Journal of Computer Science

Original Research Paper

Research on Deep Neural Network for Afaan-Oromo Language Text-to-Speech Synthesis

[1]Diriba Gichile Rundasa, [2]Arulmurugan Ramu, [3]Teshale Debushe Adugna, [1]Chala Sembeta Teshome and [3]Desalegn Tasew

[1]Department of Computer Science Information Technology, Mattu University, Ethiopia
[2]Department of Computational Sciences and Software Engineering, K. Zhubanov University, Kazakhstan
[3]Department of Information Technology, Mattu University, Ethiopia

Article history
Received: 04-11-2024
Revised: 05-12-2024
Accepted: 29-01-2025

Corresponding Author:
Diriba Gichile Rundasa
Department of Computer Science Information Technology, Mattu University, Ethiopia
Email: diriba.gichile@meu.edu.et

Abstract: Text-to-speech synthesis is the automatic translation of unlimited natural language sentences from Text to spoken form that closely mimics the spoken form of the same Text by a native speaker of the language. The purpose of a Texttext-to-speech synthesizer is to generate comprehensible, natural signalling human voice from text transcriptions. Despite the wide range of potential applications for Text-to-speech systems, the field is language-dependent, with most efforts concentrated on accessible languages, especially English. The linguistic resources required to make a speech from texts are lacking for under-resourced languages like the Afaan-Oromo language. To develop an Afaan Oromo language text-to-speech synthesizer, a speech dataset was prepared, which is 10644 text and audio pairs in numbers and assembled from dependable sources. After that, the proposed model is developed, which incorporates nonstandard terminology, including acronyms, currencies and numerals, in addition to common terms and names. The deep neural network was selected for this study because it has a good ability to convert Text into complex spoken Text. A number of experiments were carried out to find the best-performing model. To assess the performance of the model objectively, the attention mistake is used where, whereas to assess the models' performance subjectively, the Mean Opinion Score or scale (MOS) test is used. Subsequently, the objective outcomes evaluation revealed that Deep Voice (DV) 3 produced 18 of the 248 words in the evaluation sentence set. At the same time, Tacotron-2(two) made attention errors, which are two in number. Moreover, MOS scores for naturalness and intelligibility have made 4.36 and 4.33 out of five (5) for Tacotron-2 (two), respectively and 3.32 and 3.04 for Deep Voice(DV) 3, respectively. Because it can translate intricate verbal information into auditory feature parameters, the deep neural network was selected for this research. Therefore, the Tacotron-2 (two) model yielded good results and promising results compared with Deep Voice (DV) 3, making it suitable for a range of applications, such as smart education, different telephone inquiry services, and recommendation systems, which are the most common areas of the system.

Keywords: Deep Neural Network, Speech Processing, Mean Opinion Score, Afaan-Oromo, Text to Speech, Tacotron 2 (Two), Deep Voice (DV) 3

Introduction

Text-to-speech synthesis is the automatic translation of unlimited natural language sentences from Text to spoken and other human voice characteristics to produce a completely "synthetic" human voice output corresponding to input text. There are two primary stages to the Text-to-speech synthesizer system. There are several steps or processes for converting the input text into its phonemes. The first and most essential step is analyzing Text, which involves mapping the input text into its phonemes. After text analysis was done, the

second step followed what we call speech analysis, which uses prosodic and phonetic information to create speech waveforms. Digital Signal Processing (DSP) is another name for this procedure. For many years, the state-of-the-art in the field of Text-to-speech synthesis was concatenative speech synthesis, which directly concatenates the units (Phonemes) in a vast database to produce a continuous speech stream (Hunt and Black, 2002). Concatenated units can be words, morphemes, phrases, sentences, syllables, diaphones, half-phones, or other phonetic forms. Despite offering excellent voice output, it frequently has audible (Glitches) in the output because of the elements' less content concatenation. This typically costs money and requires more storage to contain every utterance in the Afaan Oromo language. In synthesizing Text-to-speech, statistical parametric speech synthesis that makes use of deep neural network architectures has proven to be more effective than earlier methods. The Pixel CNN serves as the foundation for WaveNet, a neural-based audio generation model (Oord *et al*., 2020). In terms of naturalness, it can produce audio that is extremely similar to a human voice. Modern vocoders like Wave-Net are not completely functional in end-to-end systems because they need linguistic information from another text-to-speech system component that uses the auto-regressive nature of the architecture to predict features and generate speech progressively. With the advent of end-to-end architectures like Tacotron, a lot of the time-consuming tasks involved in speech synthesis, like feature engineering and human annotation, have been reduced (apart from the compilation of text and audio pairs for training (Wang *et al*., 2017). Tacotron is an integrated end-to-end generative Text-to-speech model that generates voice from Text using a sequence-to-sequence model with an attention mechanism (Shen *et al*., 2018). Because it synthesizes talks frame by frame, it is quicker than the autoregressive technique, which involves sample-level synthesis. Tacotron includes a post-processing network, a decoder, an encoder and content-based attention (Bengio *et al*., 2015). Tacotron surpasses a working parametric system in terms of naturalness, achieving a subjective mean opinion score of 3.86 in the United States (US) English. So, it achieves better in terms of naturalness than a working parametric system beyond the naturalness of a working parametric system. Elements used by Tacotron, including the method for the reconstruction of the Griffin-Lim signal. Tacotron-2 (two) offers cutting-edge sound quality that is comparable to natural human speech by utilizing hybrid attention, which combines location-based and content-based attention (Arık *et al*., 2017). The Mean Opinion Score or scale (MOS) of 4.56 for the Tacotron-2 (two) is fairly comparable to an MOS of 4.62 for speech that has been professionally recorded. A full convolutional attention-based sequence-to-sequence model called Deep Voice (DV) 3 (Ping *et al*., 2018) can translate textual cues, including phonemes, stresses and characters, into vocoder parameters. The audio signals are produced by the audio waveform synthesis model using the expected vocoder parameters as an input (Sutskever *et al*., 2014). Three main parts comprise the Deep Voice (DV) 3 architecture: The final vocoder parameters are predicted by a fully convolutional post-processing network called Converter. A decoder is a fully convolutional system that decodes the learnt representations in an autoregressive or based immediately preceding value fashion. An encoder is a fully convolutional network that transforms textual features into an internal feature representation. The results of Tacotron-2 (two) obtained indicate good performance, but compared with the Deep Voice (DV) 3, the Deep Voice (DV) 3 trains more quickly.

