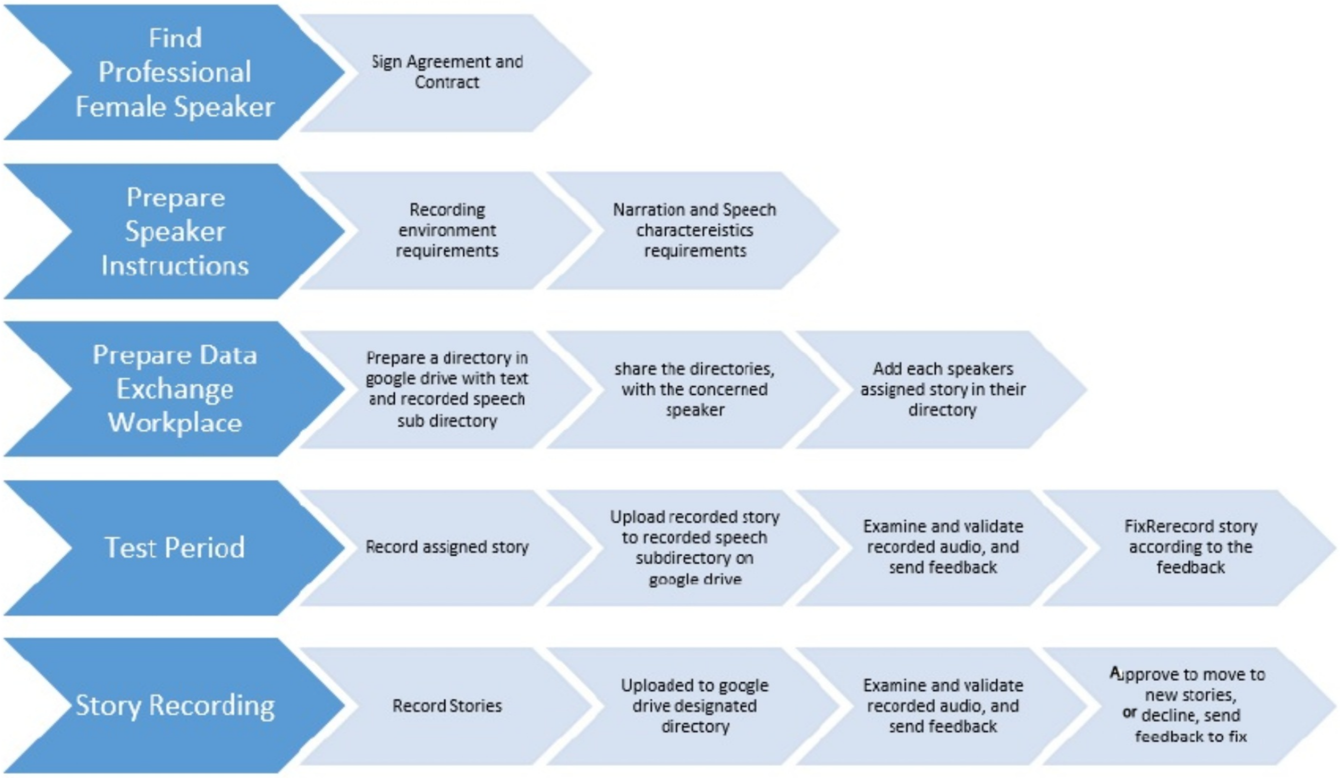

Find Professional Female Speaker
Sign Agreement and Contract
Prepare Speaker Instructions
Recording environment requirements
Narration and Speech charactereistics requirements
Prepare Data Exchange Workplace
Prepare a directory in google drive with text and recorded speech sub directory
share the directories, with the concerned speaker
Add each speakers assigned story in their directory
Test Period
Record assigned story
Upload recorded story to recorded speech subdirectory on google drive
Examine and validate recorded audio, and send feedback
FixRerecord story according to the feedback
Story Recording
Record Stories
Uploaded to google drive designated directory
Examine and validate recorded audio, and send feedback
Approve to move to new stories, or decline, send feedback to fix

Figure 2. Story Audio Recording Task Flow Diagram

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.

The labeling and metadata of the segmented audio records have four sections of information. The starting section is formed by the letter B followed by a one-digit number indicating the ID of the narrated book. The second section contains the speaker ID, which is retrieved from a file that keeps the speaker information. The third section holds the number of the stories in the book. The last section is a three-digit number assigned to 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.

3.4. Preparing Dataset

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

For Tacotron2, we split the recorded stories into segments of 1 to 10-second length, using Audacity , that are aligned with their corresponding transcribed text in a comma-separated-value (CSV) file.

For Festival, we divide the recorded files into chunks five to six words long and label them to create a tab-separated-value (TSV) file containing the audio file labels and their corresponding transcribed text. Then, we construct a lexicon, transliterate it into its 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 evaluated in terms of intelligibility and naturalness by two target audiences, children and adults, including the parents and instructors of the children. Initially, the children's teachers reviewed the collected stories to ensure their suitability for the target age range. Afterward, we examined samples of narrated stories against the criteria stated in the data collection section to provide feedback to the narrators to improve the recording quality.

The quality and performance of the model could be evaluated both objectively and subjectively. Two primary techniques exist for objective evaluations: the Character Error Rate (CER) can be determined using an automatic speech recognition (ASR) model, and the Mel Cepstral Distortion (MCD) measures the difference between target audible speech and the spectral feature (Diener et al., 2015).

Subjective evaluation of a TTS system relies on human perception, often by obtaining feedback from subjective listening tests (Wang et al., 2017; Shen et al., 2018; Arık et al., 2017; Gibiansky et al., 2017). In this test, participants listen to several samples and answer a questionnaire to evaluate 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). They help to conclude what should be updated to improve the model.

4. Experiments, Results, and Discussion

This section reports the experiments, presents the results, and discusses the outcome of 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 of the targeted age group. The following are the criteria for the textual corpus collection:

- The text must be in Sorani Kurdish.
- 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.
- Each story must include at least two illustrations.

The first step of the data collection resulted only in having a few online stories that met the criteria for inclusion in the corpus. While venues such as LanguageLizard (2021) offer a few English-Kurdish storybooks for children interested in learning a second language, they provide it solely in hard copies, which we could not have within the timeline of the experiments. Consequently, we collected the books from the local bookshops.The kindergarten teachers assessed the books to verify their suitability for the target age range. Out of eight books, six passed the verification. The rejected ones were lengthy, contained hard-to-understand vocabulary, and were less illustrative. Figure 3 presents the coverage of the collected books used to create the corpus.

Figure 3. List of Collected Books for the Corpus

For the image conversion, we used 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.

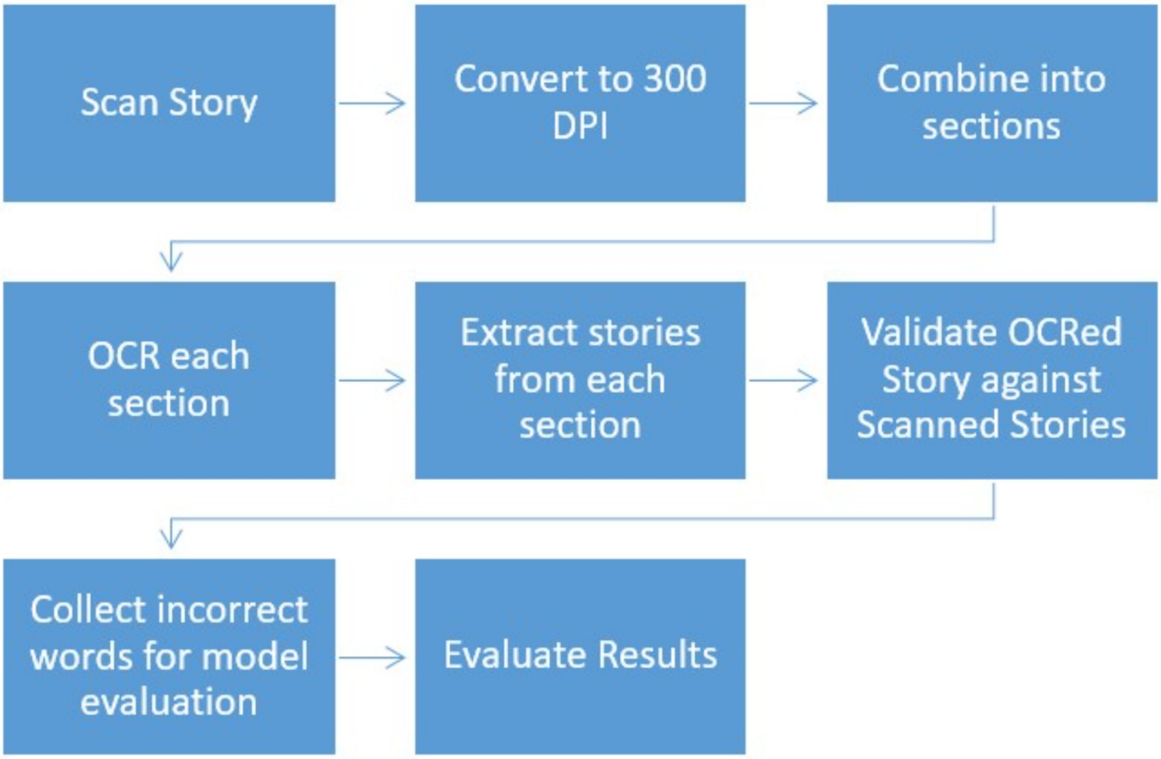

Scan Story
Convert to 300 DPI
Combine into sections
OCR each section
Extract stories from each section
Validate OCRed Story against Scanned Stories
Collect incorrect words for model evaluation
Evaluate Results

Figure 4. Story OCR-ing Process Flow Diagram

We scanned the collected (300 DPI), manually converted them to black and white, cropped them to text only, and exported them to PNG file format to be used by the OCR system. The stories were scanned and sectioned based on their corresponding book number. Once the data were ready, the scanned images were OCR-ed, and we reviewed the output to find cases such as {\Kurdishfont{میەرەبانییەوە}} that we replaced with {\Kurdishfont{میهرەبانییەوە}}..

4.1.1. Recording and data preparation

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 if the dataset is phonetically balanced and includes good-quality recordings. Consequently, we considered the synthesis system's requirements for collecting the speech corpus as follows:

- The stories must be narrated by trained professionals.
- Speakers will narrate the stories in an animated manner appropriate for children's story reading.
- The narrators must record the stories in WAV (16 PCM) format with a 22050 KHz sample and a quiet environment.

We hired two female speakers, a freelance digital creator, and a novel narrator at a radio station, both of whom were suitable for the task, and they could keep a balanced voice while speaking for long periods. Both speakers went through a test period of 3-5 days to ensure the collected data met the specifications required for high-quality speech. For the recording, the narrators used a Rode NT1-A, with a pop shield to mitigate the impact of exhaled air, with a reputation for producing high-quality output. The process produced 209 files, 115 by the first narrator and 94 by the second, containing 34530 words (after transcript revisions). The first speaker produced that added up to seven hours of speech. Table 2 shows the statistics of the collected data before processing and segmentation.

