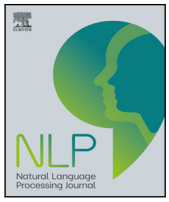
Natural Language Processing Journal 8 (2024) 100096

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp



Kurdish end-to-end speech synthesis using deep neural networks

Sabat Salih Muhamad [a], Hadi Veisi [b,*], Aso Mahmudi [c], Abdulhady Abas Abdullah [d], Farhad Rahimi [e]

[a] Computer Science Department, Faculty of Science, Soran University, Soran, Kurdistan, Iraq
[b] College of Interdisciplinary Science and Technologies, University of Tehran, Tehran, Iran
[c] School of Computing and Information Systems, Faculty of Engineering and Information Technology, The University of Melbourne, Melbourne, Australia
[d] Artificial Intelligence and Innovation Centre, University of Kurdistan Hewlêr, Erbil, Kurdistan Region, Iraq
[e] Department of Computer Engineering, Faculty of Engineering and agriculture, Islamic Azad University Arak Branch, Arak, Iran

ARTICLE INFO

Keywords:
Central Kurdish language
Deep learning
Speech synthesis
Tacotron2
End-to-end
WaveGlow vocoder

ABSTRACT

This article introduces an end-to-end text-to-speech (TTS) system for the low-resourced language of Central Kurdish (CK, also known as Sorani) and tackles the challenges associated with limited data availability. We have compiled a dataset suitable for end-to-end text-to-speech that includes 21 h of CK female voice paired with corresponding texts. To identify the optimal performing system, we employed Tacotron2, an end-to-end deep neural network for speech synthesis, in three training experiments. The process involves training Tacotron2 using a pre-trained English system, followed by training two models from scratch with full and intonationally balanced datasets. We evaluated the effectiveness of these models using Mean Opinion Score (MOS), a subjective evaluation metric. Our findings demonstrate that the model trained from scratch on the full CK dataset surpasses both the model trained with the intonationally balanced dataset and the model trained using a pre-trained English model in terms of naturalness and intelligibility by achieving a MOS of 4.78 out of 5.

1. Introduction

Speech synthesis, commonly referred to as Text-To-Speech (TTS), is a technology that converts input text into speech (Li et al., 2021), a natural-sounding voice that can communicate with humans (Mundada et al., 2014). The earliest approach for TTS was a rule-based technique for characterizing the resonance frequencies of the vocal tract, called formant synthesis, which was used for a long time. This strategy uses a language output source-filter model. An artificial speech waveform is produced using a combination of variables throughout time, including fundamental frequencies, voicing, and noise levels (Kayte et al., 2015). Concatenative synthesis by unit selection, a method of joining smaller units of previously recorded waveforms, was state-of-the-art for a while Black and Taylor (1997), Hunt and Black (1996). Concatenative methods for speech synthesis have several drawbacks: (a) they require large databases to cover various unit sizes, (b) noise captured during unit recording can degrade the quality of synthesized speech, as the recorded units are used as-is in synthesis, (c) they demand extensive labeling and recording efforts, and (d) they offer low flexibility in modifying the generated wave signals (Fahmy et al., 2020). Later, statistical parametric speech synthesis was introduced (Ze et al., 2013; Zen et al., 2009), which eliminated many of the boundary artifacts associated with concatenative synthesis. This approach created smooth trajectories of speech features, which could then be directly synthesized by the vocoder. Hidden Markov Model (HMM)-based synthesis has several downsides, including (a) requiring a lot of feature engineering and domain expertise and (b) producing speech that sounds more robotic than speech produced via unit selection voice synthesis.

In recent years, TTS methods relying on end-to-end neural network architecture have dominated both the market and research community (Sotelo et al., 2019). Tacotron, employing a sequence-to-sequence architecture (Sutskever et al., 2014), represents a fully end-to-end model capable of instantaneously converting input text into a Mel-spectrogram (Ning et al., 2019). Its successor, Tacotron2, builds upon this foundation. Tacotron2 comprises two essential components: (a) a recurrent seq-to-seq generative model with attention mechanisms, and (b) a modified WaveNet serving as a vocoder for generating voice signals.

The development of voice synthesis and other natural language processing (NLP) applications in some languages has been significantly hindered by the lack of resources, such as speech and text corpora. Kurdish language, which is spoken by millions across Western Asia, mainly in Turkey, Iraq, Iran, and Syria is divided into three major branches:

* Corresponding author.
E-mail addresses: sabat.muhamad@soran.edu.iq (S.S. Muhamad), h.veisi@ut.ac.ir (H. Veisi), amahmudi@student.unimelb.edu.au (A. Mahmudi), abdulhady.abas@ukh.edu.krd (A.A. Abdullah), fa.rahimi@mci.ir (F. Rahimi).

https://doi.org/10.1016/j.nlp.2024.100096
Received 14 December 2023; Received in revised form 30 May 2024; Accepted 3 August 2024

Central (also known as Sorani), Northern (also known as Kurmanji), and Southern Kurdish. Despite its widespread usage, Kurdish faces a shortage of linguistic resources crucial for computer processing.

Previous efforts in Kurdish TTS have relied on outdated techniques like concatenation. However, due to the scarcity of language resources and linguistic expertise in Kurdish, feature engineering for such approaches is a challenging task that demands extensive language knowledge. To circumvent this issue, we aim to leverage deep neural networks to map Kurdish language features directly to acoustic data with minimal human intervention. Our deep neural network-based TTS system takes Kurdish text as input, eliminating the need for an external phoneme dictionary or a pre-trained grapheme-to-phoneme model.

In this article, we focused on utilizing end-to-end deep learning techniques to enhance the naturalness of a Kurdish TTS system. Since text and speech corpora are necessary for acoustic modeling and developing a practical TTS system (Veisi et al., 2022), our primary goal was to create and collect a CK dataset and create the naturalness of a Kurdish text-to-speech system based on the Tacotron2 neural network architecture. This study shows how to synthesize Mel-spectrograms from CK text as an intermediate feature representation, then utilize a WaveGlow architecture as a vocoder to produce a high-quality Kurdish voice using a modified deep architecture from Tacotron2.

Our research seeks to innovate beyond existing TTS models like Tacotron2 and WaveGlow by integrating novel strategies to enhance performance and adaptability, especially for CK. We started our work with transfer learning, fine-tuning pre-trained English models on a small, high-quality CK dataset to grasp phonetic and linguistic nuances of the language. Subsequently, we used domain adaptation techniques to refine the model, improving performance and reducing data requirements.

We developed a custom symbol set to accurately represent CK phonetics, enhancing computational efficiency. Our prosody-aware model incorporates rhythm, stress, and intonation for expressive speech synthesis. Additionally, an adaptive noise reduction mechanism in the WaveGlow vocoder improves speech clarity.

These innovations significantly enhanced our TTS system's naturalness and intelligibility for CK and provide a scalable framework for other low-resource languages, advancing the field of speech synthesis technology.

The rest of this work is divided into six sections. The work with text-to-speech technology is described in Section 2. The creation of the Kurdish speech corpus is explained in Section 3. The architecture of the applied model for the Kurdish text-to-speech system is explained in Section 4. The experimental results are presented in Section 5, and the study is concluded in Section 6.

2. Related works

2.1. TTS in other languages

In recent years, deep learning has emerged as a transformative approach within the field of machine learning. A significant development in this domain is the introduction of WaveNet, a deep neural network designed to generate raw audio waveforms, as presented in van den Oord et al. (2016). This model was trained and tested using 24.6 h of English data and 34.8 h of Mandarin Chinese data, achieving a Mean Opinion Score (MOS) of 4.0.

Subsequently, Tacotron, an end-to-end text-to-speech (TTS) model, was proposed by the authors in Wang et al. (2017). The model was trained on 24.6 h of North American English speech data, attaining a MOS of 3.82.

Additionally, the researchers in Sotelo et al. (2019) introduced Char2Wav, another end-to-end voice synthesis model. Char2Wav comprises two components: the readers and a neural vocoder. The system utilized normalized WORLD vocoder features both as targets for the readers and as inputs for the neural vocoder. The model was trained using the VCTK and DIMEX-100 datasets.

Deep Voice 2 (Arik et al., 2017) represented an upgraded architecture from Deep Voice 1, introducing multi-speaker capabilities via speaker embeddings. The model was trained on two datasets: 20 h of English single-speaker speech and 44 h of multi-speaker VCTK data. Utilizing an 80-layer WaveNet vocoder, Deep Voice 2 achieved a MOS of 3.53.