The fundamental contributions of this study are the following:

- The Afaan Oromo language text analysis front was created in order to solve the text analysis problem
- Creation of a deep neural network-based Speech synthesis model for the Afaan Oromo language
- For this study, 10644 text and audio pairings of a single male speaker make up the phonetically rich and balanced Afaan Oromo language speech dataset that must be gathered and prepared for Text-to-speech challenges.

Related Works

The author developed the first Afaan Oromo language TTS system by using the concatenative speech synthesis technique and diphones as the basic concatenation units to synthesize sample Afaan Oromo language words (Morka, 2001). The Centre for Spoken Language Understanding's and the Hidden Markov Toolkit's inaccessibility and consideration for speech synthesis were noted in the study. According to the research article, native speakers were 43.33% successful in identifying the transcribed phonetic unit utterance, whereas listeners who heard the speech in many ways, which can expressed at least three times in an unrelated fashion, were 83.33% successful. In order to reduce the inaccuracies that result from segmentation, the author ultimately suggested using smoothening or making horizontal techniques such as linear-predictive-coding to horizontal the diphones' transition- points.

The second author has developed a concatenative-based Text to Speech synthesizer for Afaan Oromo language (Samson, 2011). According to the study, 75 and 54% of the words in the data set, respectively, are pronounced correctly using the diphone and triphone speech units.

Despite the fact that the number of rule-based diphone database entries employed in the study was too small. The system's performance suffered when huge units were

concatenated. For the diphone and triphones, the author obtained intelligibility rates of 3.03 and 2.2, respectively and for each speech unit, the system's naturalness was 2.65 and 2.02, respectively. However, the system performs poorly overall as a result of using a rule-based approach. To incorporate every conceivable speech, the concatenative method requires a larger database.

The Third study for the Athe faan Oromo language has developed an HMM-based statistical parametric speech synthesis model faan Oromo language (Wosho, 2021). According to the study, the author acquired scores of 4.1 and 4.3 out of 5 for intelligibility and naturalness, respectively. This study does not take into account nonstandard terms like acronyms, punctuation, numerals, currencies, or abbreviated words. It is advised to use deep neural networks or deep learning techniques.

The most recent study acoustic for the Athe faan Oromo language has developed the eep literacy model grounded on BLSTM RNN for the Athe faan Oromo language (Tamrat and Muluken, 2022). The RNN model is sed from a given input point sequence to the displaced period and aural model. The RNN-based BLSTM implementation uses the Pytorch library, which was modified on the Jupyter Notebook to create speech samples and duration from the trained acoustic model. By listening to the audio that was recorded during the test, the native speech evaluator used the Mean Opinion Score or scale (MOS) evaluation technique to evaluate subjectively the synthesized speech performance; the subjective test, which consists of thirteen sentences, yielded ratings for naturalness and intelligibility of 3.76 and 3.77 out of 5, respectively.

As previously mentioned, efforts were made to create an Afaan Oromo speech synthesizer. Their initiatives, nevertheless, are predicated on traditional techniques. Despite the fact that the HMM technique solves the huge database need, it necessitates a thorough feature engineering procedure and in-depth subject knowledge.

In addition, in the HMM technique linguistic features need to first be mapped into probability densities, for this reason, it was difficult to incorporate the complex linguistic features didn’t consider NSWs such as abbreviated words, numbers, currency, acronym and punctuation marks and used a very small speech dataset (Wosho, 2021).

Materials and Methods

Data Collection and Preparation

A speech corpus is a collection of Text at the sentence level, along with studio recordings that match the Text in that format. There isn't a substantial, open-sourced voice synthesis corpus in the Afaan Oromo language due to the expense of annotation. The research does not focus on the potential issues with dialectal variations in Afaan Oromo because even different dialects are available, namely Western (including Wallagga), Eastern (Harar), Southern (Boorana) and Central (Shawa). They have the same standard book and pronunciation, which is used in every region for teaching and learning, and for reading text for different purposes like text-to-speech synthesis. Therefore, no more focus on dialect variation. For this reason, the dataset has been collected, and the dataset prepared by this study can be used for similar research works found on: https://data.mendeley.com/datasets/mpy85ns82z.

Text Data Collection

For this study, the dataset has been collected from various sources. Because of its wide range of sources and size, the corpus gathered for this study is regarded as phonetically rich and balanced. The corpus can integrate a wide range of domains by using a diversity of language structures of sentences, kinds and durations, which improves the text-to-speech system's quality. As a result, the following three main sources were used to create the dataset. These sources are News Media sources like BBC Afaan-Oromo, Oromia Broadcasting OBN, Fana Broadcasting Corporate and Ethiopian News Agency, which are the news organizations that served as text sources for this study. Also, the dataset was collected from non-fiction Books because the sources are written by experts and are available in both electronic and manual form. A total of 10644 sentences are collected in the text dataset, which includes two non-fiction books, "Ida'amu" and "Coqorsa Abdii". The last source of Text used for this study has been collected from the Holy Bible, which is written in the Afaan Oromo language.

Process of Audio Recording

Since the proposed model makes use of the pair dataset, the text corpus needs to be converted into the proper audio format. Speech synthesis is harder with child and female voices because female voices feature a pitch of almost twice and three times as high as male and children voices respectively (Lemmetty, 1999). Because of this, we selected a qualified male speaker, and a professional studio with a background noise cancelling microphone was used for the audio recording from Oromia Institute of Art was 28 years old was supposed to read the words and so that it would eliminate most of the noise in the background, it may not Limiting diversity in phonemes and prosody representation. After two weeks of audio recording, 10644 recordings totalling 18 (h) and 30 min of speech data were collected and Wav files were created. The non-fiction works "Ida'amu" and "Coqorsa Abdii" are used as text sources under this titles. Figure (1) shows the summary of dataset created for the Afaan Oromo.

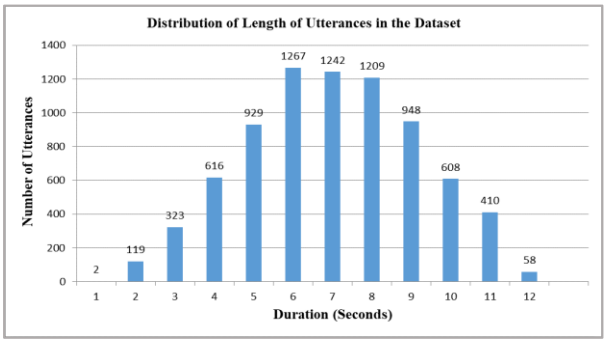

Fig. 1: Summary the of dataset created for the afaan-oromo

Text Pre Processing

Before the dataset is input into the deep neural network models, a variety of preprocessing techniques are applied to improve the quality of the Text and audio recordings. Many other text corpora into speech systems use similar text processing techniques to the ones employed in this study. The text data in this study is subjected to the following text processing techniques.