Table 2. Story Duration Statistics Before Pre-Processing

	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

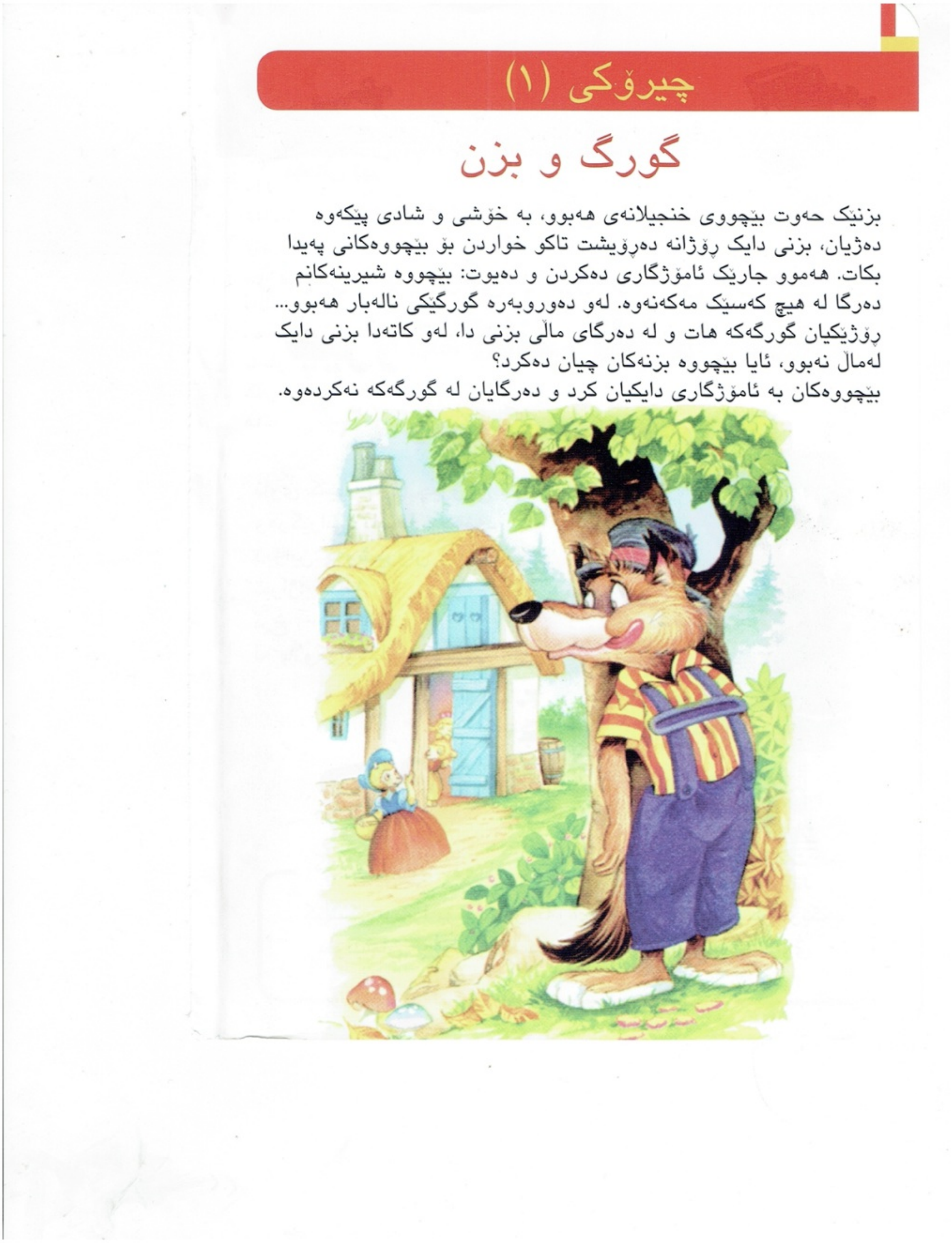
We manually segmented the obtained files to a maximum of 10-second pieces based on the length of each textual sentence and aligned it with their related text. To do so, we used Audacity and checked the sample rate of the recording to be 220500KHz and single-channel audio. If a segment had two channels, we applied a built-in function in Audacity to split stereo to mono, removed the second channel, and labeled them accordingly. 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 input to end-to-end speech synthesis is a <text, audio> pair, where the first item is a folder with all the audio files, and the second is a spreadsheet that uses the utterance segment file name and its related transcript. The following sections provide the details of the process.

4.2.1. Text Pre-Processing

The OCR process converted books into text files. Figure 5 shows a sample page of a book in its original format, and Figure 6 shows the output.

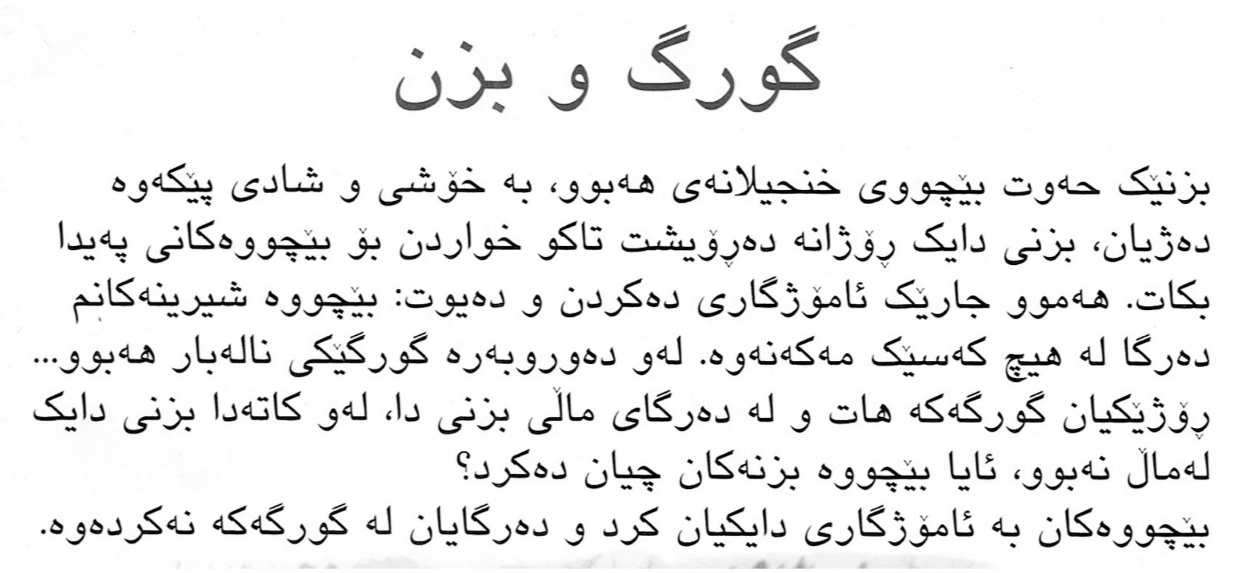


چیرۆکی (١)

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

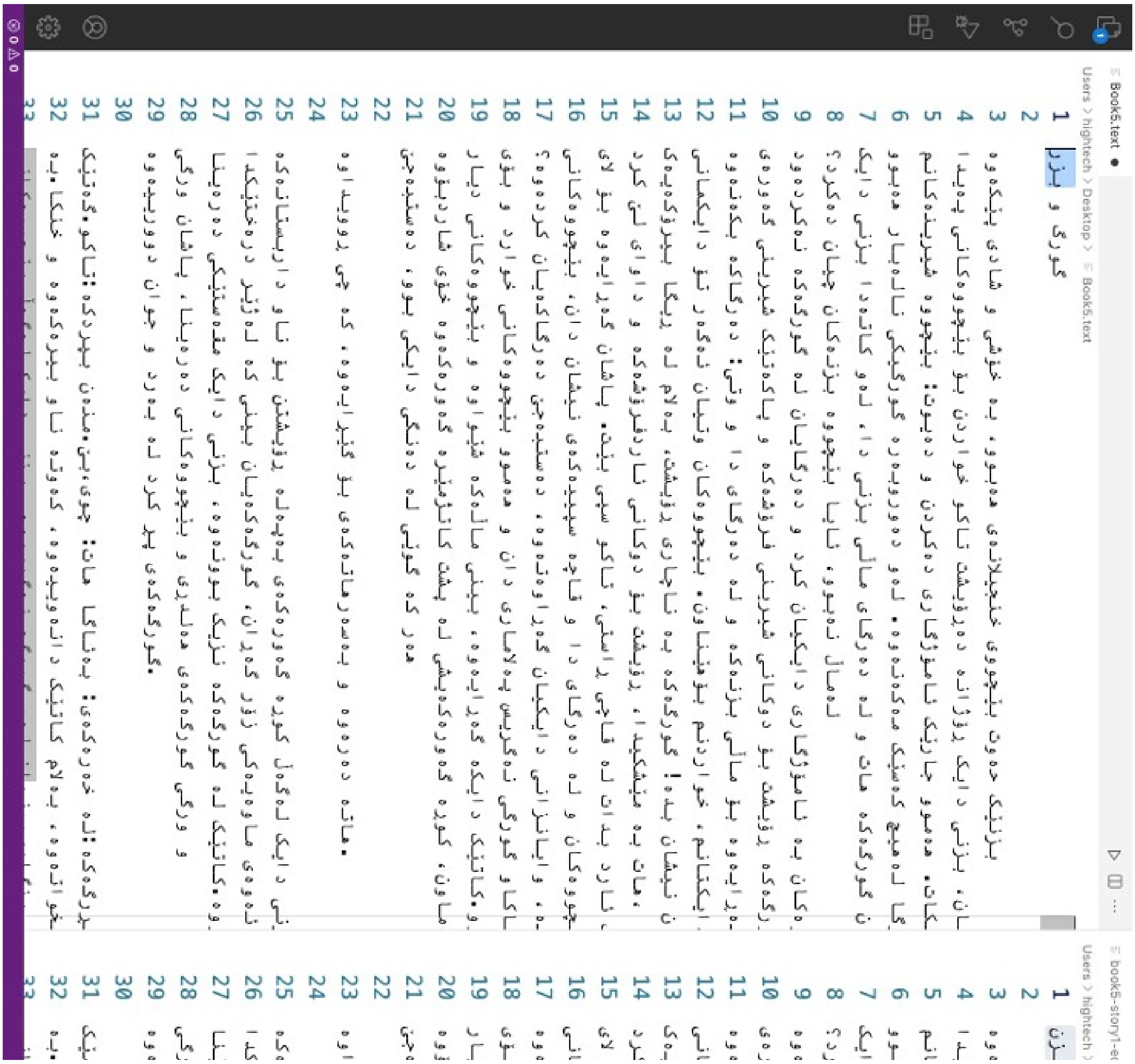


گورگ و بزن

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 دەرگای ماڵی 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



Desktop > book5-story1-edited.txt
Ln 1, Col 11 (3 selected) Spaces: 4 UTF-8 LF Plain Text

Figure 7. Reviewing OCR-ed Story

4.2.2. Speech Pre-Processing

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

- Include brief pauses at the beginning and end of sentences. That is required to provide context for each recorded utterance.
- To execute ”Dynamic Range Compression” (intensity) on each utterance loudness by multiplying the signal by a dynamic gain to maintain the signal within a predetermined limit. That is 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. While we did not suggest an exact pause duration, we gave them feedback to reduce long pauses to maintain acceptable variability in pause length without jeopardizing the precision of the alignment.