Subsequently, Deep Voice 3 (Ping et al., 2018) was proposed as a fully convolutional attention-based sequence-to-sequence model, with a comparative analysis against recurrent neural network (RNN) architectures. Various waveform generation techniques were explored, with WaveNet consistently yielding superior performance. The model was trained independently on three datasets: 20 h of English single-speaker speech, 44 h of multi-speaker VCTK, and 820 h of LibriSpeech data. When combined with the WaveNet vocoder, Deep Voice 3 obtained an MOS of 3.78.

Tacotron-2 (Shen et al., 2018a) was trained on 24.6 h of USA-English speech data, achieving a mean opinion score (MOS) of 4.53. ClariNet (Ping et al., 2019) introduced a novel parallel wave generation approach based on the Gaussian inverse autoregressive flow (IAF), representing the first fully convolutional text-to-wave neural network for speech synthesis, with an MOS of 4.15.

In the context of rhythmic and natural Chinese voice synthesis, Zhang et al. (2019) proposed a Tacotron model that incorporated prosodic annotations, trained on 10.38 h of the BZSYP database. Their approach outperformed the baseline system trained without prosodic annotations. Additionally Lu et al. (2019) explored end-to-end Chinese speech synthesis using Tacotron2 and WaveNet vocoder, trained on 31 h of Chinese data, yielding a statistically significant improvement (p-value = 0.001).

Authors in Li et al. (2019) presented a transformer network for neural speech synthesis, converting all text inputs to phonemes. They trained their model using a 25-hour dataset of US English female speech, achieving 4.39 by using MOS evaluation.

A rapid and robust approach similar to transformers, known as FastSpeech, was proposed by Ren et al. (2019). Using the 24-hour LJSpeech dataset, FastSpeech achieved a MOS of 3.84. However, FastSpeech had some limitations, such as a complex and time-consuming teacher–student distillation pipeline, imprecise length predictions from the teacher model, and information loss in target Mel-spectrograms due to data simplification. These issues affected the overall voice quality.

Authors of Ren et al. (2022) presented FastSpeech-2, a quick and high-quality end-to-end text-to-speech model. The model was trained on the LJSpeech dataset, which contained about 24 h of speech dataset. FastSpeech 2 achieved a MOS of 3.83.

Deepmind developed EATS-end-to-end adversarial text-to-speech generative model (Donahue et al., 2022). They utilized a private dataset comprising 260.49 h of speech from 69 male and female North American English speakers. The model achieved a MOS of 4.083.

In Vainer and Dušek (2020) SpeedySpeech, a convolutional system for generating phoneme-based spectrograms, was developed. This system enables quick training and synthesis while maintaining superior voice quality compared to robust baseline methods. The model was trained and tested on the LJSpeech dataset, which includes 13,100 text-audio pairs. To evaluate the model, a survey based on MUSHRA (Schoeffler et al., 2018) was conducted with 40 participants. The results showed that SpeedySpeech, with MelGan, significantly outperformed Tacotron2, achieving an average score of 75.24.

Authors of Liu et al. (2020a) suggested the novel two-task learning scheme. 17 h of Mongolian speech data and TH-CoSS (TsingHua-Corpus of Chinese Speech-Synthesis) were used. They used the Griffin-Lim algorithm for waveform generation in all schemes. They achieved a MOS of 3.91 for Chinese and a MOS of 3.83 for Mongolian. In Liu et al. (2020b), for the neural end-to-end TTS framework, they present a Tacotron-2-KD (knowledge-distillation) framework, the teacher

-student-training system. They utilized Chinese and English datasets. The suggested approach reached a MOS of 3.93 for English and a MOS of 3.94 for Chinese. He et al. (2020) suggested the DOP Tacotron module. 12 h of the biaobei speech corpus in Mandarin were utilized in their experiment. 50 sentences were chosen to evaluate the model. The MOS of DOP Tacotron is 3.683, which is higher than that of the original Tacotron.

In Hayashi et al. (2020), ESPnet-TTS, an expansion of the free and open-source ESPnet speech processing tools, was released. Their toolbox supports E2E-TTS models in addition to a number of TTS recipes with a design that is consistent with automatic speech recognition (ASR) recipes, which offers good repeatability. The MOS for their model on the LJSpeech dataset was 4.25. In Fahmy et al. (2020), Tacotron2 is used to create high-quality and human-like Arabic speech. They trained the model on 2.41 h of the Nawar Halabis Arabic dataset by utilizing the pre-trained English model. They achieved a MOS of 4.21. In Weiss et al. (2021), the Wave-Tacotron sequence-to-sequence neural network introduced which directly converts text inputs into voice waveforms. The approach extends the Tacotron model by adding a normalizing flow into the autoregressive decoder loop. Hundreds of samples are included in each of the output waveforms' non-overlapping, fixed-length blocks. Two single-speaker datasets were used to train and test their model, including the public LJ speech dataset and a private dataset with around 39 h of speech. A MOS of 4.47 was attained. A TTS system based on Tacotron2 was proposed by Naderi et al. (2022) for Persian. They created 21 h of Persian speech dataset to train their model. They obtained different values when evaluating their model from MOS 3.01 to MOS 3.97.

Tacotron2 was utilized by Win and Masada (2020), who trained their model on 5 h of Myanmar corpus. Their result was MOS = 3.89. Similarly, Tacotron2 and DeepVoice 3 were implemented on low resource Afan Oromo dataset by Shifera (2021). As a result, they found that the Tacotron2 model with an MOS of 4.32 on 5 outperformed the DeepVoice 3 model with an MOS of 3.28.

The problems with human-level quality in TTS were the subject of a thorough investigation in Tan et al. (2022). To reach human-level quality, they developed a TTS system called NaturalSpeech. Experiment evaluations on the well-known LJSpeech dataset reveal that their proposed NaturalSpeech achieves 0.01 CMOS (comparative mean opinion score) to human recordings at the sentence level, with Wilcoxon signed rank test at p-level 0.05, which for the first time on this dataset demonstrates no statistically significant difference from human recordings.

2.2. Kurdish TTS works

Research on Kurdish TTS is still in its early stages compared to other languages. Kurdish, a language within the Indo-European language family (Thackston, 2006), utilizes two scripts: a modified Arabic alphabet and a modified Latin alphabet (Sejnowski and Rosenberg, 1987).

Several synthesis models for Kurdish TTS, including allophone, syllabic, and diphone-based approaches, were developed (Bahrampour et al., 2009). The allophone-based approach yielded the poorest quality and proved the most challenging to implement. In contrast, the syllable-based approach demonstrated excellent overall quality and intelligibility. However, the diphone-based TTS system provided the highest quality among the three. In the same year, a comparative study of these three Kurdish TTS systems—utilizing concatenation—was conducted in Barkhoda et al. (2009). The diphone-based TTS system achieved the highest quality, with a Diagnostic Rhyme Test (DRT) score of 97%. Additionally, in Daneshfar et al. (2009), an allophone unit was used in a concatenative synthesis approach, which also scored well for intelligibility.

A further enhancement in natural-sounding CK speech synthesis was achieved through a concatenative synthesis approach utilizing diphone units to smooth transitions between phonemes, as detailed in Hassani et al. (2011). This method resulted in speech with high intelligibility ratings.

In a prior work (Muhamad and Veisi, 2022), a Central Kurdish TTS system was developed using transfer learning from an English pre-trained model, supplemented by 10 h of CK speech data. As we will describe in the next sections, we similarly trained our first model using transfer learning from an English pre-trained model, but we utilized a larger dataset (21 h) and we employed WaveGlow vocoder instead of HiFi-GAN to produce higher-quality, more natural-sounding speech. The relevant papers are summarized and their important points are shown in Table 1.

3. Creation of central Kurdish speech corpus

Speech corpus is one of the most important requirements for data-driven text-to-speech synthesis systems. A speech corpus is a collection of text and equivalent audio pairs. Since there were no previously available large-scale speech Kurdish corpora, for this study, we created a new one from scratch. In this section, we will explain the details of the design, compilation, and challenges of this process.