Text Cleaning

Before beginning the audio recording procedure, the gathered Text must be sanitized. The following are the primary cleaning tasks carried out on the text data:

- All kinds of typo errors, unnecessary white spaces and repeated words are fixed
- All characters are converted to lowercase
- Period (.) has been added at the end of the sentences that do not have end of sentence markers such as
- All characters other than a-z, ␣ and “’ -, ()!? are stripped
- Very long sentences were segmented

Text Normalization

The process of creating a normalized expression or words from nonstandard words or expression is known as text normalization., It goes without saying that there is nonstandard vocabulary in Afaan Oromo texts that cannot be directly translated into phonemes. The following texts have been subjected to the normalization procedures in this study: A complete textual representation that matched the recordings was used to substitute sentences that contained numbers, years and currencies. Acronyms and abbreviations are listed in writing.

Text Segmentation

After normalization breaking written Text into meaningful units, such as words, sentences, or topics are important. Because sentences have become longer after normalization, very long sentences that require more than Thirty words to be spoken and short sentences that contains less than four words were eliminated because they may hinder the effectiveness of the training process.

Audio Pre Processing

The next audio processing steps were practical to the audio signals before to the extraction of the acoustic feature and at beginning of the training model.

Normalization: Speaker's loudness may vary from word to word during the audio recording process. Because of this it is difficult to handle the dataset or the file availabel. To solve the above-mentioned issue, audio files' volume should be adjusted as it’s necessities.

Preemphasis: The audio signals it should be pre emphasized before beginning the extraction of acoustic characteristics. The audio signal's high-frequency component-were amplified in order to accomplish the process.

Silence trimming. At the beginning and the end audio signal has silence. The audio samples in dataset have had their opening and closing silences removed. It was demonstrated that Cutting has been shown to facilitate the alignment of written utterances with audio samples, because it reduces the amount of time needed for training.

Results

To choose the model that performs the best for this study, comparison of Deep Voice (DV) 3 and Tacotron-2 (two), Text to-Speech synthesizer is conducted by using the open-source model implementations.

Tacotron-2 (Two) Implementation

Tacotron 2 (two) was implemented using Rayhane M^3 open-source Tensor flow framework. To address the time and computing capabilities, the Wave Net portion has been stripped with a minor adjustment to the default hyperparameters. The implementation is separated into two phases. The first phase is training phase which involves training the feature prediction network and the second phase is synthesis phase which involves feeding the model new Text and using the learnt model with the Griffin-Lim method to produce a synthesized voice.

Training the Feature of the Model

To train the model a single GPU was used to train prediction model including the feature for 110K steps with a batch size of 16 over the course of sixteen days of computing. The linear and Mel representations of audio signal features are used to train the model. Figure (2) shows how attention changes during the training period

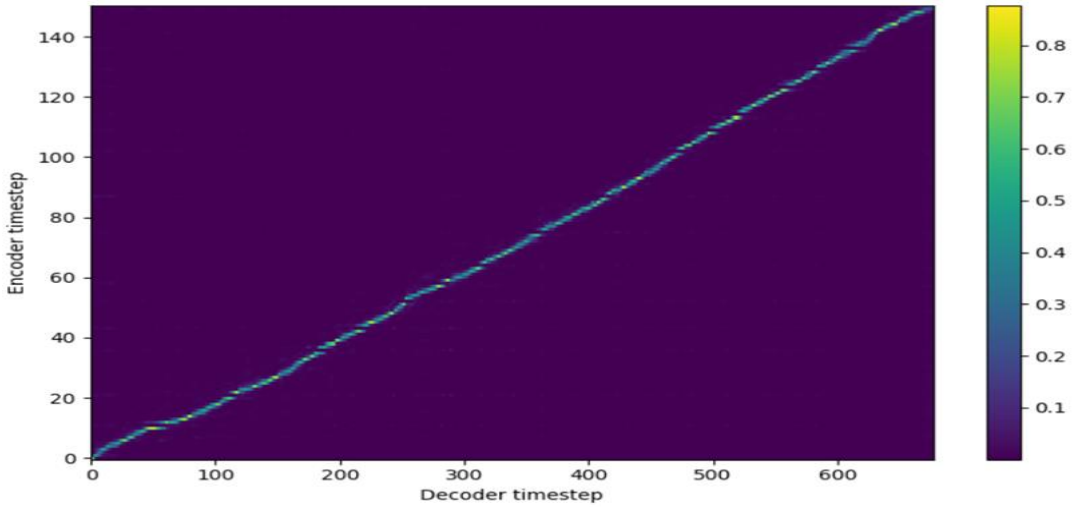

Fig. 2: This figure summarizes how the attention changes during the training

As with the open-source implementation, the amount of frames to be generated at each decoding stage is set to one. The ground truth spectrogram's preceding frame, rather than the predicted one, is fed into the pre-net when the model is in Ground-Truth-Aligned mode during training. By matching the prosody and pitch of the generated-spectrogram with the ground truth, GTA enables faster convergence and a shared framework between the prediction and the ground-truth. Without GTA, given a fixed text input, the synthesizer would produce several versions of the same phrase. In this experiment, there were challenging to offer a measurable evaluation which mean quantitative of the model's performance. The alignments produced by the attention module were taken into consideration as an alternative method of evaluating the outcomes.

Figure (3) illustrates, comparing the Ground Truth Aligned expected spectro-gram with the ground-truth spectro-gram (Right) and aligning the encoder and decoder steps (Left).

On the dataset's validation section, the model was assessed every 2500 steps. Figure (4) shows the evolution of the anticipated attention alignments. During synthesis phase, the alignments were expected for varying numbers of preparation steps.

The assessment metrics that explain on Fig. (5) displays the assessment or evaluation metrics that account for the training models' performance. The loss function exhibits the expected behaviour; after 110 K steps proceed, it converges to a value of 0.887 after decreasing steadily. After 110K steps, we took different hyperparameter like learning rate, batch size during training because adjusting hyperparameters like the learning rate, batch size and choice of optimizer is vital for optimizing the performance of neural networks. The Learning Rate (LR) drops to 0.00099, which was the initial value. Additionally, the gradient norm is acting as intended; it remained neither too high nor too low during the training phase. Because of the feeding of ground-truth frames throughout training model, another statistic is the assessment loss, which is significantly greater than the training loss.Through the process,the evaluation loss was convergent, indicating that there is no overfitting the model.