To further refine the segmentation process, it is necessary to identify the beginning and ending points of sentences. We use Microsoft Excel functions to create the segment labels from the story labels. Figure 8 demonstrates that the audio recording has stereo channels, and Figure 9 shows its mono-track form after the conversion process.

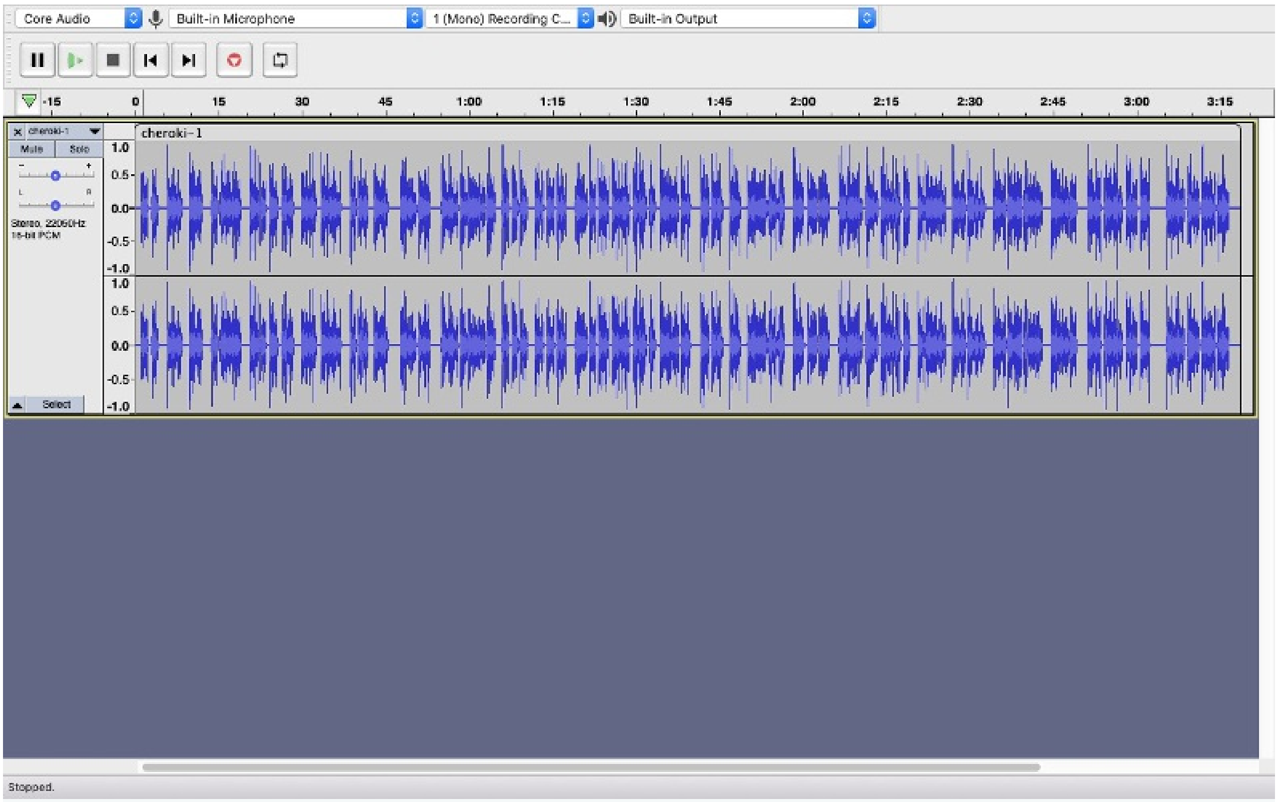


Figure 8. Raw Recorded Stereo Channeled Story

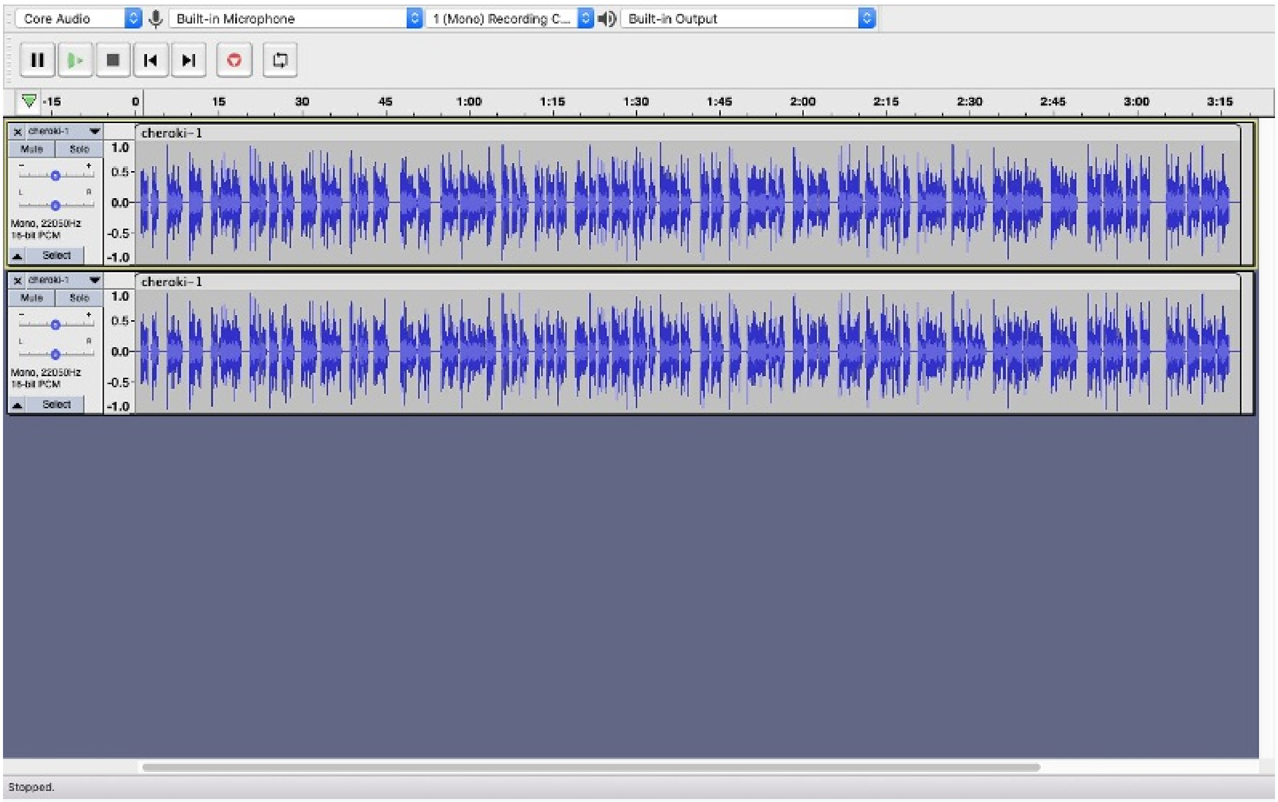


Figure 9. Raw Recorded Story with Separated Stereo and Mono Channel

Figure 10 and 11 show a labeled segment and a series of labeled segments, respectively.