3.1. Text data collection

In this study, we aimed to construct a phonetically rich and balanced CK speech corpus by gathering a large amount of raw text from diverse sources. The resulting corpus represents a wide array of linguistic features and domains, enhancing the quality of the TTS system. Our primary sources for the raw text compilation were various media outlets such as www. Rudaw.net, www.gksat.tv, www.xendan.org, www.nrttv.com, and www.kurdistantv.net, along with select textbooks on General Psychology and Psycholinguistics.

From these sources, we collected a total of 10,979 utterances, spanning 14 distinct categories including news, sports, linguistics, psychology, poetry, health, scientific topics, general knowledge, interviews, politics, education, literature, narratives, tourism, and miscellaneous subjects. The deliberate selection of these diverse categories ensures comprehensive coverage of CK sentences, aligning with the multi-faceted applications of TTS systems. Table 2 provides a breakdown of the utterance categories and the corresponding counts, illustrating the breadth of subjects covered in our corpus.

Furthermore, we curated a test set comprising 110 sentences sourced from a diverse array of texts representing 17 distinct subject areas. These sentences were carefully chosen to ensure differentiation from those in the training set. Initially gathered from various websites, the selected phrases underwent further refinement. The topics of the chosen test sentences are listed in Table 3, along with the number of sentences for each category.

Subsequently, the texts underwent normalization in accordance with CK orthographic standards Automated_Kurdish_Text_ Normalization. Additionally, non-Kurdish (Arabic and English) words within the texts were transliterated. Cardinal and fractional numbers, dates, times, and currencies were then substituted with their Kurdish equivalents utilizing an open-source library.[1] For instance, the date "٢٠٢٢-٩-١" was transformed into "یەکی نۆی دوو هەزار و بیست و دوو". A few samples of normalized Kurdish text are shown in Table 4.

3.2. Audio recording

The collected sentences were recorded using professional voice recording equipment in a studio environment. The speaker, a female in her thirties originally from Sulaimania city, delivered the sentences phonetically close to the dialect of that city. The audio files were recorded in 22050 Hz sampling rate, 16-bits depth, mono-channel, and stored as .wav format. The audio recording process took 41 days. After

[1] github.com/AsoSoft/AsoSoft-Library

Table 1
Summarizing the main points of the literature review.

No.	Ref.	Year	Method	Dataset	Result
1	van den Oord et al. (2016)	2016	WaveNet	-North American English (24.6 h) -Mandarin-Chinese (34.8 h)	MOS 4.0
2	Wang et al. (2017)	2017	Tacotron	-North American English dataset	MOS 3.82
3	Sotelo et al. (2019)	2017	CHAR2WAV	The VCTK and DIMEX-100 datasets	Their result was sufficient
4	Arik et al. (2017)	2017	Deep Voice 2- Multi speaker neural TTS	English single-speaker (20 h) VCTK dataset multi-speaker (44 h)	MOS 3.53
5	Ping et al. (2018)	2017	Deep Voice 3- a fully convolutional attention-based TTS	English single-speaker (20 h) VCTK dataset multi-speaker (44 h) Librispeech multi-speaker (820 h)	MOS 3.78
6	Shen et al. (2018b)	2018	Tacotron-2	US English (24.6 h)	MOS 4.526
7	Ping et al. (2019)	2019	Clarinet: Parallel WaveNet.	An internal-English-speech dataset (20 h)	MOS (4.15)
8	Zhang et al. (2019)	2019	Tacotron	BZSYP-Chinese database (10,000 audio samples).	84%
9	Lu et al. (2019)	2019	Tacotron-2 and Wavenet vocoder.	Chinese dataset (31 h)	Significant (p value = 0.001)
10	Li et al. (2019)	2019	Transformer-based TTS	US English dataset	MOS 4.39
11	Ren et al. (2019)	2019	FastSpeech- non-autoregressive End to End	LJSpeech dataset (24 h)	MOS 3.84
12	Ren et al. (2022)	2020	FastSpeech 2- End to End TTS	LJSpeech dataset (24 h)	MOS 3.83
13	Donahue et al. (2022)	2020	EATS-end-to-end adversarial text to speech	North American English dataset (260.49 h)	MOS 4.083
14	Vainer and Dušek (2020)	2020	SpeedySpeech	LJSpeech dataset	75.24
15	Liu et al. (2020a)	2020	MTL-Tacotron	TH-CoSS (TsingHua- Chinese Corpus). (9 h) Mongolian speech data (17 h)	MOS-Chinese 3.91 MOS-Mongolian 3.83
16	Liu et al. (2020b)	2020	Tacotron-2-KD	English and Chines datasets	MOS-English 3.93 MOS-Chinese 3.94
17	He et al. (2020)	2020	DOP-Tacotron	Biaobei speech corpus, Mandarin (12 h).	MOS 3.683
18	Hayashi et al. (2020)	2020	ESPnet-TTS	LJSpeech dataset	MOS 4.25
19	Fahmy et al. (2020)	2020	Tacotron2- Transfer Learning	Nawar Halabi's Arabic Dataset (2.41 h)	MOS 4.21
20	Win and Masada (2020)	2020	Tacotron2	Myanmar corpus (5 h)	MOS 3.89
21	Shifera (2021)	2021	Tacotron2 and Deepvoice3	Afaan Oromo (17 h)	MOS 4.32
22	Weiss et al. (2021)	2021	Wave-Tacotron	LJSpeech dataset private dataset (39 h)	MOS 4.47
23	Naderi et al. (2022)	2022	Tacotron2	Persian dataset (21 h)	MOS 3.01 – 3.97
24	Tan et al. (2022)	2022	NaturalSpeech	LJSpeech dataset	CMOS 0.01
25	Bahrampour et al. (2009)	2009	Concatenative (Allophone, Syllable, and Diphone)	Kurdish	Allophone MOS 2.45 Syllable MOS 3.02 Diphone MOS 3.51
26	Barkhoda et al. (2009)	2009	Concatenative (Allophone, Syllable, and Diphone)	Kurdish	Best quality score 3.5 Best DRT 97%
27	Daneshfar et al. (2009)	2009	Concatenative (Allophone)	Kurdish (2100 words)	Best quality score 2.4
28	Hassani et al. (2011)	2011	Concatenative (Diphone)	Kurdish (2100 words)	Best quality score 55%
29	Muhamad and Veisi (2022)	2022	Tacotron2-transfer learning	Kurdish (10 h)	MOS 4.10

Table 2
The total of the train utterances.

Category	No. of Utterances
linguistics	1760
questions and exclamation	1393
story	1092
poem	916
tourism	782
miscellaneous	700
sport	683
education and literature	619
news	608
science	543
health	483
politics	483
general information	461
interview	456
Total	10,979

Table 3
The test set sentences were distributed throughout many areas.

Topics	No. of Sentences
news	10
formal letter	10
sport	9
poem	8
questions	7
psychology	6
health	6
science	6
miscellaneous	6
general information	6
story	6
tourism	6
linguistics	5
interview	5
politics	5
education and literature	5
exclamation	4
Total	110

meticulously refining the recordings, the duration of the final speech data is 21 h. This collection is denoted as the Sabat speech corpus, serving as both our text and audio datasets for training and testing our model.

The lengths of recorded audio files are between one and twelve seconds. The average audio length is 6.89 s. Fig. 1 displays the distribution of the dataset according to the length.

After the recordings were made, we identified minor pronunciation or intonation errors, as well as dialectal variations. Rather than necessitating a return to the studio for speaker correction, we opted to adjust the text to match the speaker's voice. This involved modifying the spelling and punctuation of certain utterances to align with the speaker's articulation, intonation, and speech junctures.

Table 4
Examples of Kurdish text normalization.

Input text	Normalized text	English
صلى الله عليه وسلم	سەڵڵەڵاهو عەلەیهی وسەڵەم	sallallahou alayhe wasallam (Arabic blessing after mentioning prophet Muhammad)
iOS	ئای ئۆ ئێس	iOS (a mobile operating system)
١٥	پانزە	fifteen
٤.٦٣	چوار پۆینت شەست و سێ	four point sixty-three
١-١١-٢٠٢١	یەکی یانزەی دوو هەزار و بیست و یەک	first of the eleventh month of two thousand twenty-one

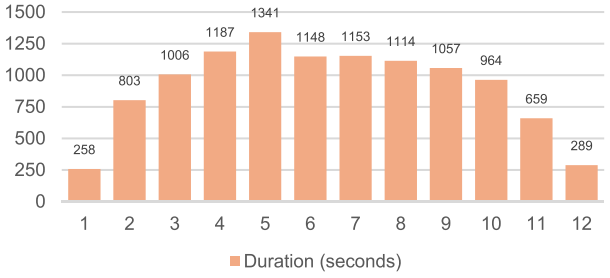


Fig. 1. Distribution of length of sentences of the dataset.