Synthesis Phase

Tacotron-2 (two) model was trained without Wave-Net, the Griffin-Lim Algorithm (GLA) shall be presented as a vocoder to evaluate the growth of the trained feature prediction model. In this study, Wave Net was not used to train the Tacotron-2 (two) model and the Griffin-Lim – Algorithm (GLA) used as a vocoder to assess the trained feature prediction model's development. Even when complex and uncommon words are present, the audio produced by the Tacotron-2 (two) accurately corresponds to the Text. However, there are instances where the prosody is not natural, with silences or pauses occurring at unpredicted points in the sentence.

Both linear and Mel spectra-gram representations of audio inputs were used to train Tacotron-2 (two) model. Consequently, both linear and Mel spectrograms were used to evaluate this model. Ten additional sentences in both forms have been synthesized using this model for the user evaluation.

Deep Voice (DV) 3 Implementation

The Pytorch from Ryuichi-Y1 was used for the implementation of Deep Voice (DV) 3 implementation. This implementation mainly has two (2) phases or stages. The first phase or stages is training phase where the training of Deep Voice (DV) 3 in which the speech dataset has conducted and the second pahse is called synthesis phase where new Text has conducted or given to the model as an input and make the synthesis. During this phase the trained model was used with Griffin Lim Algorithm (GLA) to perform the synthesis properly. To discuss each of Deep Voice (DV) 3 implementation, it’s discussed as below.

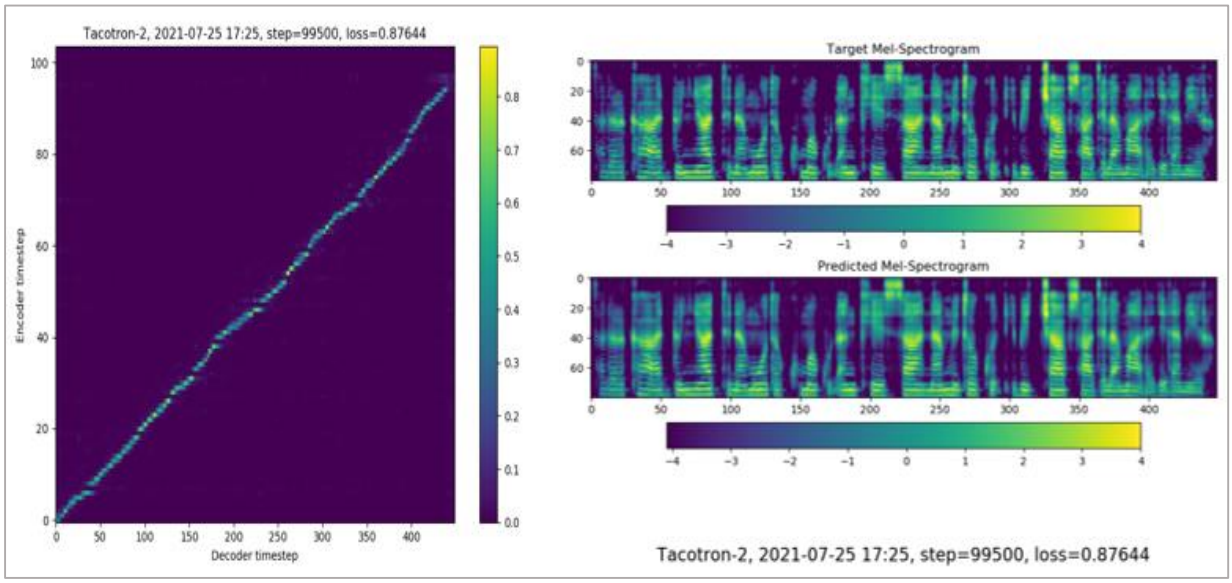

Fig. 3: Sample of training alignment

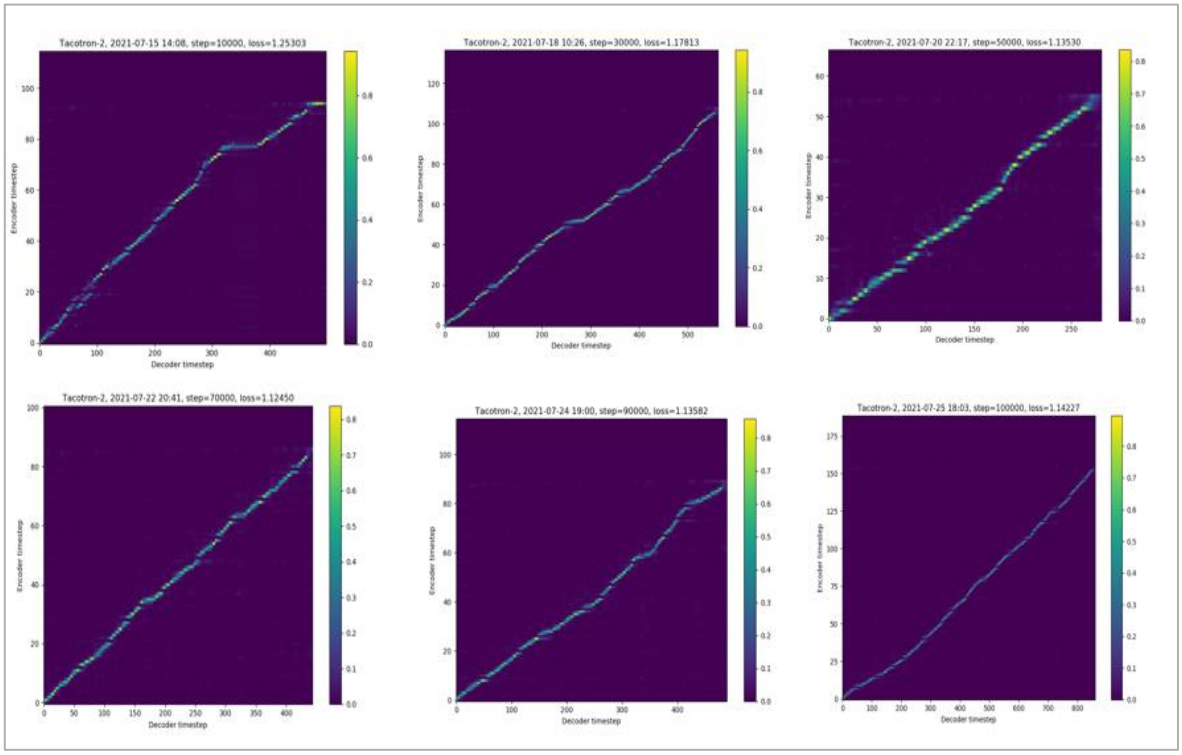

Fig. 4: Show-attention changing during the evaluation

Training Phase of Deep Voice (DV) 3

The NVIDIA Tesla T4 GPU with 16GB of RAM was used to train Deep Voice (DV) 3. Using a batch size of eight, the entire training period lasted twelve days and sixteen hours and fifteen minutes. After several steps, the training was finished. Figure (6) shows how attention changes throughout the training period. After 44 k steps the model began to generate speech that was understandable, comprehensible and somewhat human-like, as shown in the picture. Afterwards, the model improved by becoming less robotic in the speech synthesized.

