INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Text to Speech Synthesizer for Tigrigna Linguistic using Concatenative Based approach with LSTM model

Mezgebe Araya[1], **Minyechil Alehegn**[1]*

1 School of Computing and Informatics, Mizan Tepi University, Ethiopia

OPEN ACCESS

Received: 13.11.2021

Accepted: 30.12.2021

Published: 21.01.2022

Citation: Araya M, Alehegn M (2022) Text to Speech Synthesizer for Tigrigna Linguistic using Concatenative Based approach with LSTM model . Indian Journal of Science and Technology 15(1): 19-27. https://doi.org/10.17485/IJST/v15i1.1935

* **Corresponding author**.

minyechil21@gmail.com

Funding: Article Processing Charges was defrayed partially by iSEE

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The purpose of this study is to describe text-to-speech system for the Tigrigna language, using dialog fusion architecture and developing a prototype text-to-speech synthesizer for Tigrigna Language. **Methods :** The direct observation and review of articles are applied in this research paper to identify the whole strings which are represented the language. Tools used in this work are Mathlab, LPC, and python. In this paper LSTM deep learning model was applied to find out accuracy, precision, recall, and Fscore. **Findings:** The overall performance of the system in the word level which is evaluated by NeoSpeech tool is found to be 78% which is fruitful. When it comes to the intelligibility and naturalness of the synthesized speech in the sentence level, it is measured in MOS scale and the overall intelligibility and naturalness of the system are found to be 3.28 and 3.27 respectively. Based on the experiment LSTM Deep learning model provides an accuracy of 91.05%, the precision of 78.05%, recall of 86.59 %, and F-score of 83.05% respectively. The values of performance, intelligibility, and naturalness are inspiring and show that diphone speech units are good candidates to develop a fully functional speech synthesizer. **Novelty:** The researchers come up with the first text to speech LSTM deep learning model for the Tigrigna language which is critical and will be a baseline for other related research to be done for Tigrigna and other languages.

Keywords: LSTM; speech synthesis; Tigrigna syllables; TexttoSpeech; Concatenative approach

1 Introduction

Deep learning is one type of machine learning which is a subset of Artificial intelligence. It is about computer learning to think using structures modeled on the human brain. DL can analyze videos, images, and unstructured data in various ways ML can't easily do. Every industry will have career paths that involve machines and deep learning. Language is an important part of everyday life of human beings. Whether we are using speech, sign language, passion, or a coding system that conveys meaning through touch. We use language to express our thoughts, intentions, reactions, and experiences[1]. Text-to-speech synthesizer transforms language information stored text into speech. It is most widely used in audio reading devices for visually impaired people nowadays. TTS is one of the major applications of NLP. The NLP module of the general TTS synthesizer consists of the Pre-processor, text analyzer, and contextual analyzer(2). Synthesized speech can be created by concatenating part of recorded speech which is stored in a database. Speech is often based on concatenation of natural speech that is the units, which are taken from natural speech put together to form a word or sentence(3).text-to-speech synthesis system has a wide range of applications in everyday life and a text-to-speech synthesizer is used for vocalization processed content(4). In the last decade, a great deal of TTS-Synthesis system has done much work in various languages as well as different synthesis techniques such as Unit-selection, Formant, Hidden Markov Model, and Articulatory synthesis was done by researchers(5). To make the computer systems more interactive and helpful to the users, especially physically and visually impaired and illiterate masses, the TTS synthesis systems are in great demand for the Ethiopian languages(6). Research in the area of speech synthesis has been worked by the growing importance of many new applications which includes information retrieval services over the telephone such as banking services, public announcements at places like train stations, and reading out manuscripts for gathering(7). Special tackle for the physically defied, such as word workstations with reading-out capability and book-reading aids for visually challenged and speaking aids for the vocally defied also use speech synthesis(8). The perceived quality of standard general-purpose TTS systems is not good enough, which forces application developers to use pre-recorded prompts, drastically reducing the text generation plasticity. Recent improvements in limited-domain amalgamation have been in the context of concatenative synthesis, with a focus on methods for combining whole phrases and words with sub word units for infrequent words. As the complexity of the sphere increases, there is more room for the prosodic inconsistency that must be accounted for to achieve natural speech(9). According to Ethnology, there are 2.4 million Tigrigna speakers in Eritrea(10). The orthographic representation of the language is organized into orders. Each of the 35 consonants has seven orders (derivatives). Out of the 35 consonants, four of them are diphthongs, six of them are CV combinations while the 7th is the consonant itself. The way Tigrigna orthographic characters are written is very similar to the way they are spoken. It means Tigrigna is a phonetic language. The mapping of the written form and the spoken form is one to one except the epenthetic vowel. The main purpose of this study is to apply the state of art of deep recurrent neural network model long short-term memory (LSTM) for text to speech system for Tigrigna language and adapting the new model on Tigrigna language using the collected dataset.