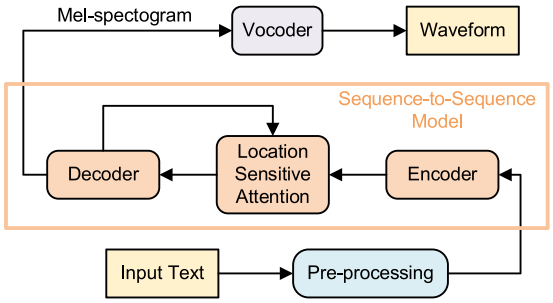


Fig. 2. Block diagram of our approach.

Table 5
Specifications of "Sabat Speech Corpus".

Feature	Sabat dataset
Total audio length	21 h (75,626 s)
Audio file format	.wav (22.05 kHZ, 16-bits, mono)
Text file format	.txt (UTF-8)
Sampling rate	22,050 Hz
Number of audio files	10,979
Longest audio length	12.97 s
Shortest audio length	1.01 s
Average audio length	6.89 s

Furthermore, since the audio files were recorded on different days, variations in volume were observed across some recordings. To ensure uniformity, manual adjustments were made to equalize the loudness of these recordings. Table 5 shows the general specifications of our final speech corpus.

4. Research method

End-to-end neural network designs offer a significant advantage over traditional voice synthesis approaches by eliminating the need for extensive domain expertise and labor-intensive feature engineering. These networks require minimal human annotation and can be trained to respond to any language, gender, or emotion. In contrast, traditional TTS synthesizers operate through multiple phases, each requiring independent training. This multi-stage process can lead to error propagation across stages. End-to-end architectures, being designed as a unified system, are inherently more resilient to such issues.

4.1. System architecture

Outlined in Fig. 2, our proposed methodology proceeds through several key stages. We begin by curating a dataset of "text, audio" pairs featuring a single female speaker, as detailed in Section 3. Through preprocessing steps, including text normalization, we prepared the input data for our TTS system. A sequence-to-sequence synthesis network based on Tacotron2 (Shen et al., 2018a) was used to predict the mel spectrograms, which are then transformed into audible sounds using the WaveGlow (Prenger et al., 2019) vocoder.

The choice of Tacotron2 as the base model for our speech synthesis system is driven by its efficient and effective architecture, which is well-suited to address the specific challenges of CK. Tacotron2 employs an end-to-end approach, utilizing only two stages: an acoustic model and a vocoder. This simplifies the training process and reduces the complexity typically associated with TTS systems that use multiple stages. The acoustic model, a recurrent neural network, predicts a sequence of Mel spectrograms from an input letter sequence, while the WaveGlow vocoder (Prenger et al., 2019) generates time-domain waveforms from these spectrograms. Given that CK has a nearly one-to-one correspondence between phonemes and letters, the simplicity of its phonetic structure makes the creation of a TTS system more straightforward and requires less data compared to languages with more complex phonetics. Tacotron2's ability to learn directly from speech and corresponding text, without the need for extensive phonetic annotations or multiple processing stages, makes it an ideal choice for developing a high-quality TTS system for CK.

As shown in Fig. 3, the Tacotron2 architecture is an encoder-attention-decoder paradigm that takes advantage of "location-sensitive attention". First, an encoder creates a word embedding vector from the input character sequence. From the embedding vector, the decoder generalizes the corresponding spectrograms. The WaveGlow vocoder constructs the real voice waveform from the predicted spectrograms produced by Tacotron2.

Tacotron2 and WaveGlow networks are trained independently in our approach. The Tacotron2 model is specifically trained on our CK speech corpus. For the WaveGlow vocoder, we utilize a pre-trained model based on the LJSpeech corpus, which is then adjusted to synthesize CK speech. Notably, the WaveGlow vocoder has demonstrated effectiveness with unseen languages and speakers (Hsu et al., 2019).

In this research, we have developed and trained three distinct models. Detailed descriptions of these models are provided in the following subsections.

4.2. Model 1: Transfer learning to Kurdish from an english model

Transfer learning, which involves starting with a pre-trained model on a large dataset and fine-tuning it on a smaller, domain-specific

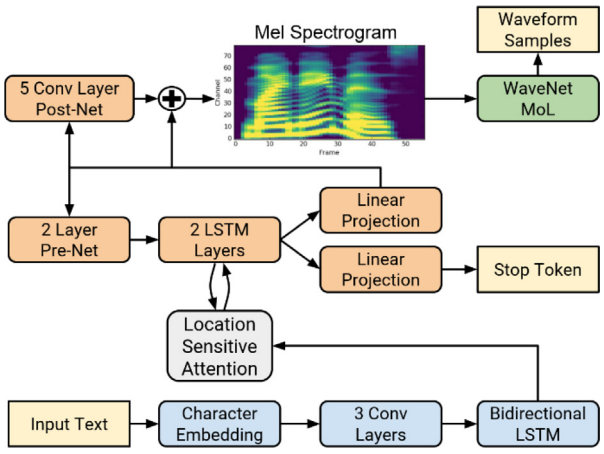


Fig. 3. The Tacotron2 model's block diagram (Shen et al., 2018a).

dataset, offers significant advantages. We implemented our first model by leveraging a pre-trained English model to utilize the learned linguistic features and general speech patterns, to reduce the computational resources needed to achieve high-quality results. Implementing this model involves several steps:

1. **Preparing the Pre-trained Model**: The first step is downloading an English pre-trained model[2] and then uploading it into the model.
2. **Fine-Tuning Dataset Preparation:** The dataset collected for CK TTS is used to train the model where the texts are character-level. Furthermore, the dataset is split into two (train and validation) files, %95 of the data used for the train and %5 of the data used for the validation file, and updated both of them in the *hparams* file. The text preprocessing step encodes the input text into a list of symbols. We define the set of symbols because the pre-trained Tacotron2 model requires a specific set of symbol tables. For instance, we can use "_-!'(),.:;?" and "abcdefghijklmnopqrstuvwxyz". Each character of the input text is mapped to its corresponding symbol's index in the table.
3. **Spectrogram Generation:** The Tacotron2 model generates spectrograms from the encoded text.
4. **Waveform Creation:** Finally, the generated spectrogram is converted into a waveform using the WaveGlow vocoder.

4.3. Model 2: Trained from scratch using full dataset

The second model was trained using Tacotron2 from scratch on our entire speech dataset. For implementation of this model, following steps have been done:

1. **Initialization of Model Weights:** Since we train this model from scratch, the English pre-train model is not downloaded for the training process. We prepare the initial weights randomly and then get the weights during training.
2. **Dataset Preparation:** We use the collected dataset to train the model where the texts are character-level. Furthermore, the dataset is split into two (train and validation) files, %95 of the data used for the train and %5 of the data used for the validation, and updated in the *hparams* file. In the text preprocessing step, we need to define a set of symbols. For example, in the symbol file, we use "'\n! "(),-./:_|«»,-;?'", and "'ابتجحخدرزسشعغفقلمنهوپچژڤکڵۆیە'" . The next step is mapping each character from the input text to its corresponding symbol's index in the table, indicating that the input text is encoded into a set of symbols.
3. **Spectrogram Generation:** A spectrogram is created from the encoded text. For this purpose, we employed the Tacotron2 model.
4. **Waveform Transformation:** The final step is to transform the spectrogram into a waveform. The term "vocoder" also refers to the process of producing speech from a spectrogram. For this purpose, a WaveGlow vocoder was employed.

4.4. Model 3: Trained from scratch using intonationally balanced dataset

One of the limitations of prosody-aware models is their complexity and the extensive amount of data required to accurately capture prosodic features, including sentence intonation, which can be particularly challenging for less-resourced languages like Kurdish. Our Tacotron2-based approach overcomes these limitations by leveraging its end-to-end architecture, which simplifies the process and reduces the need for large datasets. By focusing on character-level text and utilizing a straightforward mapping to symbols, we can efficiently train the model from scratch, achieving high-quality speech synthesis without the intricacies associated with prosody-aware models.