DOI: 10.3844/jcssp.2025.1037.1048

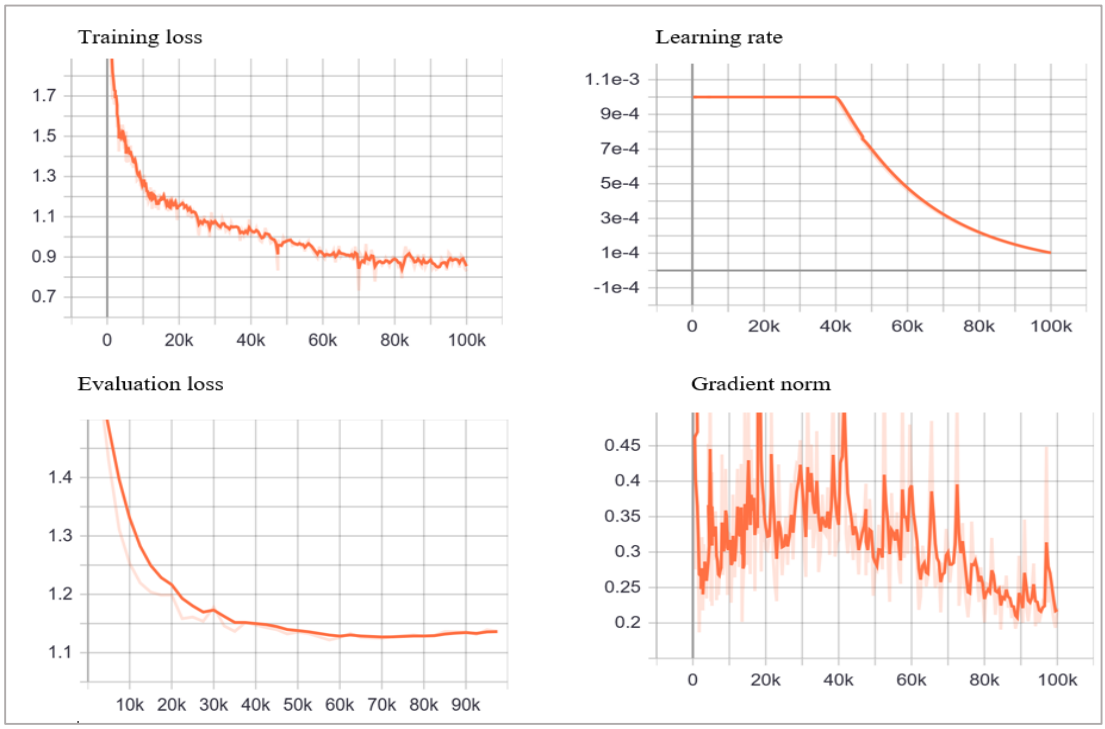

Fig. 5: Illustrates the performance of the Tacotron-2 (two) model

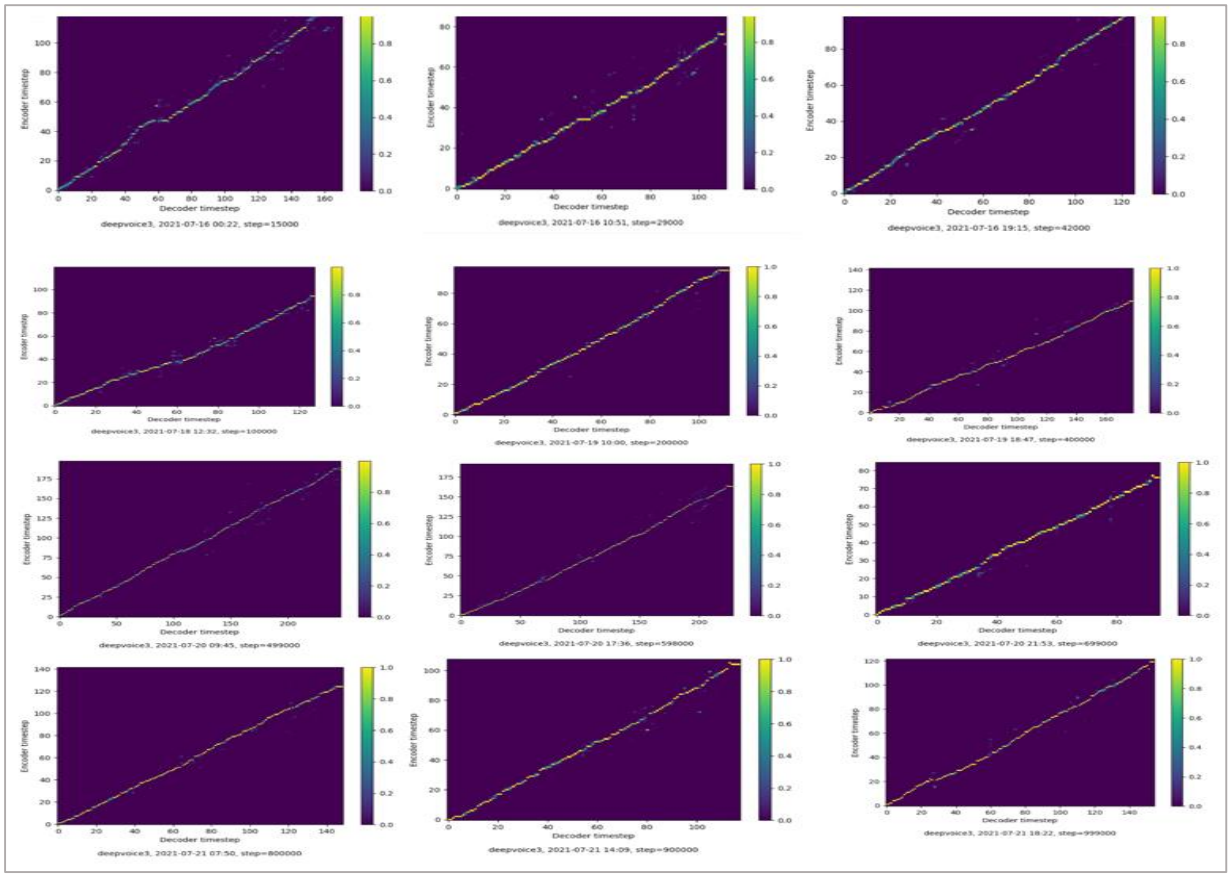

Fig. 6: Shows how the DV3 model's attention changes through training

Before the training process began, the model's training was assessed at 10-K stages. Six unseen statements that were deliberately selected to be decisive in order to assess the models' abilities were used in the evaluation. Special cases in the Afaan Oromo language are covered in the sentence examples, including long words, compound words, glottal stop (Huda) characters and commas to check if the model pauses appropriately when synthesizing the speech, as well as sentences that end with question mark(?) a full stop(.) and exclamation mark(!) to check if the model can alter the intonation in the synthesized speech.

The Fig. (7). Illustrates the shift in focus for the following sentence: "Kunoo, Waaqayyo gooftaan keenya ulfina isaa fi guddina isaa nu argisiiseera, nuyis ibiddicha gidduudhaa sagalee isaa dhageenyeerra;"

Figure (8) displays the evaluation metrics that interpretation for the training models' performance. An indicator of how distant an expected value deviates from its actual value is loss function. Since minimizing the loss function is the primary goal, the deep neural network algorithm iterates as numerous times as necessary to achieve the flattest possible loss shape. In this instance, the loss functions exhibit the intended behavior; they progressively decline and, following 96 K steps over four days and eleven hours (h) of training, converge to a value of 0.209.

In this case, the learning rate is crucial to minimizing the Loss-Function (LF). It determines the rate at which the model must acquire or learn from. It is important to set this parameter appropriately since, for example, if it the input is set too much, the model took long time to learn anything. In this instance, the starting Learning Rate (LR) was set to 0.0005 in accordance with the deep voice V3 architecture. It drops to 0.00010300 after 98 K steps, or four days and 10 hours and 30 minutes of training. The gradient norm displayed in the same image determines the L2 (two) norm of the inclines of the deep learning network's last layer. By dynamically adjusting gradient weights, it automatically balances training in deep learning models. It may suggest a disappearing gradient if its value is too low.