2 Related Work

Speech synthesis is the progression of adapting a written manuscript into linguistic and this technology has the ability to adapt arbitrary text into distinct speech, intending to be able to afford textual data to people via speech messages(1). NLP module of the general TTS synthesizer consists of the Pre-processor, text analyzer, and contextual analyzer(2). The components are automatic phonetization, text analysis, and prosody generation(3,4). There are several factors which is affected NLP and the final output of digital signal processing work like, quality of the microphone, environmental echo,noise, and sampling frequency. The formant method acting uses the source-filter possibility of voice communication production, where auditory communication is modeled by Deck lung of the filter model. Rule-based formant analysis can produce quality dialog which Timbre peculiar(5). DSP component are the computer analog of dynamically controlling the articulatory muscles and the vibratory frequency of the vocal folds so that the output signal matches the input requirements(6). Synthesized speech can be produced by employing several different techniques to find natural human-like sounds(7). The speech synthesizer depends on the TTS synthesizer architecture inculcated to produce intelligible and natural sounds from the synthesizer(8). Text-to-phoneme (T2P) is beleaguered to harvest phonetic transcription of the text, together with the wanted prosodic features(9). A speech synthesizer is one such interface facilitating people to amalgamate with the digital era(10). Pronunciation synthesis tries to worthy the human spoken language production awkward to utilize(11). Speech technologies are vastly used and has unlimited uses(12). Concatenative talking is producing intelligible & natural synthetic speech, usually close to a real voice of a person(13). Concatenative synthesizers are modest to only one speaker and one sound. The difference between earthy variation in speech signals and the nature of the automated techniques is segmenting the waveforms from the audible output(14). Text-to-speech conversion (TTS) has a wide variety of applications(15). Recently, synthesized speech quality based on deep neural networks was found as intelligible as human voice. (16). Our knowledge of human speech processes is still incomplete, the quality of text-to-speech is far from natural-sounding(17). Here the researcher generates and analyze the prosodic information from the recorded Sindhi sounds using the back propagation neural network(18).

3 Material and methods

3.1 Propose Method

This study describes the TTS system for the Tigrigna linguistic, using speech analysis architecture. When we use concatenative speech synthesis where the segments of recorded speech are concatenated to produce the desired output.