Our dataset initially consisted of 10,979 text-audio pairs, with the majority being declarative sentences. To improve prosody in question and exclamation sentences, we reduced the training speech data from 21 h to 12 h. This adjustment aimed to strike an equilibrium between questions, exclamations, and declarative sentences, ensuring a more balanced learning experience for the model. Following the dataset's rebalancing process, we arrived at 7079 text-audio pairs. Each category – news, sports, poetry, questions, and exclamations – now comprises 1416 sentences. We then trained the Tacotron2-balanced data model from scratch. WaveGlow vocoder was utilized to synthesize the speech sound.

5. Experimental results

5.1. Training setup

We trained a neural network-based TTS system using our CK dataset, which comprises approximately 21 h of female speaker voice data. As detailed in Section 3, this dataset consists of text and audio pairings. The input text is in Kurdish characters, and the audio is sampled at a rate of 16-bit, 22050 Hz. The audio segments vary in length from 1 to 12 s. The models were initially trained entirely on Google Collaboratory Pro+ using a Tesla V100-SXM2-16 GB GPU.

Experiment 1 (Model 1): Using english pre-trained model

In the first experiment, we divided the dataset into training and validation sets. The training set contained 10,529 transcript lines, while the validation set contained 450 lines. Importantly, no training data was included in the test set, which is detailed in Section 3.1. All texts underwent normalization, where numbers and non-Kurdish characters were converted to Kurdish.

After preprocessing, we obtained numerical sequences and mel-spectrograms, stored in NumPy arrays and saved as .npy files. The implementation consists of two phases: a training phase and a synthesis phase, with the latter utilizing a WaveGlow vocoder to generate synthetic speech.

Training was conducted using the following steps:

- **Character Conversion and Transfer Learning:** CK words were converted into English characters, and transfer learning from English models was utilized.

[2] Download link: https://drive.google.com/file/d/1bwL6Bz8Yohs_iCjWCk0JRPruBZUVsXH4/view?usp=sharing.

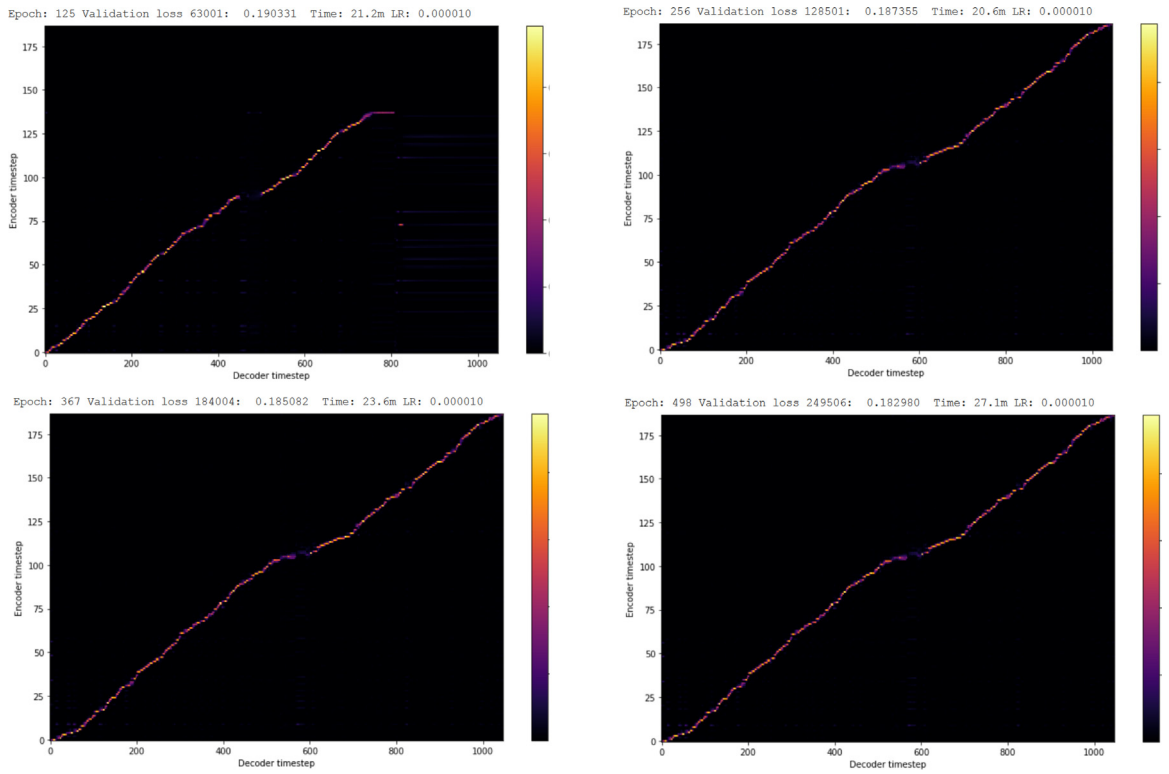


Fig. 4. Examples of alignments at different iterations of the training model for the pre-trained TTS model (Model 1).

Table 6
Final value of the training model hyper-parameters.

Hyper-parameter	Values
Epochs	2000
Batch-size	8
Attention dropout	0.1
Decoder dropout	0.1
Decay start	15,000
Learning rate	1e−5
Weight decay	1e−6

- **Attention Mechanism Training:** The attention mechanism was fully trained using the pre-trained English model, incorporating the learned English character integration.

The audio training samples had a sampling rate of 22050 Hz to match the audio parameters of an open-source implementation trained on the LJS-speech dataset, including hop length and filter length.

The alignment graph is used to assess the precision of the alignment between input characters and output waveform structures. A diagonal alignment map indicates that the model can generate understandable speech, having effectively learned to solve the sequence-to-sequence problem between the input text (encoder stages) and output spectrogram (decoder stages). Monitoring the alignment plots throughout the training process is crucial; if the plots do not appear linear, retraining is necessary. Alignment graphs were taken at various stages of training, with examples shown every 25,000 steps. On average, each epoch of training took approximately 15 min, while generating a waveform took about 2 s. Some alignment graphs from our proposed model are displayed in Fig. 4.

The values of the dropout and learning degree are obtained in a try-and-error manner, and the values of other parameters are taken from previous similar works (Shen et al., 2018a). The training parameters for Model 1 are presented in Table 6.

In the synthesis phase, the pre-trained WaveGlow model is utilized in the vocoder. This model is only used during inference and is robust to variations in gender and language, making its direct application a sound choice. Consequently, this approach reduces both the computational load and the overall training time.

Experiment 2 (Model 2): Train the model from scratch

By adjusting the hyperparameters for our dataset, we trained the Tacotron2 feature prediction model, which is based on a recurrent neural network (RNN). Tacotron2 was trained from scratch on the 21-hour CK dataset using an open-source implementation. We used a batch size of 40 for our training. Choosing the right batch size is crucial, as a smaller batch size led to deteriorated results. Therefore, we opted for a larger batch size to achieve high-quality results and accelerate model convergence.

Training was notably slow due to the use of an RNN. The training process utilized mel spectrogram representations of the raw audio waves. Each experiment involved training the model for 500,000 steps. Throughout the iterative training process, we generated alignment graphs, and by 10,000 steps, the model began to produce acceptable results.

By adjusting the hyperparameters, we trained the Tacotron2 feature prediction model, which is based on a recurrent neural network, using our dataset. This training was conducted from scratch on the 21-hour CK dataset utilizing an open-source implementation.[3]

During the iterative training process, we used alignment graphs to monitor the model's progress, as depicted in the figures. Notably, the experimentation yielded acceptable results at around 10,000 steps (iterations). The alignment graph demonstrates how accurately the input characters align with the output waveform structures. As mentioned, in previous subsection, a diagonal alignment map indicates that the model can produce intelligible speech, having successfully learned to solve the sequence-to-sequence problem between the input text (encoder stages) and output spectrogram (decoder stages). Throughout the training