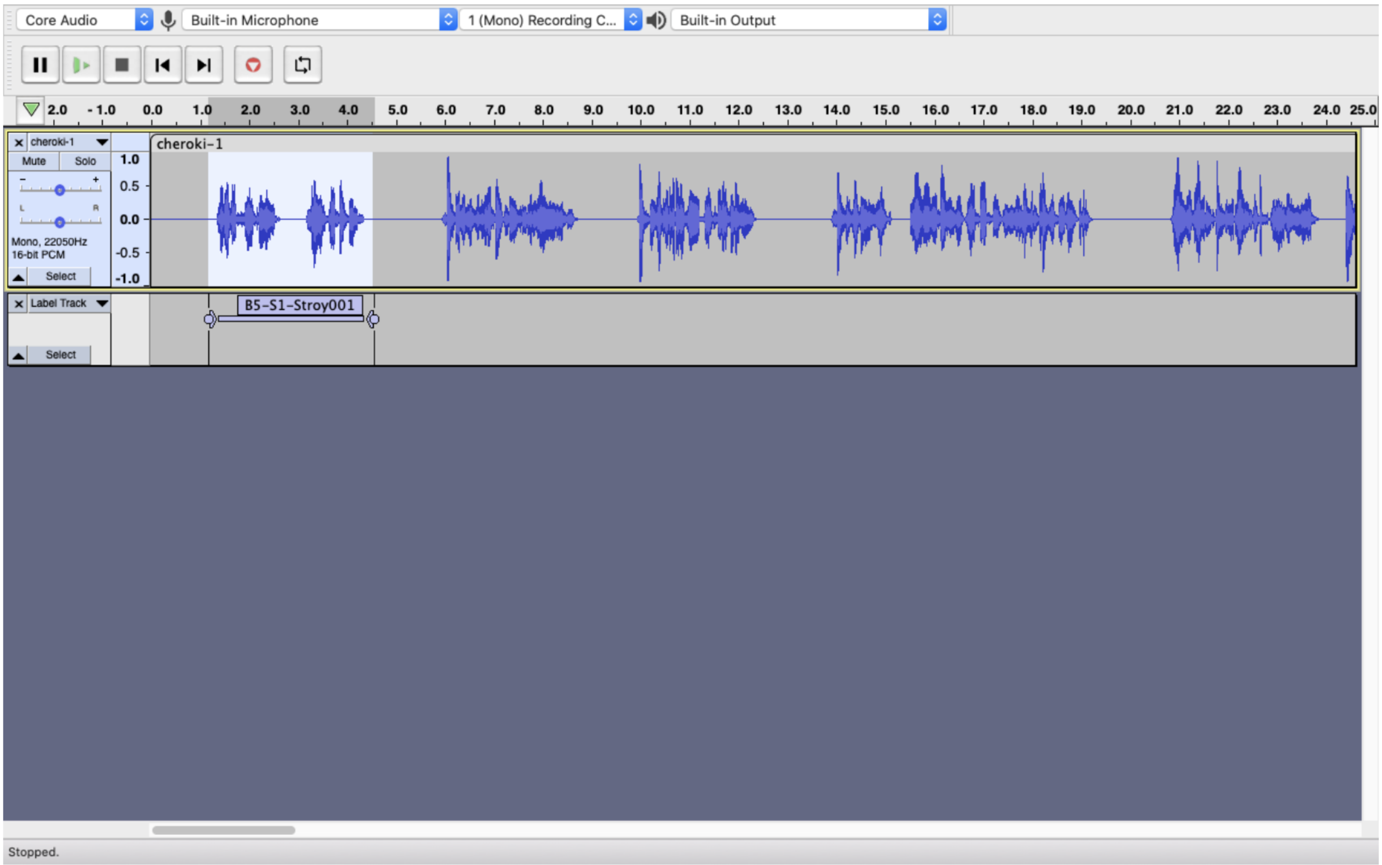


Figure 10. Creating Label for a Selected Utterance on Audacity

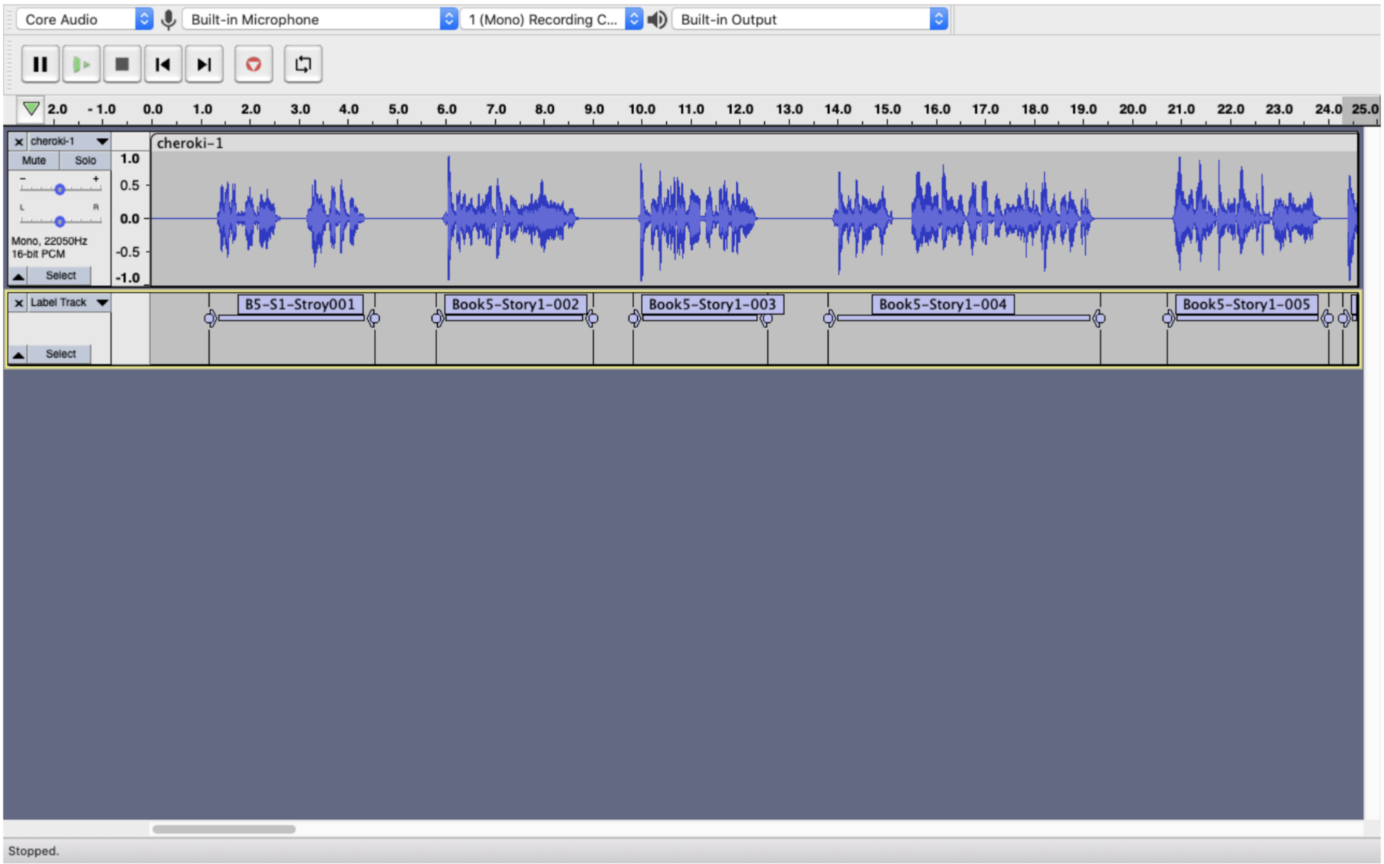
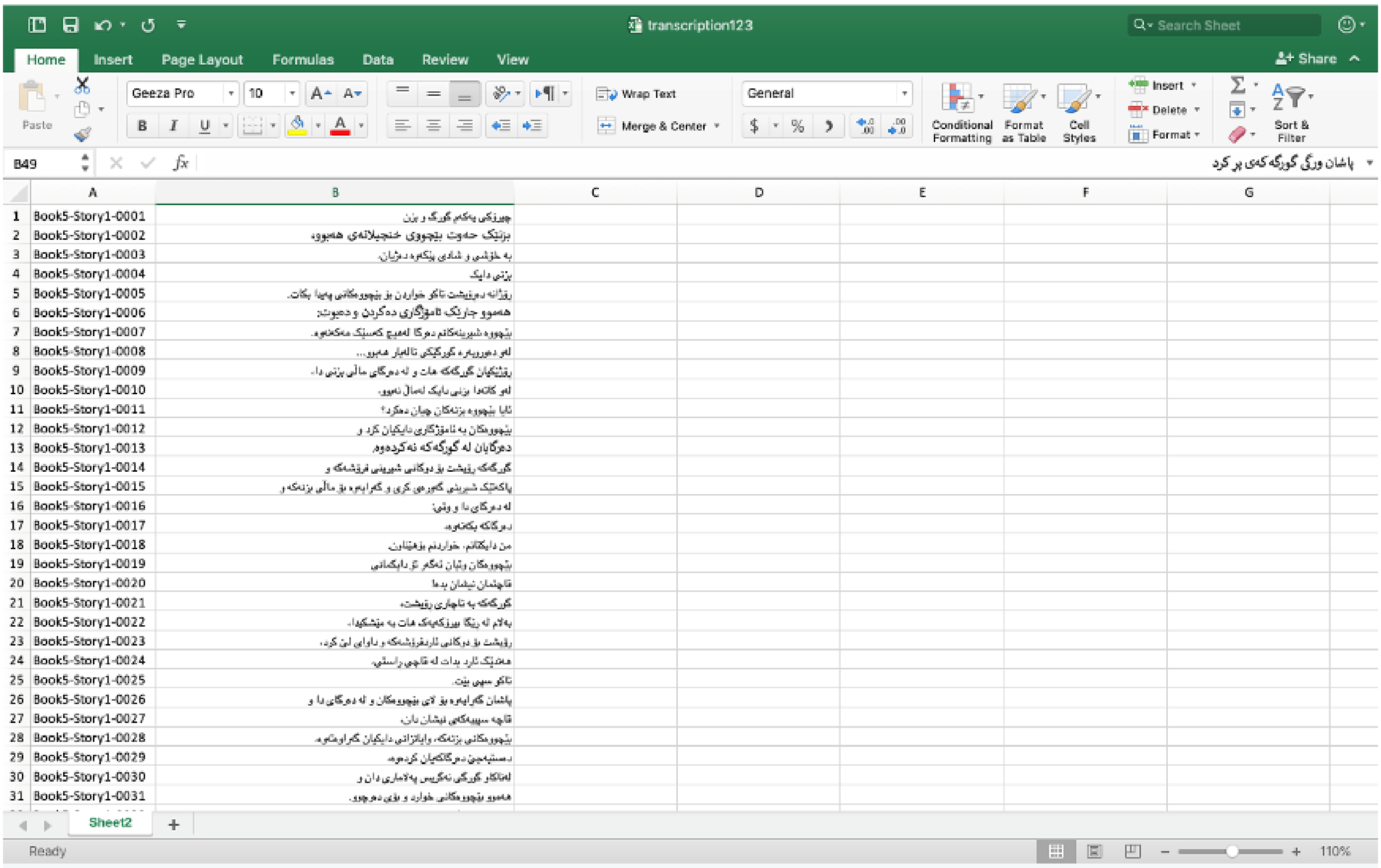


Figure 11. An Example of a Labeled Story

A single spreadsheet contains all labels to keep track of the labels for each segmented utterance and the transcripts that correspond to those utterances. Figure 12 shows the transcripts and their corresponding labels.



B49 پاشان ورگی گورگەکەی پر کرد

	A	B
1	Book5-Story1-0001	چیرۆکی یەکەم گورگ و بزن
2	Book5-Story1-0002	بزنێک حەوت بێچووی خنجیلانەی هەبووە
3	Book5-Story1-0003	بە خۆشی و شادی پێکەوە دەژیان.
4	Book5-Story1-0004	بزنی دایک
5	Book5-Story1-0005	ڕۆژانە دەڕۆیشت تاکو خواردن بۆ بێچووەکانی پەیدا بکات.
6	Book5-Story1-0006	هەموو جارێک ئامۆژگاری دەکردن و دەیوت:
7	Book5-Story1-0007	بێچووە شیرینەکانم دەرگا لەهیچ کەسێک مەکەنەوە.
8	Book5-Story1-0008	لەو دەوروبەرە گورگێکی ئالەبار هەبوو...
9	Book5-Story1-0009	ڕۆژێکیان گورگەکە هات و لە دەرگای ماڵی بزنی دا.
10	Book5-Story1-0010	لەو کاتەدا بزنی دایک لەماڵ نەبوو.
11	Book5-Story1-0011	ئایا بێچووە بزنەکان چیان دەکرد؟
12	Book5-Story1-0012	بێچووەکان بە ئامۆژگاری دایکیان کرد و
13	Book5-Story1-0013	دەرگایان لە گورگەکە نەکردەوە
14	Book5-Story1-0014	گورگەکە ڕۆیشت بۆ دوکانی شیرینی فرۆشەکە و
15	Book5-Story1-0015	پاکەتێک شیرینی گەورەی کری و گەڕایەوە بۆ ماڵی بزنەکە و
16	Book5-Story1-0016	لە دەرگای دا و وتی:
17	Book5-Story1-0017	دەرگاکە بکەنەوە.
18	Book5-Story1-0018	من دایکتانم، خواردنم بۆهێناون
19	Book5-Story1-0019	بێچووەکان وتیان ئەگەر تۆ دایکمانی
20	Book5-Story1-0020	قاچمان نیشان بدە
21	Book5-Story1-0021	گورگەکە بە ناچاری ڕۆیشت.
22	Book5-Story1-0022	بەڵام لە ڕێگا بیرۆکەیەک هات بە مێشکیدا.
23	Book5-Story1-0023	ڕۆیشت بۆ دوکانی ئاردفرۆشەکە و داوای لێ کرد،
24	Book5-Story1-0024	مشتێک ئارد بدات لە قاچی ڕاستی.
25	Book5-Story1-0025	تاکو سپی بێت.
26	Book5-Story1-0026	پاشان گەڕایەوە بۆ لای بێچووەکان و لە دەرگای دا و
27	Book5-Story1-0027	قاچە سپییەکەی نیشان دان.
28	Book5-Story1-0028	بێچووەکانی بزنەکە، وایانزانی دایکیان گەڕاوەتەوە.
29	Book5-Story1-0029	دەستبەجێ دەرگاکەیان کردەوە.
30	Book5-Story1-0030	لەناکاو گورگی ئەگریس پەلاماری دان و
31	Book5-Story1-0031	هەموو بێچووەکانی خوارد و بزی دەرچوو.