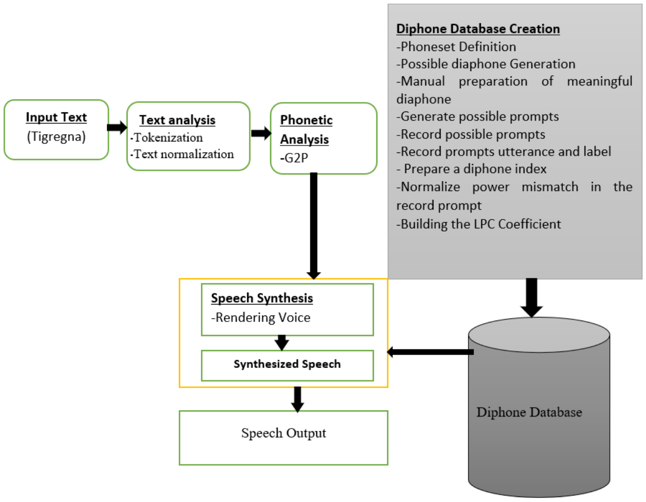

Input Text
(Tigregna)
Text analysis
-Tokenization
-Text normalization
Phonetic Analysis
-G2P
Diphone Database Creation
-Phoneset Definition
-Possible diaphone Generation
-Manual preparation of meaningful diaphone
-Generate possible prompts
-Record possible prompts
-Record prompts utterance and label
- Prepare a diphone index
-Normalize power mismatch in the record prompt
-Building the LPC Coefficient
Speech Synthesis
-Rendering Voice
Synthesized Speech
Speech Output
Diphone Database

Fig 1. Proposed work Architecture

3.1.1 Data base construction

I. First a phone-set that consisted every phones of the language are prepared.

II. Along with the phone symbols, features such as vowel height, place of articulation and voicing are defined.

III. The synthesis quality is naturally bound to the speech database from which the units are selected.

IV. This set of selected segments is recorded in a quiet room in a typical multimedia computer system.

V. Annotated speech data is taken and manually segmented the syllable boundaries using Praat tool for accuracy.

VI. Annotation of the recorded sound file is done using the TextGrid objectfinally syllables are recorded with a sampling frequency of 16 KHz and represented using 16-bits.

```
Algorithm to read file
Began
Step1: Create a database of various wave files
Step2: Create a text file (.txt)
Step3: Open the .txt file in matlab.
Step4: Read the file opened.
Step5: For every character read, play the corresponding wave (.wav) file.
End
```

3.1.1.1 Text Analysis

. The first step of the Text-to-Speech system designed is text analysis. That means analysis of raw text into pronounceable word. It involves the task of cutting down sentences posed to the system into individual unambiguous words. For instance, if one poses the sentence ‘′, ‘⊠⊠ ⊠⊠⊠ ⊠⊠⊠’, which means ‘he eat a food’ the tokenizer, which is part of the text analysis component, identifies three tokens, ‘⊠⊠’, “⊠⊠⊠”, and ‘⊠⊠⊠’.

3.1.2 Data Collection and preparation

In this work 35x 35 diphones are recorded to create the TTS model for Tigrigna language communication. In addition, to test the text to speech model 10 sentences and 100 words were used.to check the performance. 20 native speakers participated from them 12 men and the remaining 8 are women.

3.1.3 Tools and Techniques

Since connecting prerecorded natural utterances using concatenative synthesis is the easiest way to produce intelligible and natural sounding speech[4], a prototype TTS for Tigrigna language is developed using a concatenative synthesis. The toolkit that is used for this speech synthesis model is called MATLAB and WAVESURFER 1.8.8 and PRAAT. For the purpose of recording wave files microphone was used. PRAAT, used to record and analyze strings of Tigrigna linguistic, MATLAB used for implementation of TTS synthesizer, and Neospeech and python applied for testing purposes to check performance. In this paper, the collected data was analyzed using PRAAT. Natural sounds are gathered from various articles, and newspapers of Tigrigna linguistic and analyzed to words, phones, phrases, and sentences for checking of performance the TTS synthesizer.

3.2 Models

3.2.1. LSTM

One kind of RNN.The four important layers in LSTM models are point-wise operation, vector transfer, neural network layer, concatenate and copy.