[3] https://github.com/NVIDIA/tacotron2

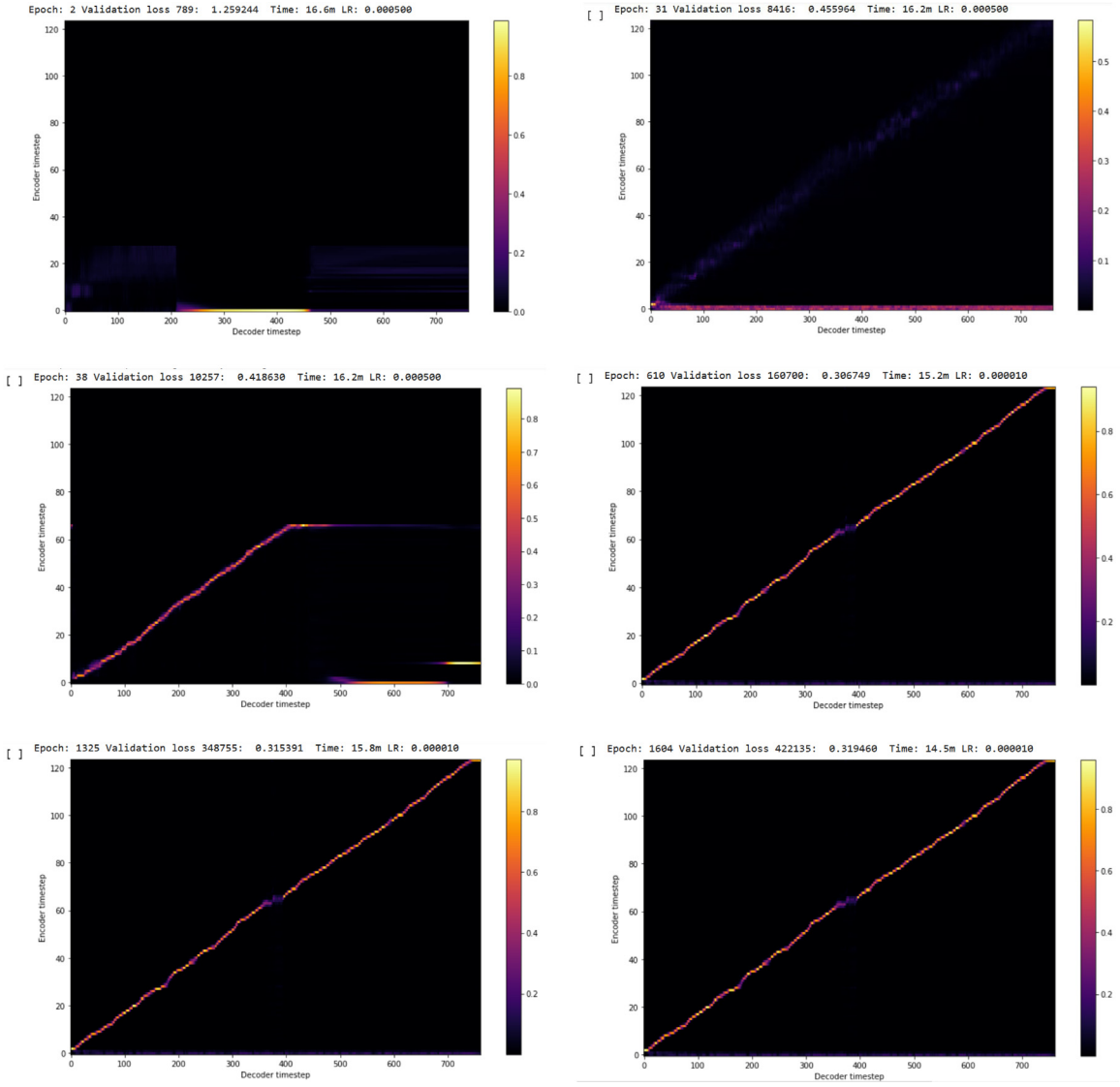


Fig. 5. Some alignments at different epochs in training of experiment 2.

procedure, it is critical to keep an eye on the alignment plots; if they do not appear to be linear, the training should be redone. Some alignment graphs derived from our model are shown in Fig. 5.

According to the open-source implementation, each decoding step generates one frame. The values of the adopted and learning rate are obtained in a tray an error manner, while the values of other parameters are occupied from previous similar works (Shen et al., 2018a). WaveGlow was utilized in this experiment to transform the predicted acoustic features into waveforms of the audio. The subjective rating of this experiment was positive. On average, each training epoch lasted approximately 20 min, while generating a waveform typically took around 2 s. The training hyperparameters are presented in Table 7.

Table 7
Final values of the training model hyperparameters for model 2 (trained from scratch).

Hyperparameter	Value
Epochs	2000
Batch-size	40
Attention dropout	0.4
Decoder dropout	0.1
Decay start	15 000
Learning rate	1e−5
Weight decay	1e−6

Experiment 3 (Model 3): Training the model from scratch using balanced data

To enhance the prosody of the generated speech signal, particularly for exclamatory and interrogative sentences, we balanced our dataset by equalizing the number of sentences across these categories and other types (e.g., news sentences). Initially containing 10,979 (text, audio) pairs, we reduced the dataset to 7079 sentences, resulting in a total of 12 h of data. This dataset was then used to train the Tacotron2 model from scratch. The dataset reduction was aimed at achieving a balance among all sentence types to effectively capture the intonation of question and exclamation sentences. We employed the open-source TensorFlow version of Tacotron2 from NVIDIA/Tacotron2 for this purpose, utilizing the WaveGlow vocoder for waveform synthesis instead of WaveNet.

The implementation comprised two phases: in the training phase of the feature prediction network, we used a batch size of 40, an attention dropout rate of 0.4, and trained the data over 2000 epochs. Throughout the training process, alignment graphs were generated to aid model refinement. Notably, acceptable outcomes were observed after 6k steps. The attention alignment during training is illustrated in Fig. 6.

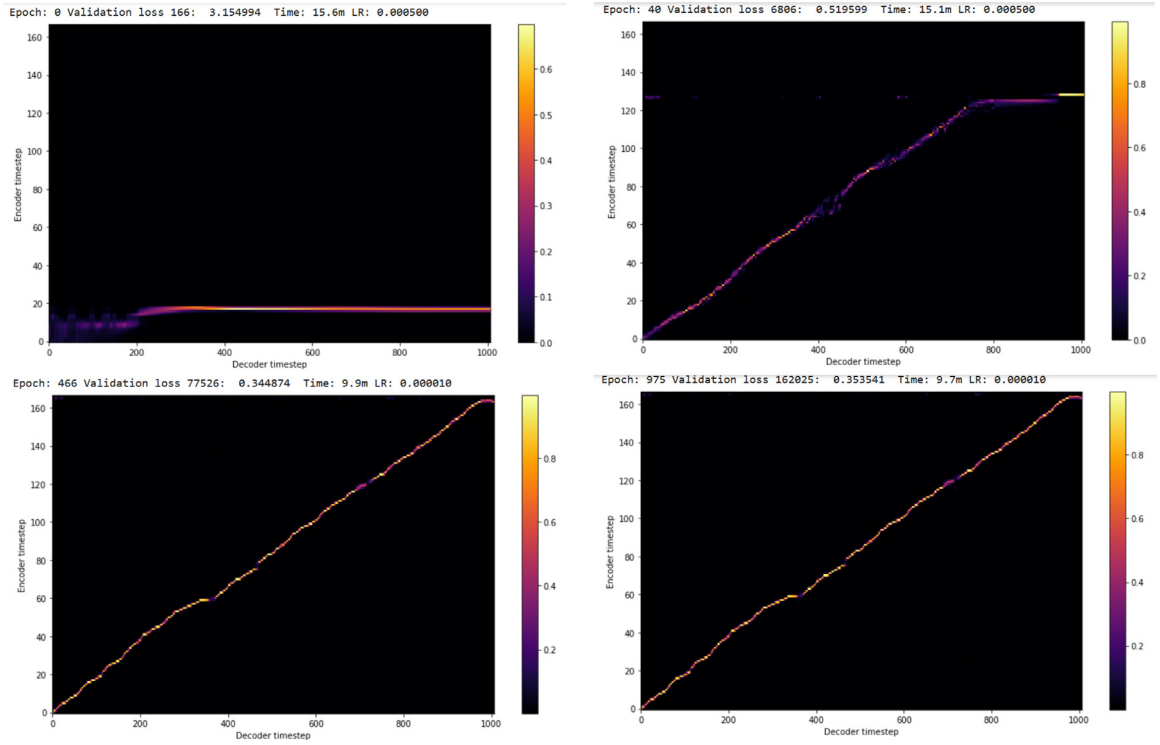


Fig. 6. Attention alignments during training for the third experiment.

In the synthesizing phase, the WaveGlow vocoder with default hyperparameters was used to synthesize audio waveforms.