Figure 12. Transcription Sheet to Map Labels and Utterance Transcripts

4.2.3. The Dataset

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

Table 3. Pre-Processed Story Duration Statistics

	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	

Removing unnecessary silence and noise decreased the dataset length from 417 minutes to 308. The dataset includes 5 hours of speech segmented into 4142 segments from approximately five hours of speech from 209 stories. Each segment is approximately 5 seconds long and contains eight words on average. Figure 12 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 summariezs the result.

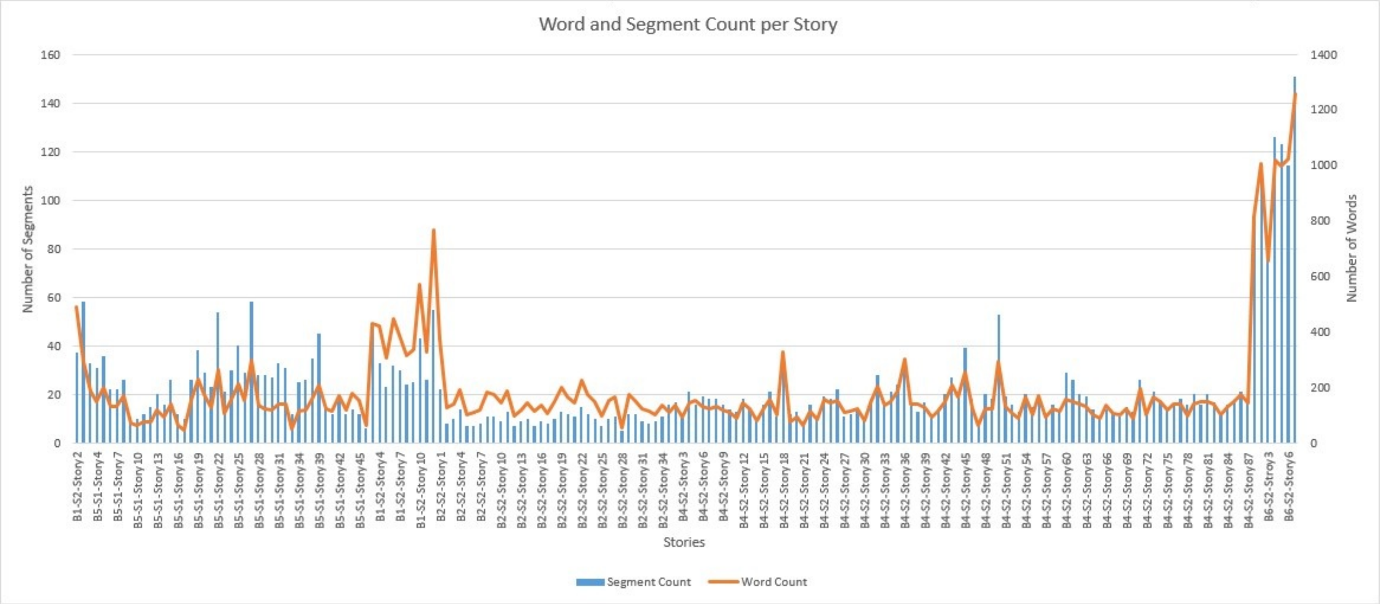


Figure 13. Word and Segment Count per Story

Table 4. Dataset Statistics

	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

4.3. Experiments

The following sections report the experiments.

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. We use a computer with 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

We used 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. Researchers and developers have used it in over 20 languages. The software for the experiment 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

Unless stated differently, the setting of hyperparameters is identical for both frameworks. For Tacotron2 training, we used a batch size of 16 with 3000 epochs and a learning rate of 0.01, while for VITS, the batch size was eight with 4000 epochs and a learning rate of 0.00001.

4.4. Training Results

The training of Tacotron2 took 125,000 steps over 17 days and 14 hours. Figures 14, 15, and 16 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.