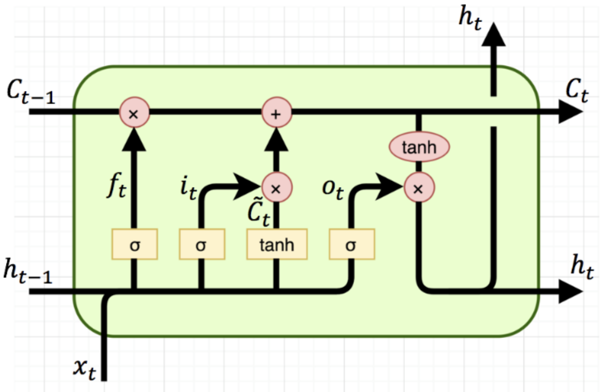


Fig 2. Long short-term memory (LSTM) structure

the three gates are computed as follow

$$f_{(t)} = \sigma\left(W_{fx}X_{(t)} + W f_h h_{(t-1)} + b_f\right) \tag{1}$$

$$i_{(t)} = \sigma\left(W_{ix}X_{(t)} + W i_h h_{(t-1)} + b_i\right) \tag{2}$$

$$o_{(t)} = \sigma\left(W_{ox}X_{(t)} + W o_h h_{(t-1)} + b_o\right) \tag{3}$$

where σis a nonlinear activation function

$$C_{(t)} = \tanh\left(W_{cx}X_{(t)} + W_{ch}h_{(t-1)} + b_c\right) \tag{4}$$

$$C_{(t)} = f_{(t)}\Theta C_{(t)} + i_{(t)}\Theta C_{(t)} \tag{5}$$

$$h_{(t)} = O_{(t)}\Theta \tanh\left(C_{(t)}\right. \tag{6}$$

Where tanh =the none linear tanh activation function

Θ is pointwise that used to denote multiplication operation for two vectors

3.2.2. Linear productive Coding model

a method to stand for and analyze human speech. Mental representation when using LPC is defined with LPC coefficients and an error signal, instead of the original speech signal.

$$S[\mathrm{n}] = \textstyle\sum_{k=1}^{n} a_k S[n-k]) \tag{7}$$

4 Results and Discussion

Based on the experiment our proposed method showed better results than the previous. The comparison of the proposed and the previous model is shown in [Table **??**] in the discussion section.

Table 1. Scales used in MOSMOS

Value	MOS
1	Excellent
2	Very Good
3	Good
4	Fair
5	Bad

Then the evaluators provide their ranks based on the MOS scale as shown in [Table 1]. To evaluate the synthesizers' intelligibility and naturalness ten sentences are prepared as a test data for the synthesizer. All words used in the sentence are found in the compiled lexicon. Then the selected individuals listen to the synthesized waveform from the synthesizer and evaluate naturalness and intelligibility based on the MOS scale. The invited native Tigrigna speakers are given with the questionnaire to evaluate intelligibility and naturalness of the synthesized speech.

Table 2. Intelligibility (MOS) Scores of Tigrigna Speech Synthesizer

Test Data (Sentence)	Ranks given by different evaluators for intelligibility																			
	p 1	p 2	p 3	p 4	p 5	p 6	p 7	p 8	p 9	p 10	p 11	p 12	p 13	p 14	p 15	p 16	p 17	p 18	p 19	p 20
1	4	3	4	2	4	3	3	4	2	3	4	3	3	3	2	4	4	4	3	3
2	3	3	3	3	3	4	3	3	3	4	4	3	3	2	3	3	4	4	3	2
3	4	4	3	3	3	4	2	3	3	3	3	4	3	2	2	3	5	5	4	3
4	3	4	3	4	3	5	2	3	3	4	3	3	4	3	4	3	3	4	4	3
5	3	3	3	4	4	3	3	3	4	3	3	4	3	3	4	4	3	5	3	3
6	3	3	4	2	3	3	2	3	3	3	2	3	2	4	3	3	4	3	3	4
7	4	3	4	3	4	2	3	4	3	3	3	4	3	2	3	3	2	4	4	2
8	2	4	4	3	3	3	3	4	4	3	2	3	4	3	4	4	4	3	4	3
9	3	3	3	4	4	3	3	4	3	3	4	3	4	3	2	4	3	3	3	3
10	4	4	3	4	3	3	3	3	2	4	4	3	4	4	3	3	3	4	4	3

Where P is the persons who are invited to evaluate the intelligibility of text to speech synthesizer for Tigrigna language.