Optimizing the learning rate, a crucial hyperparameter in neural networks, presents certain challenges. It dictates the magnitude of adjustments during training, with lower rates often necessary for accuracy enhancement. However, this necessitates longer training times and multiple iterations to prevent overfitting, particularly in datasets with limited samples. Various tools are available to aid in this process. In our case, we conducted extensive tests on the dataset, varying the learning rate. Additionally, we referenced relevant studies, such as those in Google TTS, for initial setup and testing. These studies utilized low learning rates, like 0.001 and 0.0001. Interestingly, in our experiments, a learning rate of 0.00001 yielded the best results across all three experiments. We balanced training speed and stability by setting a batch size of 8 for the first experiment and 40 for the second and third experiments. Furthermore, we maintained a sequence length typically ranging between 800 and 1200 frames, balancing memory usage and context capture. Sequences were either padded or truncated to ensure a fixed length. These hyperparameters are pivotal in optimizing training efficiency, ensuring efficient memory utilization, steady convergence, and high-quality voice synthesis.

5.2. Evaluations

We conducted a subjective assessment to measure the naturalness and comprehensibility of our trained models. This evaluation utilized the Mean Opinion Score (MOS) as the metric. MOS involves a group of evaluators, ideally native speakers or language experts, who rate various synthetic speech samples on a scale of 1 (very bad) to 5 (very good). These assessments are conducted in a controlled environment, with evaluators listening to recordings and assigning scores. The mean score is then calculated by summing all sample scores and dividing by the total number of samples. This process yields an overall MOS, providing a single numerical measure of speech quality. We recruited 12 native speakers, comprising 7 males and 5 females, with ages ranging from 21 to 46, to assist in the MOS assessments. The evaluation involved 110 sentences from the test set (outlined in Section 3.1), which were distinct from those used in training.

Table 8
The MOS results of the proposed models.

Models	Result (MOS)
Genuine Voice	4.99
Model 1 (using English pre-trained)	4.10
Model 2 (scratch)	**4.78**
Model 3 (balanced data)	4.57

Each participant was asked to listen and rate four distinct sets of sounds: the genuine voices from the test set, along with three synthesized voice sets generated by models trained over 2000 epochs (500,000 iterations). They heard each set through headphones and subsequently rated them on a scale of five points: 5 for excellent, 4 for good, 3 for neutral, 2 for poor, and 1 for very poor. Table 8 shows the overall MOS results of the evaluations.

Fig. 7 Illustrates a comparative analysis of the MOS results of four evaluations across distinct categories of the sentences.

Although the results we obtained from the first model trained using an English pre-trained model were outstanding in terms of quality, there were issues regarding the understanding and pronunciation of some words and letters. Table 9 presents several cases where the model produces incorrect sounds in the first experiment.

We conducted experiment 3 to see if our model would better read the intonation of questions and exclamation sentences compared to previous experiments. The MOS result of this experiment (MOS = 4.57) is lower than the MOS result of the second experiment (MOS = 4.78). However, the primary purpose of the third experiment was to increase the (MOS) score of the second experiment, especially the intonation of questions and exclamation sentences. The results indicated that there were no significant differences between the second and third experiments with respect to the naturalness, intelligibility, and intonation of questions and exclamation sentences, as shown in Table 10.

The third experiment did not show an improvement in MOS due to the limited amount of training data and the presence of some

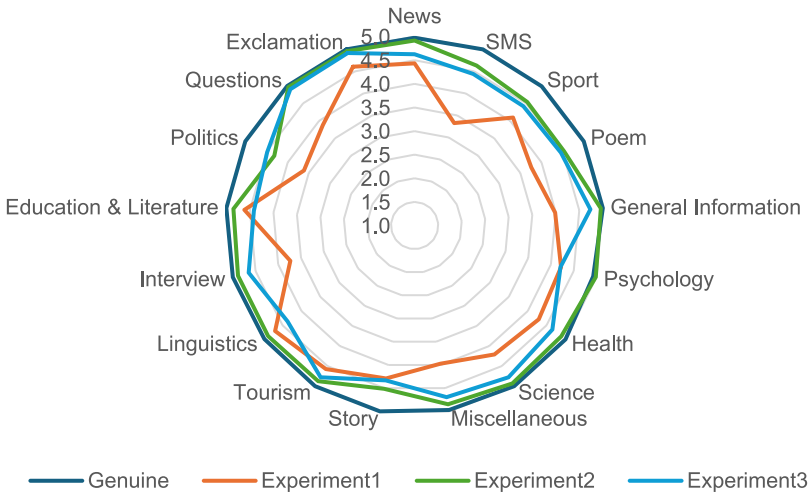


Fig. 7. Results of the genuine and the proposed models (average of MOS) about all different categories.

Table 9
Some instances where the model generates incorrect sounds in the first experiment.

Test set (sentences)	Word	Generated speech by the model
کاتێک کە مرد	مرد	مردا
بەشێک لە خەڵکی بەریتانیا حەز دەکەن بۆریس جۆنسن، واز لە پۆستەکەی بهێنێت.	حەز	هەز
ئایا دەتوانی بە عەرەبی گۆرانی بڵێی؟	گۆرانی بڵێی	گۆرانی بلێی

Table 10
The average MOS for questions and exclamation sentences in two models.

Sentence Type	Model 2: trained on 10,979 sentences	Model 3: trained on 7079 sentences (balanced data)
Question sentences	4.96	4.90
Exclamation sentences	4.95	4.91

Table 11
Comparison of MOS results between our model and the model presented in Muhamad and Veisi (2022) on identical test sets.

Method	Result (MOS)
The model presented in Muhamad and Veisi (2022)	4.10
Our model trained from scratch using the full dataset	4.78

noisy data, which included minor issues with the intonation of the recorded sentences. Consequently, the model reads some question and exclamation sentences as declarative sentences.

In summary, after evaluating all three experiments, we found that the second experiment produced the best model for CK text-to-speech in terms of naturalness and intelligibility.

5.3. Comparison results

Table 11 compares the MOS result of our best model with the previous Kurdish TTS work [42] trained on a 10-hours CK dataset and using the English pre-trained model on identical test sets. The MOS results show that our model trained from scratch on the 21-hours CK dataset, significantly outperforms in terms of naturalness and intelligibility.

6. Conclusion and future works

In this research, we assembled a comprehensive 21-hour Central Kurdish (CK) speech corpus and developed a CK text-to-speech system using Tacotron2 architecture. Tacotron2 utilizes a recurrent sequence-to-sequence feature prediction network with attention to predict mel spectrogram frames from input character sequences.

Our findings demonstrate that our Tacotron2-based system achieves both satisfactory intelligibility and naturalness in synthesized speech output. Notably, our second experiment, the model trained from scratch on the CK dataset, surpassed the results of experiment 1 (Tacotron2 model trained via pre-trained English model) and experiment 3 (same architecture as experiment 2 but using balanced dataset) in terms of both naturalness and intelligibility, as per subjective evaluations.

Based on the insights from our assessment, the Tacotron2, trained from scratch and based on a recurrent neural network model, is suitable for practical text-to-speech applications. The proposed system can be used for audiobooks, recommendation systems, phone inquiry service, and smart education.

The data used to train the models were collected in the central dialect of Kurdish. Future work will include training the model in other dialects of the Kurdish language, such as Northern, Southern, and Hawrami.

Adding sentiment analysis to future Kurdish text-to-speech systems enhances contextual awareness. By analyzing text emotions, the system adjusts intonation and emphasis, improving the synthesized speech's naturalness and contextual appropriateness.

We will consider using transformers instead of RNN-based Tactron2 to train our dataset.

CRediT authorship contribution statement

Sabat Salih Muhamad: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Hadi Veisi:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Conceptualization. **Aso Mahmudi:** Writing – review & editing, Writing – original draft, Software, Methodology. **Abdulhady Abas Abdullah:** Writing – review & editing, Software, Methodology, Formal analysis. **Farhad Rahimi:** Writing – review & editing, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Arik, S.O., et al., 2017. Deep voice 2: Multi-speaker neural text-to-speech. In: Advances in Neural Information Processing Systems.

Bahrampour, A., Barkhoda, W., Azami, B.Z., 2009. Implementation of three text to speech systems for Kurdish language. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). http://dx.doi.org/10.1007/978-3-642-10268-4_38.

Barkhoda, W., ZahirAzami, B., Bahrampour, A., Shahryari, O.K., 2009. A comparison between allophone, syllable, and diphone based TTS systems for Kurdish language. In: IEEE International Symposium on Signal Processing and Information Technology. ISSPIT 2009, http://dx.doi.org/10.1109/ISSPIT.2009.5407540.