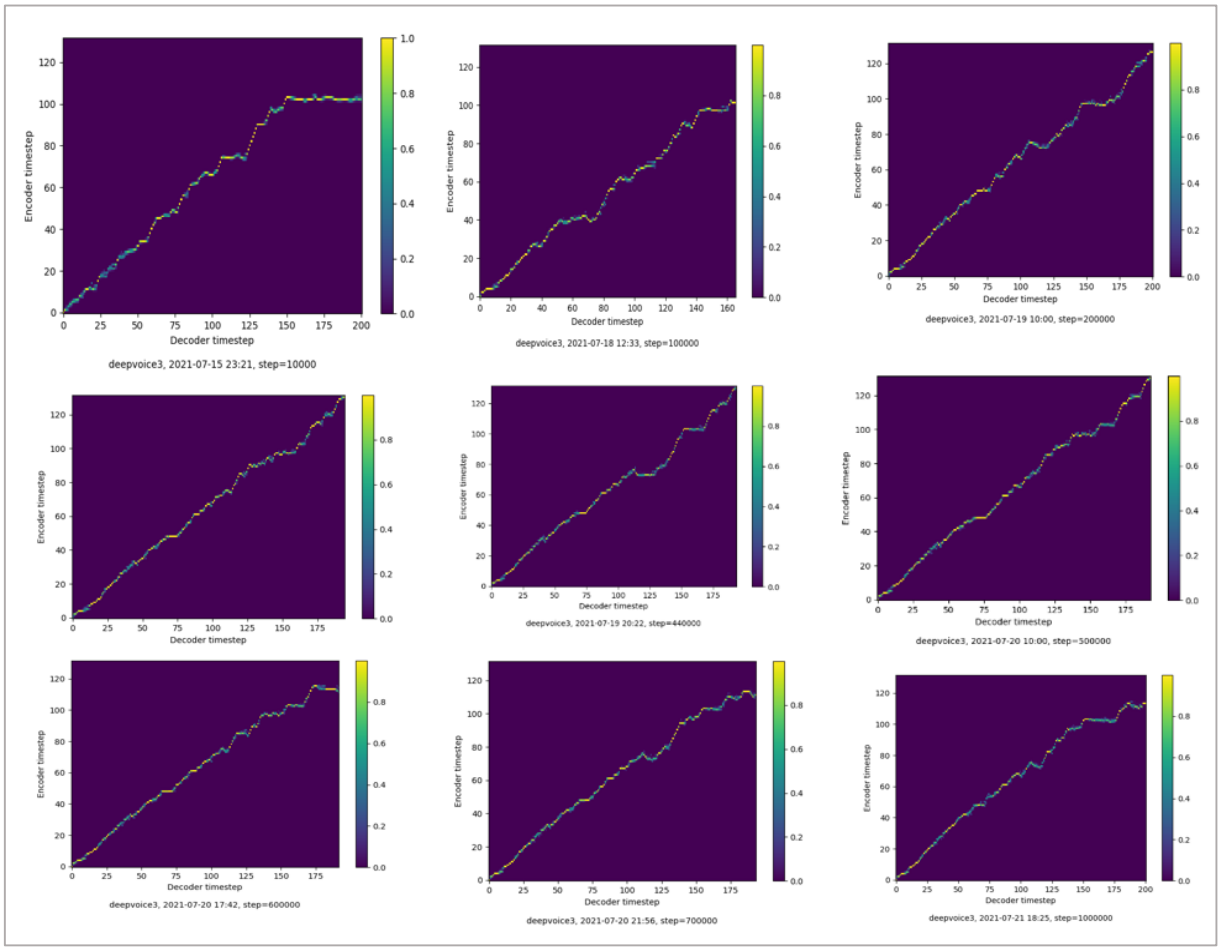

Fig. 6: Attention alignment plot for the above sentence

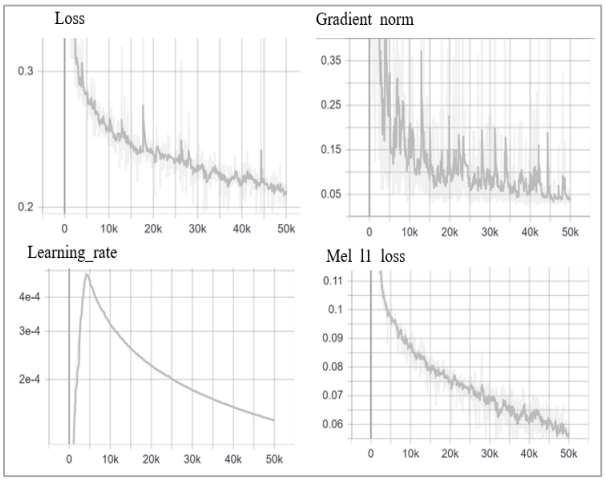

Fig. 8: Learning models' tensor board scalars

The deep learning network's upper layers are impacted by this issue, which makes it challenging for the network to learn and adjust the constraints. Further, a value that is too high may indicate that a gradient phenomenon is exploding. The buildup of significant error gradients throughout the training phase leads to very large updates in the deep learning model weights, which makes the model unstable and unable to learn from data. The L1 norm, which measures the model's capacity to forecast Mel-spectrograms, is the final performance indicator shown in Fig. (8). Over the course of the iteration steps, the metric's L1 norm decreases to 0.04098 after 110 K steps, or five days and eight hours and 30 min of training.

The language-specific front-end that normalizes uncommon or non-standards words like abbreviations, numbers, currencies and acronyms must be run through when fresh Text is fed into the model. This is discussed in the text and audio preprocessing steps. To maintain the order, lengthy input sentences are also divided into segments and batched. The last checkpoint on which a comprehensible voice is synthesized predicts each of them. The segments are combined and concatenated during audio post-processing so that they can be exported as a single.wav file. The last checkpoint for evaluation of the model performance was done by user which used to synthesize 12 additional sentences at final steps.

Evaluation and Discussion

Now days there are several methods used to evaluate text-to-speech synthesis systems. To assess the model's performance, participating native speakers provide both objective and subjective assessments were used. twelve new sentences that are not visible during training have been written specifically for this purpose. Acronyms, Ordinals, numbers, abbreviated-words and punctuation-signs were among the Nonstandard Words (NSW) that were taken into account when preparing the utterances and their impact on synthesis quality were explored. In order to verify that the model pauses appropriately when synthesizing the speech, the utterances additionally include complex and lengthy sentences, glottal stop (Huda) characters and commas at various points in the sentence. Volunteers from Gambella University (GU) and Mattu University (MaU) took part. A questionnaire focus on mainly on the Intelligibility and Naturalness of the synthesized speech has been shaped prior to the evaluation.

Objective Evaluation

The purpose of the objective evaluation is to assess how well the models function. As stated by Ping *et al*. (2018), the attention-based Text to speech systems run into attention errors such as repeated words, mispronunciations and skipped words. To measure or asses the performance of the models against the stated errors. The objective evaluation of the models has been carried out. The evaluation of speech along with its normalized sentence is given to the participants and The participants listen to the evaluation speeches and manually calculate the occurrence of the above stated errors.