Table 3. Naturalness (MOS) Scores of Tigrigna Speech Synthesizer

Test Data (Sentence)	Ranks given by different evaluators for naturalness																			
	p 1	p 2	p 3	p 4	p 5	p 6	p 7	p 8	p 9	p 10	p 11	p 12	p 13	p 14	p 15	p 16	p 17	p 18	p 19	p 20
1	3	4	3	2	3	2	4	3	3	3	3	4	3	4	3	3	4	3	4	2
2	4	3	3	3	3	3	3	3	2	3	4	4	3	3	3	4	3	5	3	3
3	4	4	3	3	4	4	3	3	4	3	4	3	4	4	3	3	4	4	3	3
4	4	4	3	4	3	4	3	4	3	3	2	2	4	3	4	4	3	3	3	4
5	3	4	3	3	3	3	2	3	3	4	3	3	3	4	3	3	4	4	4	3
6	3	3	4	2	3	2	3	4	3	3	3	4	3	3	3	4	5	3	3	3
7	4	4	4	3	2	3	4	3	4	4	2	3	2	3	4	3	3	3	5	3
8	2	4	4	3	3	3	2	3	3	2	3	2	3	3	5	3	3	4	3	3
9	2	4	3	4	4	4	3	3	3	2	3	4	3	4	3	4	4	4	3	4
10	4	3	3	3	4	3	4	4	3	3	4	4	3	4	2	4	3	3	3	2

Where P is the persons who are invited to evaluate the naturalness of text to speech synthesizer for Tigrigna language.

Table 4. Average MOS Scores of Tigrigna Speech Synthesizer

Sentence	Intelligibility	Naturalness
1	3.25	3.15
2	3.15	3.25
3	3.30	3.50
4	3.40	3.35
5	3.40	4.25
6	3.00	3.20
7	3.15	3.30
8	3.35	3.05
9	3.25	3.40
10	3.40	3.30

The overall intelligibility of the system from twenty listeners for the ten Tigrigna sentences is found to be 3.27. Which means the synthesizer is 'good' as per the scale of the MOS test. The overall naturalness of the synthesizer found to be 3.28 which also approach to 'good' MOS scale. These values of intelligibility and naturalness look encouraging to come up with a better system. Though we tried to record the corpus data in a quiet room, there was also some noises from the computer itself listened during the synthesis of the data which degrades the naturalness of the sound for the listeners.