Black, A.W., Taylor, P., 1997. Automatically clustering similar units for unit selection speech synthesis. Int. Speech Commun. Assoc..

Daneshfar, F., Barkhoda, W., Azami, B.Z., 2009. Implementation of a text-to-speech system for Kurdish language. In: Proceedings - 2009 4th International Conference on Digital Telecommunications. ICDT 2009, http://dx.doi.org/10.1109/ICDT.2009.29.

Donahue, J., Dieleman, S., Bińkowski, M., et al., 2022. End-to-end adversarial text-to-speech. arxiv.org, [Online]. Available: https://arxiv.org/abs/2006.03575. (Accessed: 22 October 2022).

Fahmy, F.K., Khalil, M.I., Abbas, H.M., 2020. A transfer learning end-to-end arabic text-to-speech (tts) deep architecture. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). http://dx.doi.org/10.1007/978-3-030-58309-5_22.

Hassani, H., et al., 2011. Kurdish text to speech (KTTS), researchgate.net. [Online]. Available: https://www.researchgate.net/profile/Hossein-Hassani-2/publication/295092948_Kurdish_Text_to_Speech_KTTS/links/59c78058458515548f37944d/Kurdish-Text-to-Speech-KTTS.pdf. (Accessed: 01 March 2022).

Hayashi, T., et al., 2020. Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. http://dx.doi.org/10.1109/ICASSP40776.2020.9053512.

He, T., Zhao, W., Xu, L., 2020. DOP-tacotron: A fast Chinese TTS system with local-based attention. In: Proceedings of the 32nd Chinese Control and Decision Conference. CCDC 2020, pp. 4345–4350. http://dx.doi.org/10.1109/CCDC49329.2020.9164203.

Hsu, P., Wang, C., Liu, A.T., Lee, H., 2019. Towards robust neural vocoding for speech generation: A survey. [Online]. Available: http://arxiv.org/abs/1912.02461.

Hunt, A.J., Black, A.W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. pp. 373–376. http://dx.doi.org/10.1109/icassp.1996.541110.

Kayte, S., Waghmare, K., Gawali, B., 2015. Marathi speech synthesis: A review. Int. J. Recent Innov. Trends Comput. Commun. 3 (6), 3708–3711.

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M., 2019. Neural speech synthesis with transformer network. In: 33rd AAAI Conference on Artificial Intelligence, AAAI 2019 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence. EAAI 2019, http://dx.doi.org/10.1609/aaai.v33i01.33016706.

Li, Y., Qin, D.H., Zhang, J.B., 2021. Speech synthesis method based on Tacotron2. In: 2021 13th International Conference on Advanced Computational Intelligence. ICACI 2021, http://dx.doi.org/10.1109/ICACI52617.2021.9435882.

Liu, R., Sisman, B., Bao, F., Gao, G., Li, H., 2020a. Modeling prosodic phrasing with multi-task learning in tacotron-based TTS. IEEE Signal Process. Lett. 27, 1470–1474. http://dx.doi.org/10.1109/LSP.2020.3016564.

Liu, R., Sisman, B., Li, J., Bao, F., Gao, G., Li, H., 2020b. Teacher-student training for robust tacotron-based TTS. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. pp. 6274–6278. http://dx.doi.org/10.1109/ICASSP40776.2020.9054681.

Lu, Y., Dong, M., Chen, Y., 2019. Implementing prosodic phrasing in Chinese end-to-end speech synthesis. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. pp. 7050–7054. http://dx.doi.org/10.1109/ICASSP.2019.8682368.

Muhamad, S., Veisi, H., 2022. End-to-end Kurdish speech synthesis based on transfer learning. Passer J. Basic Appl. Sci. 4 (2), 150–160. http://dx.doi.org/10.24271/PSR.2022.351832.1149.

Mundada, M.R., Gawali, B., Kayte, S., 2014. Recognition and classification of speech and its related fluency disorders. Int. J. Comput. Sci. Inf. Technol. 5 (5), 6764–6767.

Naderi, N., Nasersharif, B., Nikoofard, A., 2022. Persian speech synthesis using enhanced tacotron based on multi-resolution convolution layers and a convex optimization method. Multimedia Tools Appl. 81 (3), http://dx.doi.org/10.1007/s11042-021-11719-w.

Ning, Y., He, S., Wu, Z., Xing, C., Zhang, L.J., 2019. Review of deep learning based speech synthesis. Appl. Sci. (Switzerland) 9 (19), 4050. http://dx.doi.org/10.3390/app9194050.

van den Oord, A., et al., 2016. WaveNet: A generative model for raw audio based on PixelCNN architecture. arXiv.

Ping, W., Peng, K., Chen, J., 2019. Clarinet: Parallel wave generation in end-to-end text-to-speech. In: 7th International Conference on Learning Representations. ICLR 2019.

Ping, W., et al., 2018. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In: 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings.

Prenger, R., Valle, R., Catanzaro, B., 2019. Waveglow: A flow-based generative network for speech synthesis. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. http://dx.doi.org/10.1109/ICASSP.2019.8683143.

Ren, Y., et al., 2019. FastSpeech: Fast, robust and controllable text to speech. In: Advances in Neural Information Processing Systems.

Ren, Y., et al., 2022. Fastspeech 2: Fast and high-quality end-to-end text to speech. arxiv.org, [Online]. Available: https://arxiv.org/abs/2006.04558. (Accessed: 22 October 2022).

Schoeffler, M., et al., 2018. webMUSHRA - A comprehensive framework for web-based listening tests. J. Open Res. Softw. 6 (1), http://dx.doi.org/10.5334/jors.187.

Sejnowski, T.J., Rosenberg, C.R., 1987. Parallel networks that learn to pronounce English text. Complex Systems 1.

Shen, J., et al., 2018a. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. http://dx.doi.org/10.1109/ICASSP.2018.8461368.

Shen, J., et al., 2018b. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. pp. 4779–4783. http://dx.doi.org/10.1109/ICASSP.2018.8461368.

Shifera, B., 2021. 2021 Adama. Adama Sci. Technol. Univ. Sept., Ethiopia, no. September.

Sotelo, J., et al., 2019. Char2Wav: End-to-end speech synthesis. In: 5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems.

Tan, X., et al., 2022. NaturalSpeech: End-to-end text to speech synthesis with human-level quality. [Online]. Available: http://arxiv.org/abs/2205.04421.

Thackston, W.M., 2006. Sorani Kurdish—a Reference Grammar with Selected Readings. Harvard Univ.

Vainer, J., Dušek, O., 2020. SpeedySpeech: Efficient neural speech synthesis. In: Proceedings of the Annual Conference of the International Speech Communication Association. INTERSPEECH, http://dx.doi.org/10.21437/Interspeech.2020-2867.

Veisi, H., Hosseini, H., MohammadAmini, M., Fathy, W., Mahmudi, A., 2022. Jira: a central Kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon. Lang. Resour. Eval. http://dx.doi.org/10.1007/S10579-022-09594-4.

Wang, Y., et al., 2017. Tacotron: Towards end-to-end speech synthesis. In: Proceedings of the Annual Conference of the International Speech Communication Association. INTERSPEECH, http://dx.doi.org/10.21437/Interspeech.2017-1452.

Weiss, R.J., Skerry-Ryan, R., Battenberg, E., Mariooryad, S., Kingma, D.P., 2021. Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis. pp. 5679–5683. http://dx.doi.org/10.1109/icassp39728.2021.9413851.

Win, Y., Masada, T., 2020. Myanmar text-to-speech system based on tacotron-2. In: International Conference on ICT Convergence. http://dx.doi.org/10.1109/ICTC49870.2020.9289599.

Ze, H., Senior, A., Schuster, M., 2013. Statistical parametric speech synthesis using deep neural networks. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. pp. 7962–7966. http://dx.doi.org/10.1109/ICASSP.2013.6639215.

Zen, H., Tokudaa, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Commun. 51 (11), 1039–1064. http://dx.doi.org/10.1016/j.specom.2009.04.004.

Zhang, C., Zhang, S., Zhong, H., 2019. A prosodic mandarin text-to-speech system based on tacotron. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. APSIPA ASC 2019, pp. 165–169. http://dx.doi.org/10.1109/APSIPAASC47483.2019.9023283.