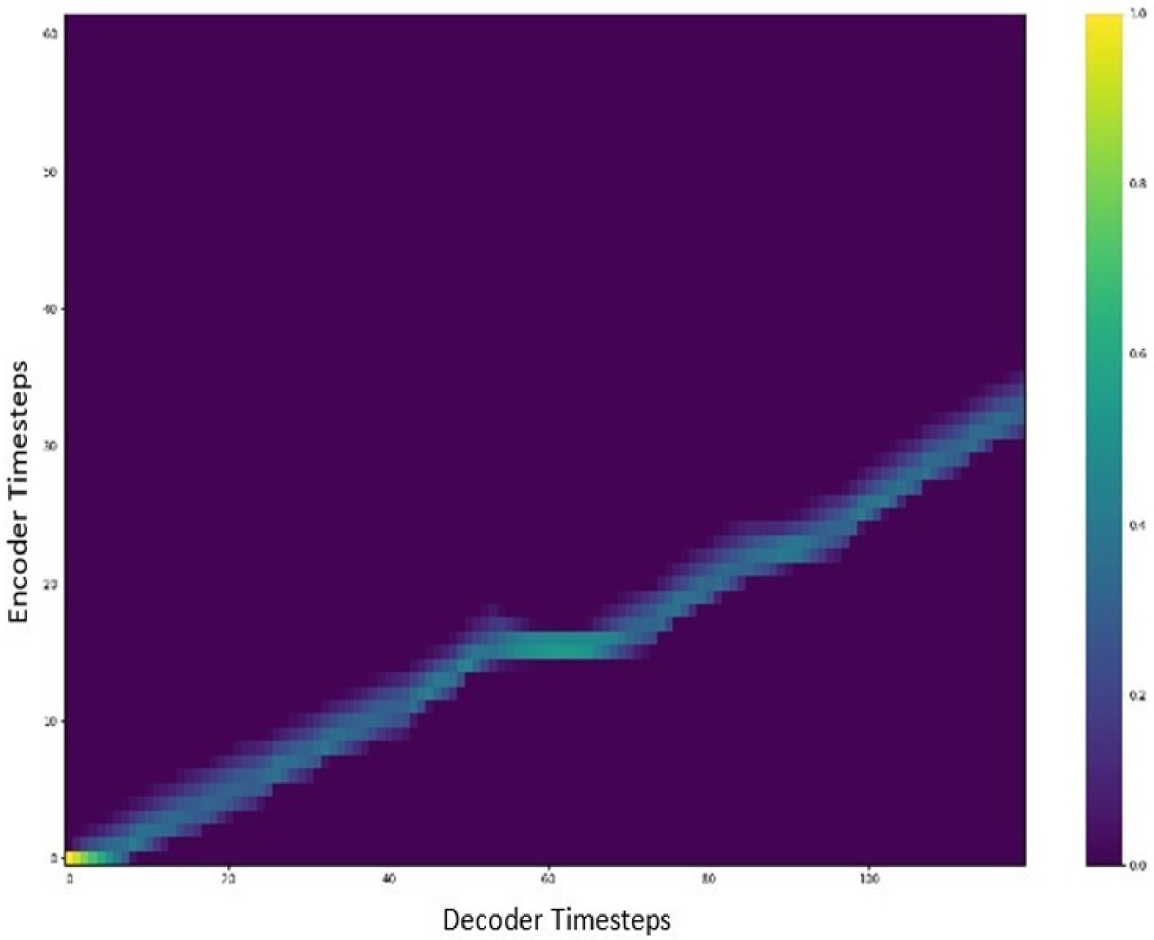


Figure 14. Predicted alignment of Tacotron2 model at 30000 steps

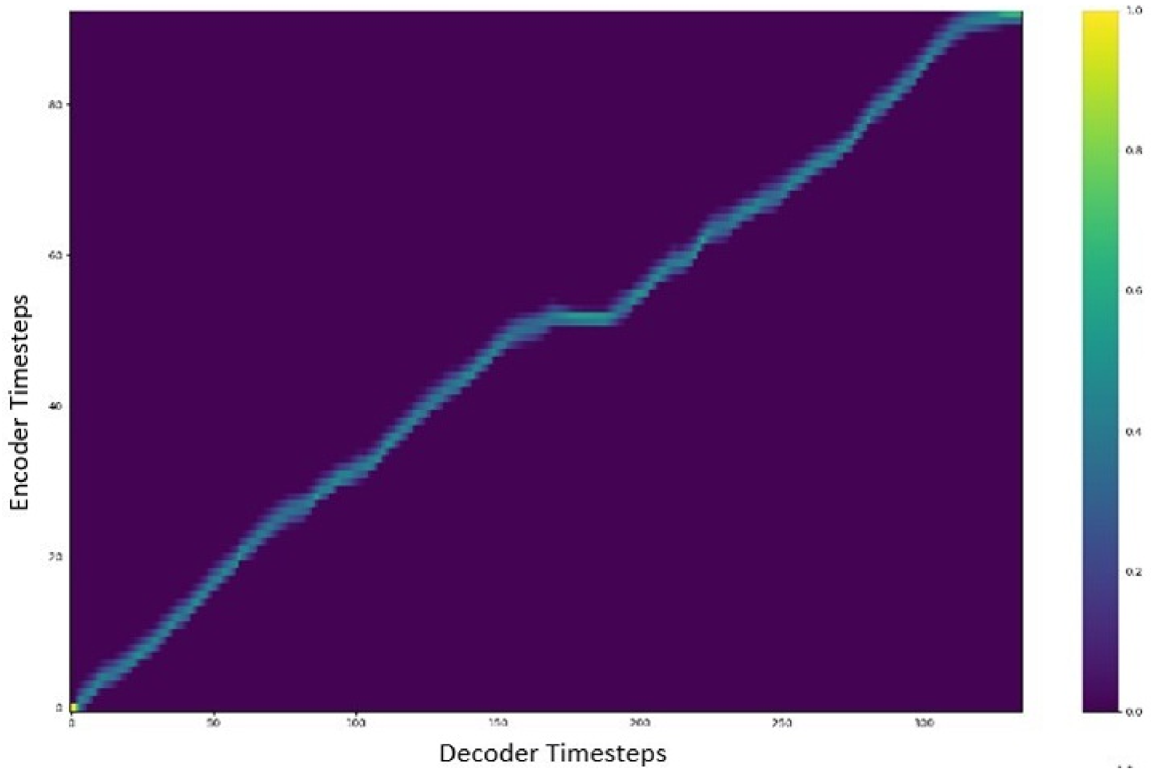


Figure 15. Predicted alignment of Tacotron2 model at 60000 steps

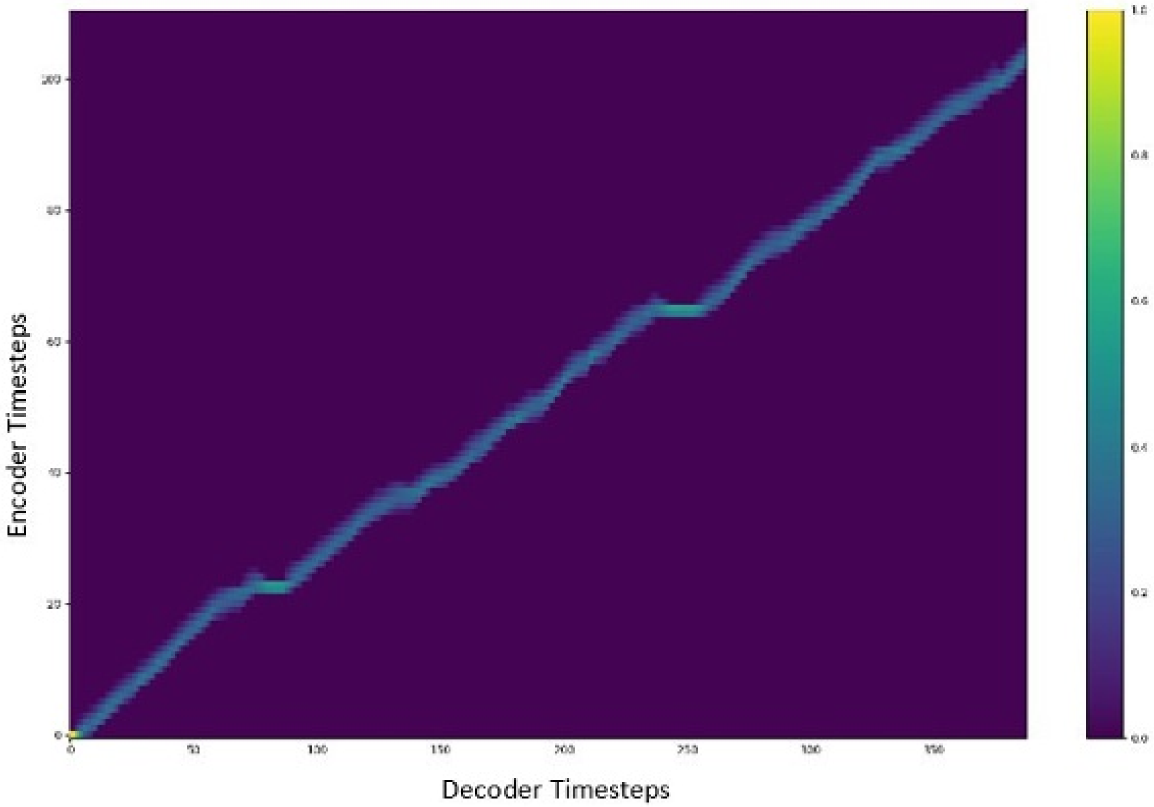


Figure 16. Predicted alignment of Tacotron2 model at 125000 steps

The training of VITS took nearly 69,000 steps over seven days and 8 hours. Figures 17, 18, and 19 illustrate the progression of the predicted alignments produced by the trained model. The total training loss for the VITS dropped from 3.6795 to 0.2425. The subjective evaluation conducted during training revealed that while the VITS model produced a reasonably understandable and natural speech, it provided at a faster pace than expected.