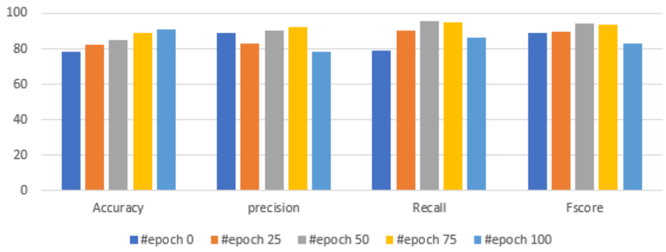


Fig 3. Performance of LSTM in different Epoch number

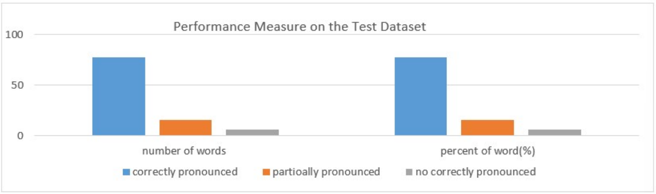


Fig 4. Performance Measure on the Test Dataset

The test consists of 100 words designated through the help of a native. The words evaluated their quality and intelligibility using Neo Speech. Therefore, the chosen words for the Neo Speech and listen to their naturalness and understandability of the sound which is played by the Neo Speech online. The overall performance was measured in terms of the number of correctly pronounced words over the number of words played. correctly pronounced the overall performance of the system is found to be 78%. The overall intelligibility of the system from twenty listeners for the ten Tigrigna sentences is found to be 3.27. This means the synthesizer is 'good' as per the scale of the MOS test. The overall naturalness of the synthesizer was found to be 3.28 which is also an approach to the 'good' MOS scale as it is shown in [Table 4]. The MOS score of the intelligibility shown in detail at [Table 2] and naturalness score of MOS described at [Table 3]. As [table 5] shown that modern deep neural network models are better than the traditional machine learning methods .using the LSTM model, it provides different results in different epoch numbers which are from 0,25,50,75,100 respectively. LSTM provides better accuracy of 91.05% at epoch 100, the precision of 92.01 at epoch number 75, recall of 95.69 % at 50, and fscore of 94.02% at epoch 50 respectively shown in [Figure 3]. Based on

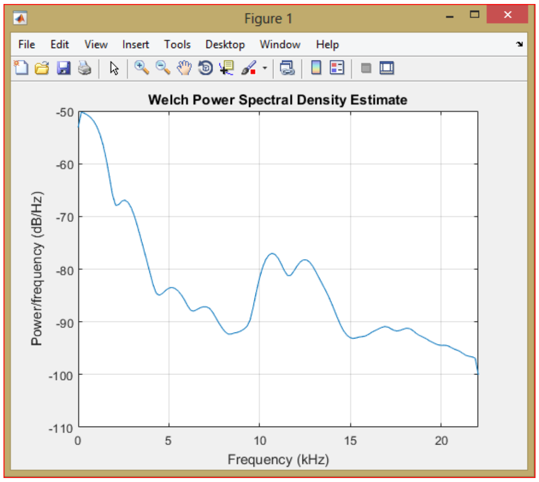


Fig 5. Formants of the word "arba"

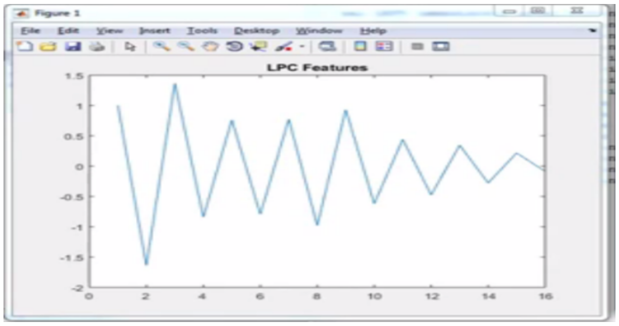


Fig 6. The LPC feature extraction of the word "arba"

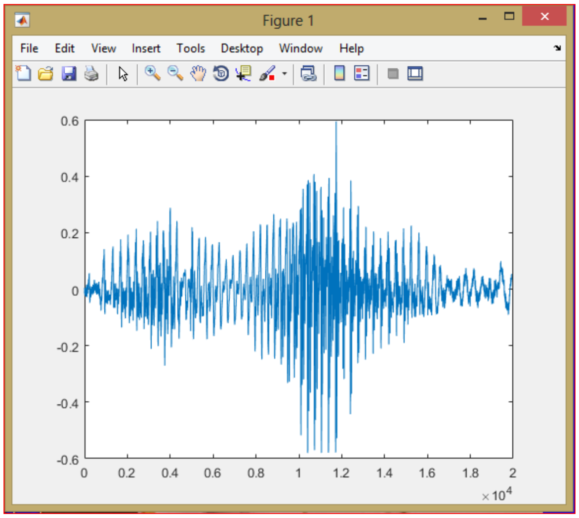


Fig 7. The original signal of the word "arba"

the experiment the proposed method provided better results than the previous work by providing 91.05% accuracy.