After both models evaluated the errors in the evaluation sentences describes as follow. In each utterance, the participants were instructed to tally the instances of mispronounced words, repeated words and words that were skipped. One error per utterance is considered to have occurred when one or more mispronunciation skips and repeats occur and Twelve in numbers normalized sentences totaling 248 words after normalization are included in the evaluation sentences list. After evaluation was performed the following results was obtained: For Deep Voice (DV) 3 model 8 Repeated words,10 Mispronunciations and 2 Skipped word. For Tacotron-2 (two) 2 Mispronunciations, 0 repeated words and 0 skipped words.

Subjective Evaluation

The naturalness and comprehensibility of the trained models have been assessed subjectively. Mostly we used subject evaluation because of fluency in the Afaan Oromo language was a prerequisite for the experiment's subjective evaluation and a Mean Opinion Score or scale (MOS) hearing tests were used to assess the Text to speech synthesis model subjectively. It has been recommended subjective performance assessment of the synthesized speech quality since 1994 (Rec, 1994; Chu *et al*., 2019). In Afaan Oromo language fluency in was a prerequisite for the experiment's subjective evaluation. Prior to taking respondents received a printed the questionnaire form, the listening test, the respondents received eleven items on the survey have the numbers 5-1 next to phrase anchors; the number 5 denotes an item's higher score. Participants were instructed to locate a

quiet- listening space and put on head-phones before the start of the test. Additionally, they were asked not to change the volume while the test was being administered and were told to set it to a comfortable level. During the test, adjusting the audio volume could have an impact on the outcomes. Eleven replies from volunteers were gathered for the MOS test and Table (1) displays the overall mean opinion score result.

Discussion

The first experiment involves adjusting the hyper parameters to train the Tacotron-2 (two) feature prediction model, which is based on a Recurrent Neural Network (RNN). Before training the latency and computational cost of real-Time Text-to-Speech (TTS) synthesis are critical considerations for practical deployment, especially in applications like voice assistants, live translations, or accessibility tools. For instance, factors influencing the latency incudes: Model complexity, inference speed, hardware constraints, input length and resource availability. The training which performed by this model was extremely slow due to the recurrent neural network. Both linear and Mel spectra-gram representations of the unprocessed audio signals are used for training. After 50 k steps, the experiment began to yield a satisfactory outcome; in total, the model was trained for 110K steps.

This inefficacy may due to different factors like hardware constraints, input length and resource availability.

In this experiment, the predicted acoustic properties were transformed into audio time-domain waveforms using the Griffin-Lim-method. The MOS test According to both objective and subjective assessments, this experiment was effective.

In the second experiment, CNN was used which stads for convolutional neural network to train Deep Voice (DV) 3 Using different convolutional layers on all of its sub components made and this trial faster than the first. To get an acceptable outcome, the experiment was run for 1 million steps and Mel spectrogram representations of the unprocessed audio sources are used to train the model.