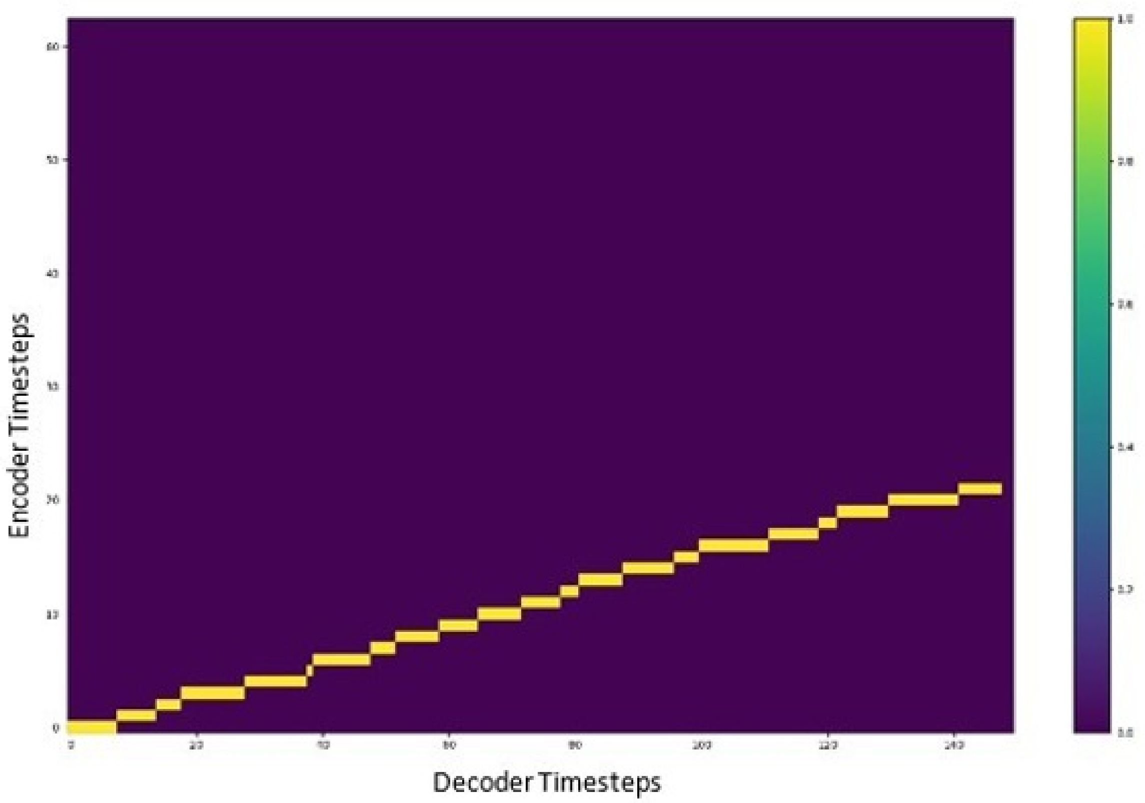


Figure 17. Predicted alignment of VITS model at 15000 steps

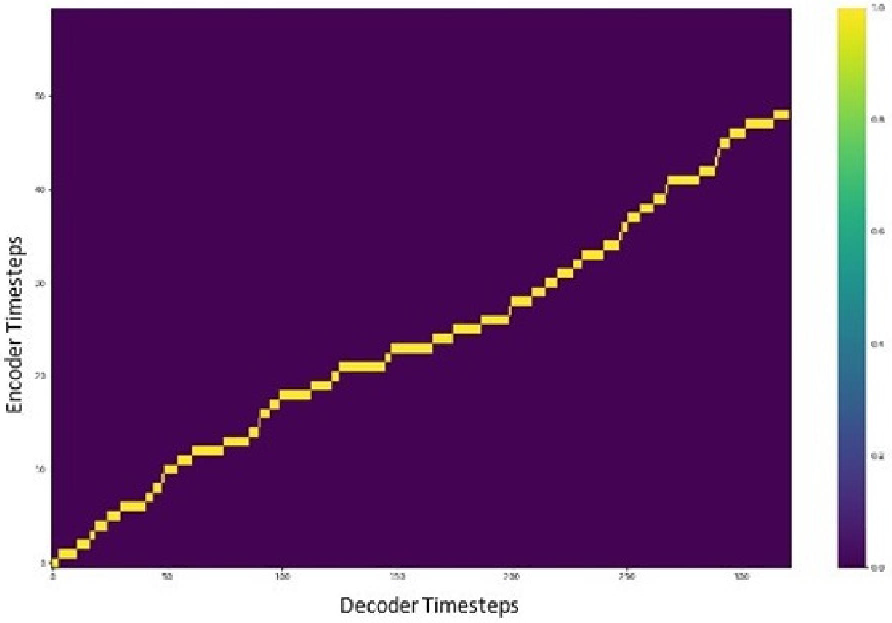


Figure 18. Predicted alignment of VITS model at 40000 steps

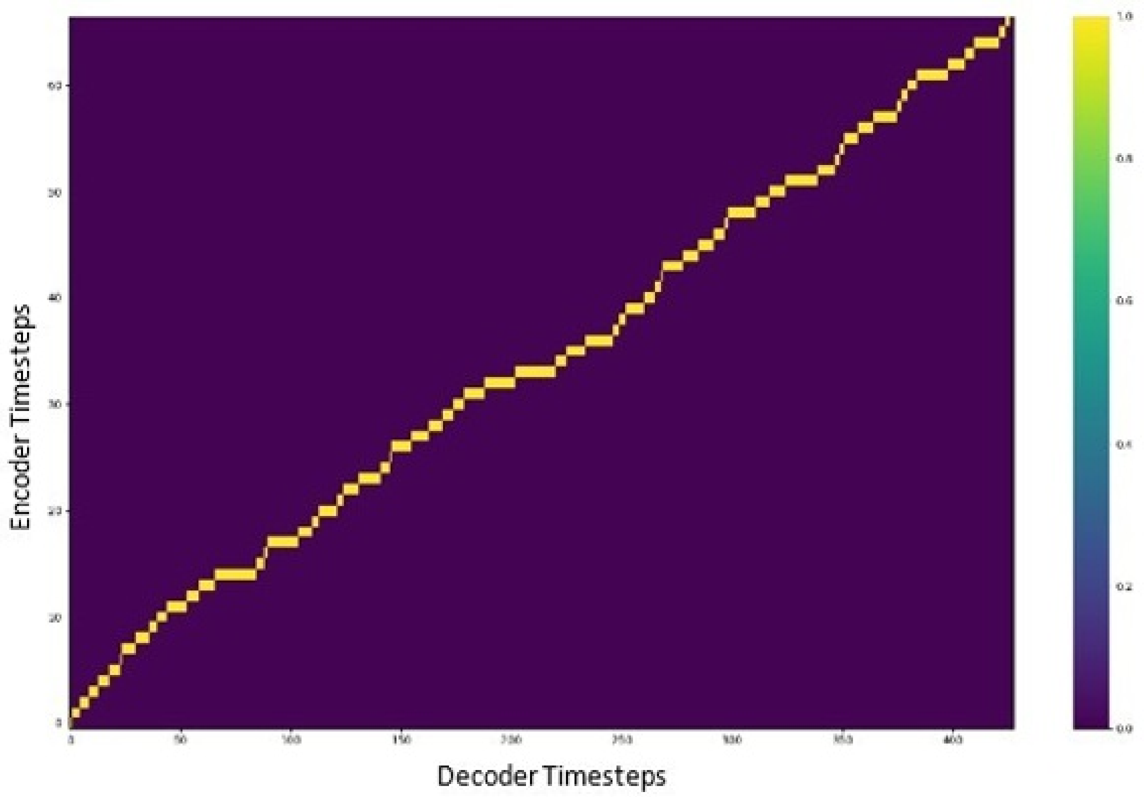


Figure 19. Predicted alignment of VITS model at 69000 steps

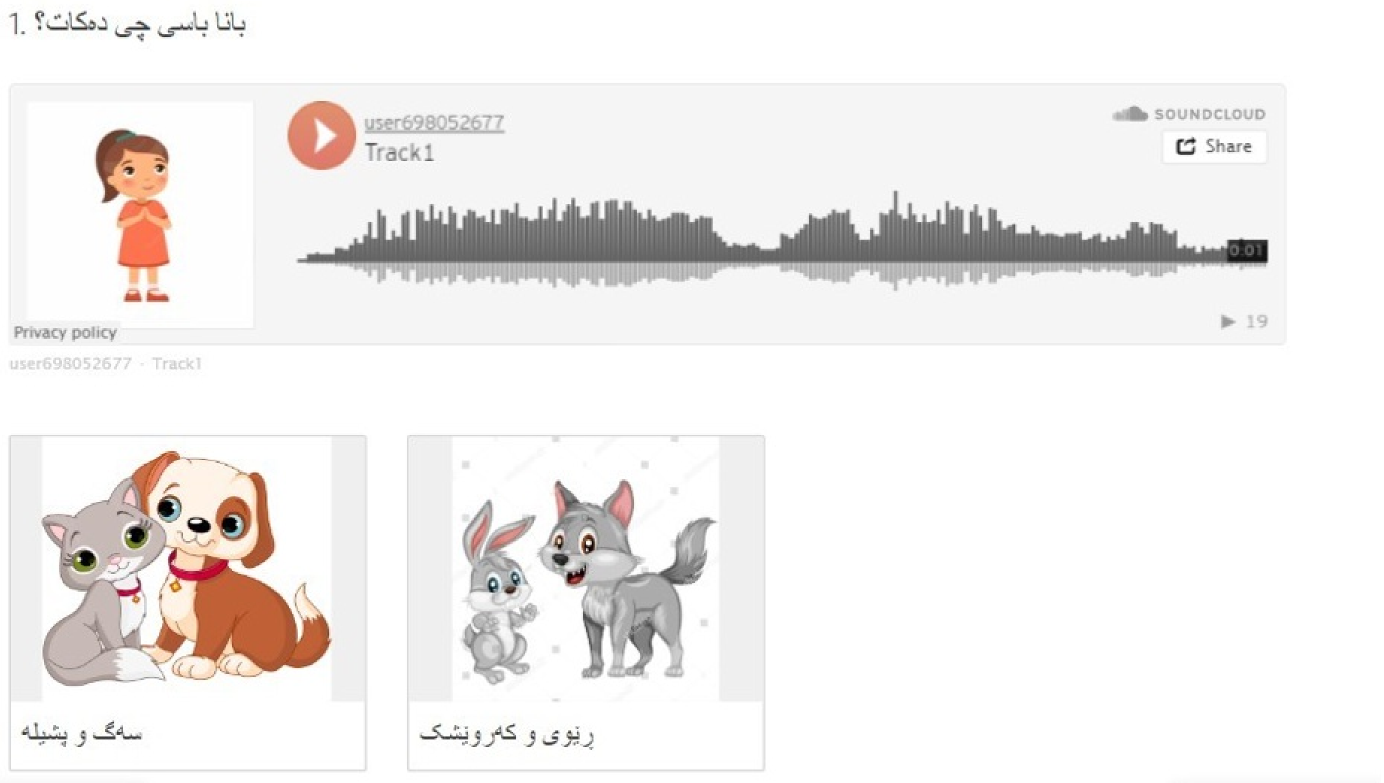
4.5. Testing and Evaluation

The objective evaluation measured and compared the MCD values of the two models using Python libraries. VITS achieved 7.91 and Tacotron2 9.89, which indicates that VITS outperforms Tacotron2 by nearly two points.

In the subjective evaluation, a group of kids, their parents, and teachers evaluated the models. For each group of evaluators, we created an evaluation form with age-appropriate language and question formats. We had 40 participants, 20 children and 20 adults (teachers and parents of the children). To design the children's survey, we consulted with four kindergarten teachers to choose appropriate vocabulary and a question format that would be simple 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 used SurveyMonkey for the interactive survey. The questionnaire had two sections: the first comprised six questions to assess the quality and intelligibility of the speech, and the second had four questions to evaluate the naturalness of the speech. Each model receives an equal number of questions. We randomly selected three questions from the first section and two from the second. We uploaded samples and made them accessible to the evaluators. Figures 20 and 21 present a sample of each question type for the kids.

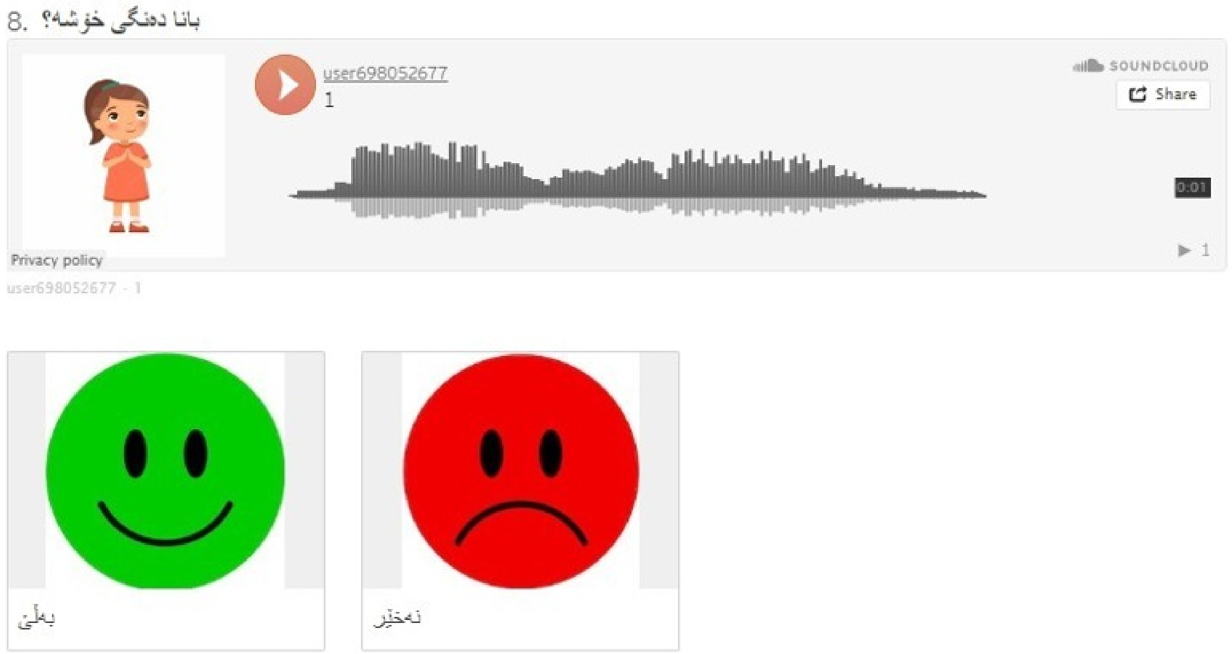


1. بانا باسی چی دەکات؟

سەگ و پشیلە

ڕێوی و کەروێشک

Figure 20. A Sample of the First Set of Questions for the Children’s Survey



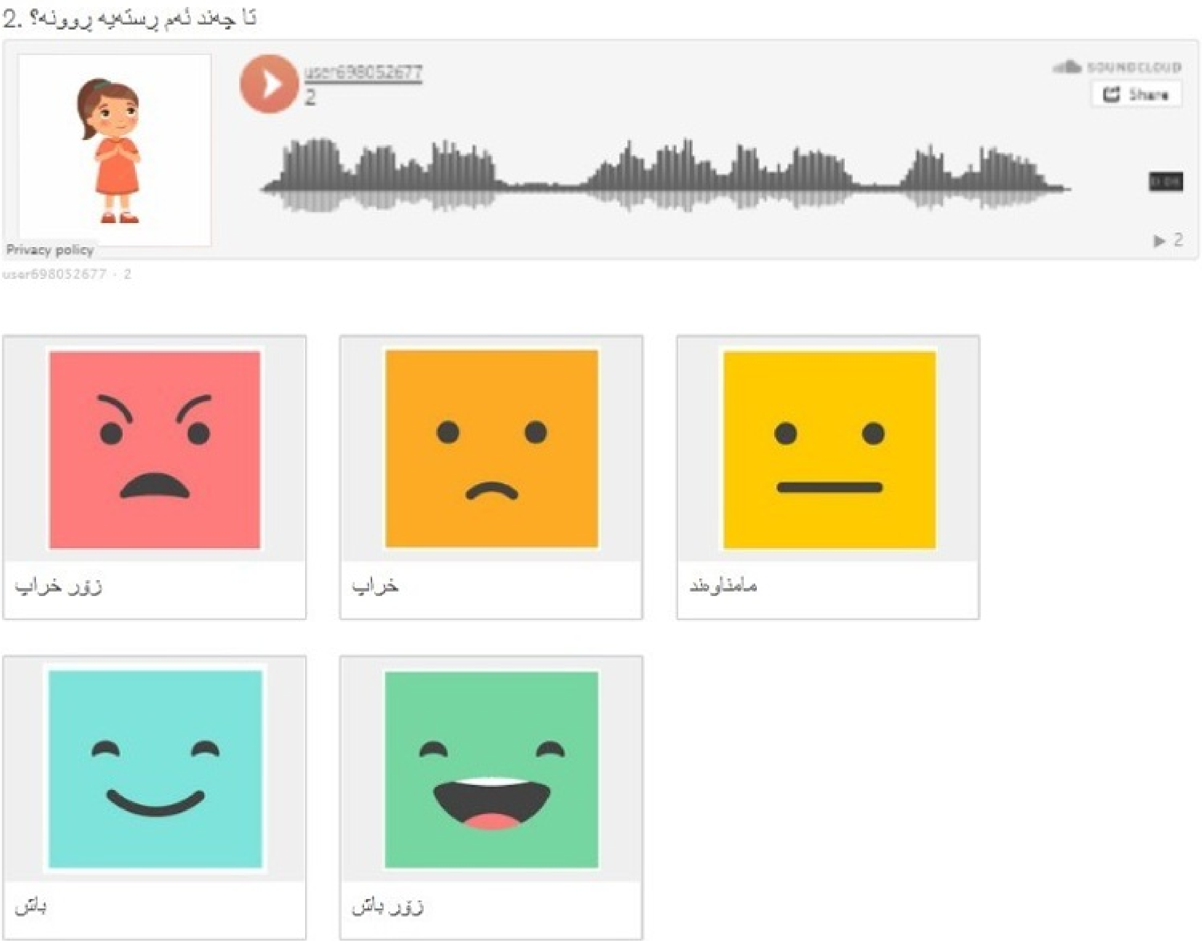
8. بانا دەنگی خۆشە؟

بەڵێ

نەخێر

Figure 21. A Sample of the Second Set of Questions for the Children’s Survey

Figures 22 and 23 show a sample of each question type for the parents and teachers, respectively.