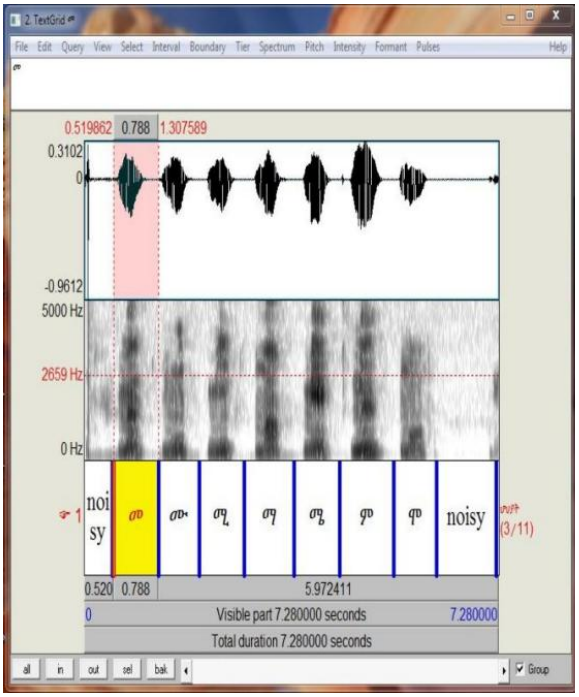
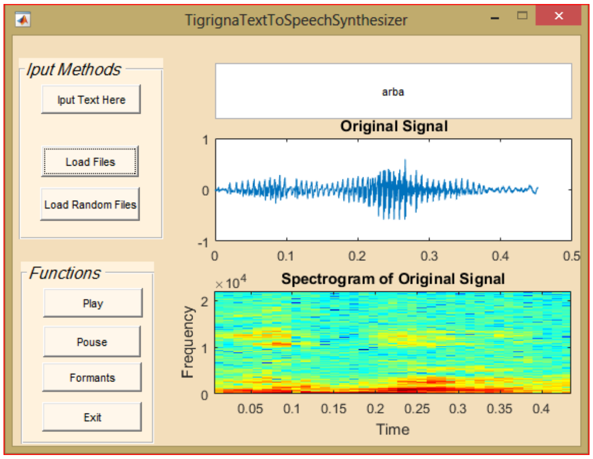

2. TextGrid
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
0.519862
0.788
1.307589
0.3102
0
-0.9612
5000 Hz
2659 Hz
0 Hz
noi sy
noisy
(3/11)
0.520
0.788
5.972411
0
Visible part 7.280000 seconds
7.280000
Total duration 7.280000 seconds
all
in
out
sel
bak
Group

Fig 8. The original signal of the word "arba" Figure 8. Segmentation of Recorded Speech using PRAAT



arba

Fig 9. GUI Text-to-Speech Synthesizer for Tigrigna language

5 Conclusion

In this article to conduct the experiments the datasets have been divided into two (training set and testing set) for testing and training purpose the experiments tested using 10 k fold cross validation. The result looks motivating and further improvement of intelligibility and naturalness depends on proper works in a different context. In this experiment when the modern deep learning model compared with the previous, the modern method shown better result than the traditional model. LSTM provides better Accuracy of 91.05% at epoch 100, Precision of 92.01% at epoch number 75, Recall of 95.69 % at 50, and Fscore of 94.02% at epoch 50 respectively.in conclusion modern LSTM method is a better deep learning method for enhancing accuracy than traditional machine learning method. Our future paintings include:(1) validate the overall performance of our proposed version on very larger datasets, (2) different textual capabilities and their fusion shall be analyzed to enhance the performance.