Table 1: Displays the overall Mean Opinion Score(MOS)

Different model type and Experiment	Intelligibility	Naturalness
Deep Voice(DV) 3	3.32	3.04
Tacotron -2 (two) (Linear spectro-gram)	4.34	4.16
Tacotron-2 (two) (Mel-spectro-gram)	4.36	4.33

The anticipated acoustic properties have been utilized to create audio files using the Griffin-Lim method. Sentences from a test list shown in Fig. (6) had attention alignment curves that were inferior to Tacotron 2's, which lowers the MOS score. There are artefacts in the Deep Voice (DV) 3 model's talks and the voice did not sound continuous and fluid. Since the Tacotron-2(two) feature prediction model makes extremely few attention errors, it performs well in terms of attention error. Tacotron-2 (two) made two quantitative errors out-of 248 words, while Deep Voice (DV) 3 made 16 attention errors out-of 248 words. The lack of an Afaan Oromo language phoneme dictionary that includes words with their pronunciation for use with an explicit grapheme to phoneme model and the attention mechanism used because Deep Voice (DV) 3 uses a dot product attention mechanism and Tacotron-2 (two) replaces it with one that is location-sensitive, which may reduce attention errors, are the two reasons for this.

The Mean Opinon Score/scale (MOS) listening test used to evaluate the synthesized speech samples' naturalness and intelligibility. Tacotron-2 (two) was trained with with both Linear Spectro-gram (LS) and a Mel Spectro-gram (MS) not only mel specto-gram and Deep Voice (DV) 3 was trained with with Mel Spectro-gram (MS). Finally, Tacotron-2 (two) trained with a Linear Spectro-gram (LS), Deep Voice (DV) 3 trained with a Mel Spectro-gram (MS) and Tacotron-2 (two) trained with a Mel Spectro-gram (MS) are used for the MOS exam. The Deep Voice (DV) 3 model received scores of 3.32 and 3.04 out-of-five for intelligibility and naturalness, respectively. The Tacotron-2 (two) trained on a linear spectrogram received a naturalness score of 4.16 out-of five and an intelligibility score of 4.34 out-of five. In contrast, the Tacotron-2 (two) trained with the Mel spectrogram received scores of 4.33 and 4.36 out-of five for naturalness and intelligibility respectively.

After all the evaluation result shows that the Tacotron-2 (two) model trained on Mel Spectro-gram (MS) representation performs better overall than the other models, as discussed in objective evaulation. A very similar result to the first one is likewise obtained using the Tacotron-2 (two) trained on linear spectrogram. To be approved as a functional text-to-Speech model, the Deep Voice (DV) 3 model must significantly enhance its performance.

Conclusion

This study presents an attempt to train a deep neural network model on text-to-speech synthesis system for Afaan Oromo language. After many experiments, the fully convolutional neural network Deep Voice (DV) 3 and the Recurrent Neural Network (RNN) based Tacotron 2 (two) models were successfully trained and presented. Quantitatively, the Tacotron-2 (two) model made only two attention errors out-of 248 words, while the Deep Voice (DV) 3 model made 16 attention errors out- of 248

words. Deep Voice (DV) 3 and Tacotron-2 (two) received MOS scores of 4.36 and 3.32 out-of five for intelligibility and also 4.33 and 3.02 out-of five for naturalness, respectively. Additionally, according to the evaluation results, the Tacotron-2 (two) model performs better than the Deep Voice (DV) 3, making it suitable for a range of applications like smart education, telephone inquiry services and recommendation systems.

Future Works

In the future research the usage of a learning-based technique, like WaveNet, which mimics the process of creating audio waveforms using a deep neural network and The Griffin-Lim vocoder introduces artifacts and limits the naturalness of synthesized audio. So training using modern Wave Glow or WaveNet vocoders and Generative Adversarial Network (GAN) model will be suggested to address the limitations of Griffin-Lim methods in future work. The machine learning classifier with numerous features can be used to incorporate all nonstandard word normalization.

The trained models and the source code are openly accessible on Google Drive for additional development: https://drive.google.com/drive/folders/18fcBCwvpL2FOy1-k9ak0cCd3bbyW1uy?usp=sharing andhttps://drive.google.com/drive/folders/18uncV0bovqBN76o2TFXmfu28QBuwS8QN?usp=sharing.

Acknowledgment

We acknowledge Dr. Arulmurugan Ramu, Associate Professor Department of Computer Science, Heriot-Watt University his advice, motivation, input and support during the creation of the manuscript.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author's Contributions

Diriba Gichile Rundasa: Data collection, analysis and design of the study.

Arulmurugan Ramu: The contributor and mainly contributed the analysis of the data and design of the study.

Teshale Debushe Adugna, Chala Sembeta Teshome and Desalegn Tasew: Collection of data and the analysis of the data for study was done by these contributors.

Ethics

This manuscript is an original work. The corresponding authors declares that no ethical concerns associated with this submission.

References

Arık, S. Ö., Arik, S., Diamos, G., Gibiansky, A., Miller, J., Peng, K., Ping, W., ... & Zhou, Y. (2017). Deep voice 2: Multi-speaker neural text-to-speech.

Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 1171–1179.

Chu, S. K. W., Ravana, S. D., Mok, S. S. W., & Chan, R. C. H. (2019). Behavior, Perceptions and Learning Experience of Undergraduates Using Social Technologies During Internship. *Educational Technology Research and Development*, *67*(4), 881–906. https://doi.org/10.1007/s11423-018-9638-2

Hunt, A. J., & Black, A. W. (2002). Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 373–376. https://doi.org/10.1109/icassp.1996.541110

Lemmetty, S. (1999). *Review of speech synthesis technology*. https://doi.org/http://www.acoustics.hut.fi/~slemmett/dippa/%5Cnhttp://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/

Morka, M. (2001). *Text to Speech system for Afaan Oromo Languageo. A Thesis Submitted in Partial Fulfilment of the Requirement for the Degree of Master of Science in Information Science*.

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2020). WaveNet: A Generative Model for Raw Audio. *A Generative Model for RawAudio*. https://doi.org/http://arxiv.org/abs/1609.03499

Ping, W., Peng, K., Gibiansky, A., Arık, S., Kannan, A., Narang, S., Raiman, J., & Miller, J. (2018). Deep Voice(DV) 3: Scaling Text to Speech with convolutional sequence learning. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 1–16.

Rec, I. T. U.-T. (1994). *P.85: A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, *27*.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4779-4783). IEEE.

Tamrat, D., Ch, A., & Muluken, H. (2022). Design and Development of a Text to Speech Synthesizer for Afan Oromo. *SN Comput. Sci*, *3*(420). https://doi.org/https://doi.org/10.1007/s42979-022-01306-7

Wosho, M. K. (2021). *Text to Speech Synthesizer for Afaan Oromo language Using Hidden Markov Model*. *02*(03), 21–25.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. *Interspeech 2017*, 4006–4010. https://doi.org/10.21437/interspeech.2017-1452