2. تا چەند ئەم ڕستەیە ڕوونە؟

زۆر خراپ

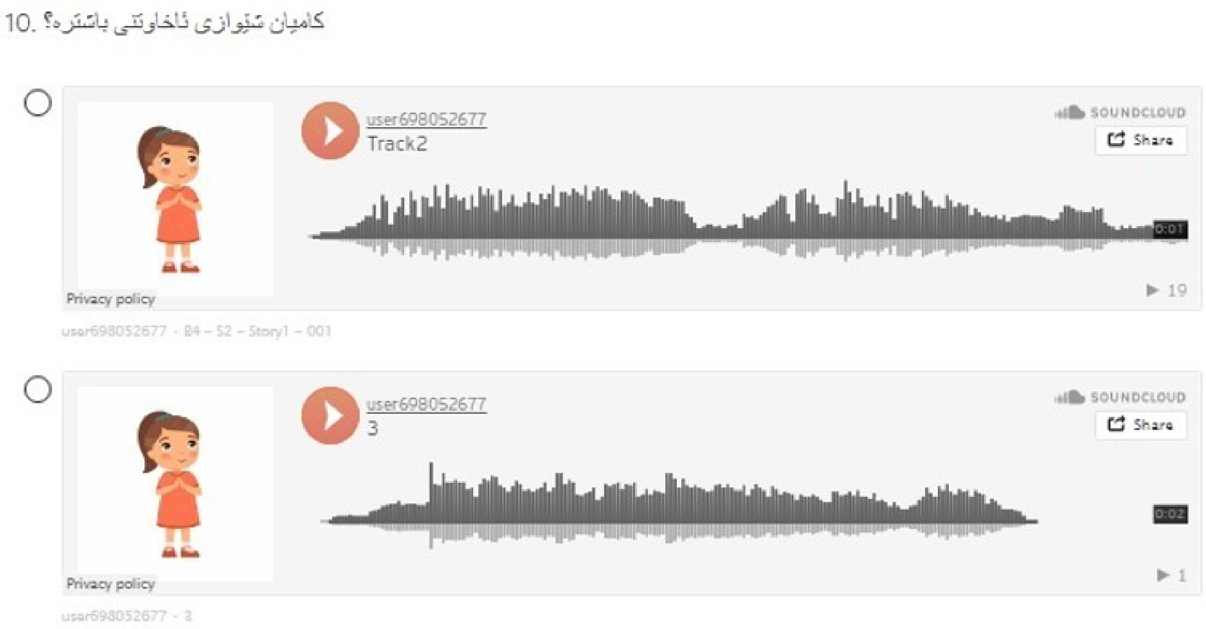
خراپ

مامناوەند

باش

زۆر باش

Figure 22. A Sample of the First Set of Questions for the Parents' and Teachers' Survey



10. کامیان شێوازی ئاخاوتنی باشترە؟

○

○

Figure 23. A Sample of the Second Set of Questions for the Parents' and Teachers' Survey

The subjective evaluation showed that VITS outperforms the Tacotron2 model for their naturalness and intelligence. It also showed that the children had more difficulty comprehending sentences than single-word audio files. Figure 22 presents the survey results for the children and parents group, respectively, and Table 5 presents their MOS score.

Table 5. MOS Scores

	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

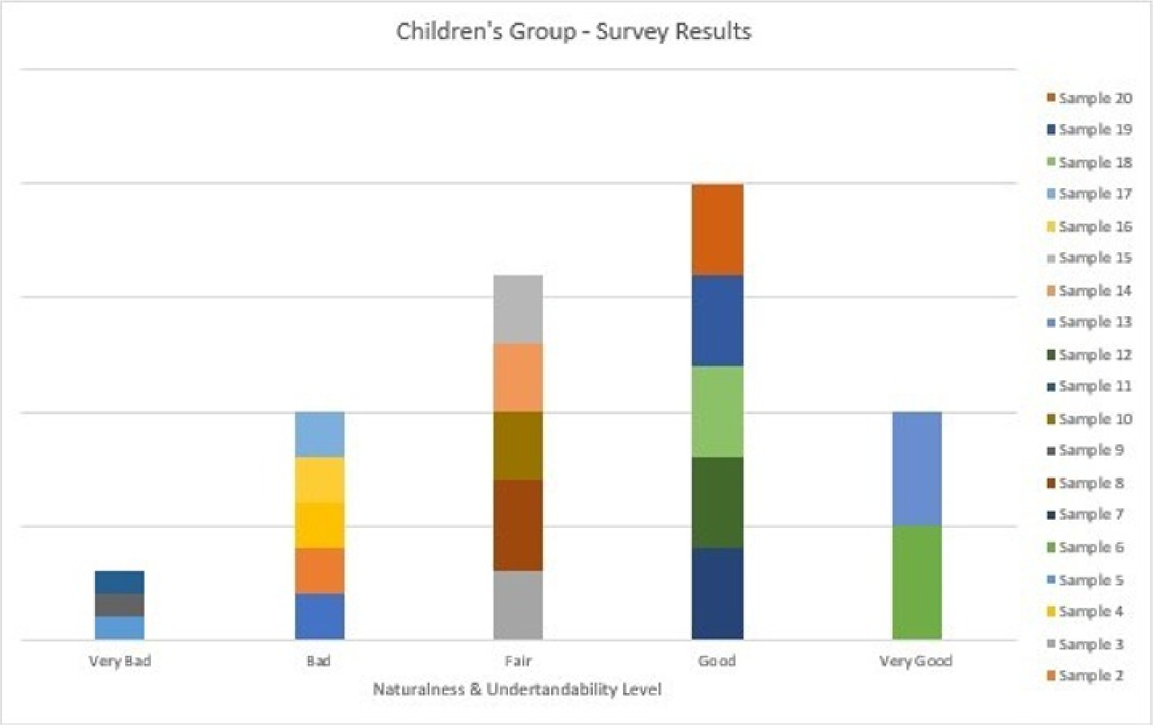


Figure 24. Children's Group - Survey Results

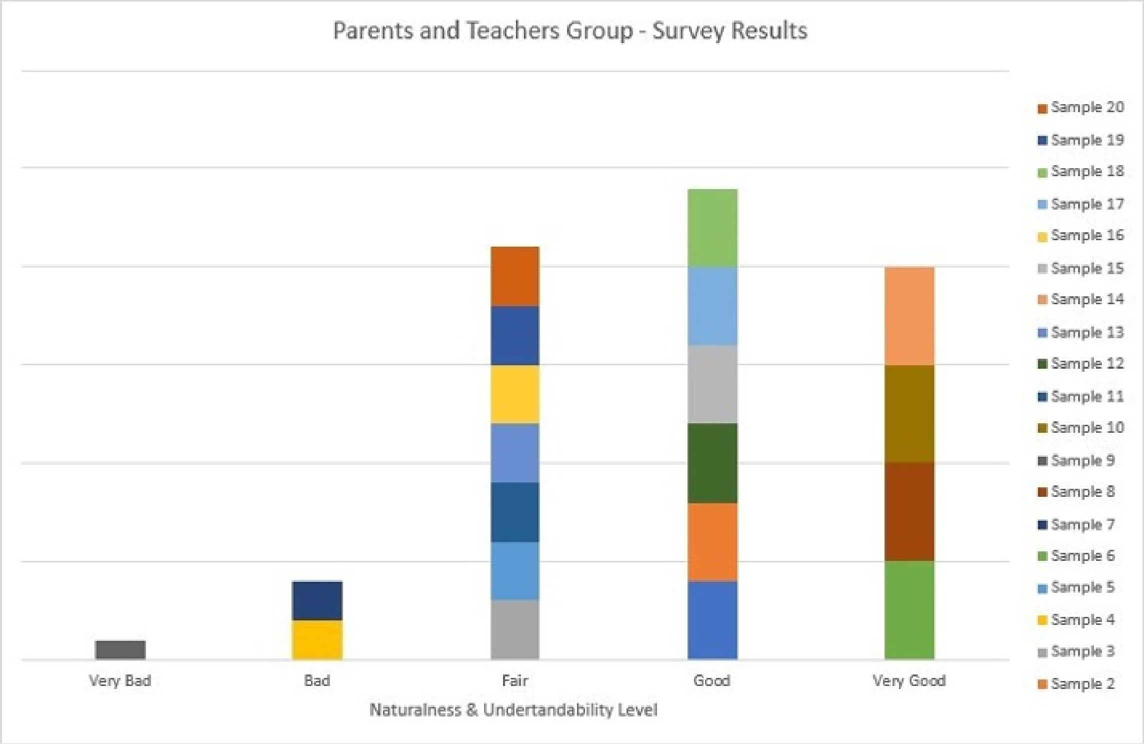


Figure 25. Parents' and Teachers' Group - Survey Results

4.6. Discussion

The evaluations showed that the speech generated by the VITS model is overall more understandable and natural-sounding than that generated by the Tacotron2 model.

We also compared the two models based on training time, evaluation loss, and alignment prediction during training. A good alignment is typically observable through a straight slope between the steps of the Y-axis encoder and its corresponding X-axis decoder. From Figures 14, 15, and 16, we can observe that the slop of the predicted alignment for the Tacotron2 model improves slightly throughout the training, although the slop appears to have been smoothed out and not deemed a good alignment slop. On the other hand, if we examine the predicted alignment between the two models, we find a significant difference between them. Compared to the VITS model, the alignment slope of the Tacotron2 model is smoother and has a smaller slope. That suggests that Tacotron2 has not yet predicted an accurate alignment. It also indicates that Tacotron2 needs additional training samples to produce more accurate results. For instance, to better align the samples and produce understandable 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 ability to keep 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 of a sentence delivery or the amount of time the model needs to generate frames, we must derive this from the training data, and the model must independently choose when to stop speech generation. That 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 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 recap, we found that training VITS on Sorani when the training dataset is small is more appropriate. 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

This study aimed 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, and it contains 209 stories from six children's books, divided into 4149 segments of speech, each between 0.1-10 seconds containing approximately 1-25 words. We trained two models, Tacotron2 and VITS, and evaluated them. The results suggested that a high-quality dataset could compensate for the lack of a large dataset in training text-to-speech models. Also, they showed that for the model to converge more quickly, the speakers must maintain a balanced speaking style and rate during the recording. Furthermore, we found that due to the lack of linguistic tools for aligning text and audio datasets, speech segmentation and alignment with the text could be the most labor-intensive tasks of the entire process.

The evaluation used MCD for the objective assessment and MOS for the subjective one. The latter used 20 samples from both models with two groups of participants: first, a group of children, and second, a group of teachers and parents. To facilitate the evaluation for the kids, we designed an interactive survey. The results showed that VITS performed better than Tacotorn2, with scores 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 for its naturalness and intelligibility on both 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, preferably from multiple speakers, the bare minimum needed to train Tacotron2. Also, the effort and time necessary for aligning and segmenting the data collection continues to be the major bottleneck, so developing the fundamental tools to automate speech segmentation and alignment becomes crucial. Finally, extending the data collection efforts to include other Kurdish dialects, such as Kurmanji (Northern Kurdish), is essential.

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.
- Diener, L., Janke, M., and Schultz, T. (2015). Direct conversion from facial myoelectric signals to speech using deep neural networks. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. 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.

- 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.
- 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*.