References

1) Weiss RJ, Skerry-Ryan RJ, Battenberg E, Mariooryad S, Kingma DP. Wave-Tacotron: Spectrogram-Free End-to-End Text-to-Speech Synthesis. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021;p. 5679–5683. doi:10.1109/ICASSP39728.2021.9413851.
2) Balyan A. An Overview on Resources for Development of Hindi Speech Synthesis System. *New Ideas Concerning Science and Technology Vol 11*. 2021;11:57–63. Available from: https://doi.org/10.9734/bpi/nicst/v11/5977D.
3) Joshi MM, Agarwal S, Shaikh S, Pitale P. Text to speech synthesis for Hindi language using festival framework. *International Research Journal of Engineering and Technology (IRJET)*. 2019;6(04):630–632. Available from: https://www.irjet.net/archives/V6/i4/IRJET-V6I4142.pdf.
4) Manoharan S. A Smart Image Processing Algorithm for Text Recognition, Information Extraction and Vocalization for the Visually Challenged. *Journal of Innovative Image Processing*. 2019;1(01):31–38. Available from: https://doi.org/10.36548/jiip.2019.1.004.
5) Herbert B, Wigley G, Ens B, Billinghurst M. Cognitive load considerations for Augmented Reality in network security training. *Computers & Graphics*. 2021. Available from: https://dx.doi.org/10.1016/j.cag.2021.09.001.
6) Tebbi H, Hamadouche M, Azzoune H. A new hybrid approach for speech synthesis: application to the Arabic language. *International Journal of Speech Technology*. 2019;22(3):629–637. Available from: https://dx.doi.org/10.1007/s10772-018-9499-4. doi:10.1007/s10772-018-9499-4.
7) Koc WWW, Chang YTT, Yu JYY, Ik TU. Text-to-Speech with Model Compression on Edge Devices. *2021 22nd Asia-Pacific Network Operations and Management Symposium (APNOMS)*. 2021;p. 114–119. doi:10.23919/APNOMS52696.2021.9562651.
8) Gujarathi PV, Patil SR. Gaussian Filter-Based Speech Segmentation Algorithm for Gujarati Language. In: Smart Computing Techniques and Applications. Springer Singapore. 2021;p. 747–756. Available from: https://doi.org/10.1007/978-981-16-1502-3_74.
9) Rajendran V, Kumar GB. A Robust Syllable Centric Pronunciation Model for Tamil Text To Speech Synthesizer. *IETE Journal of Research*. 2019;65(5):601–612. Available from: https://dx.doi.org/10.1080/03772063.2018.1452642. doi:10.1080/03772063.2018.1452642.
10) Daba E. Improving Afaan Oromo Question Answering System: Definition, List and Description Question Types for Non-factoid Questions . 2021. Available from: http://hdl.handle.net/123456789/6240.
11) Kim C, Gowda D, Lee D, Kim J, Kumar A, Kim S, et al. A Review of On-Device Fully Neural End-to-End Automatic Speech Recognition Algorithms. *2020 54th Asilomar Conference on Signals, Systems, and Computers*. 2020;p. 277–283. doi:10.1109/IEEECONF51394.2020.9443456.
12) Kodhai SDDE. Textaloud Assistant App Development for Multilanguage. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*. 2019;8(7s). Available from: https://www.ijitee.org/wp-content/uploads/papers/v8i7s/G10010587S19.pdf.
13) Kewley-Port D, Nearey TM. Speech synthesizer produced voices for disabled, including Stephen Hawking. *The Journal of the Acoustical Society of America*. 2020;148(1):R1–R2. Available from: https://dx.doi.org/10.1121/10.0001490. doi:10.1121/10.0001490.
14) Nadig PPS, Pooja G, Kavya D, Chaithra R, Radhika AD. Survey on text-to-speech Kannada using Neural Networks. *International Journal of Advance Research*. 2019;5(6):128–128. Available from: https://www.ijariit.com/manuscripts/v5i6/V5I6-1159.pdf.
15) Narvani V, Arolkar H. Information and Communication Technology for Competitive Strategies (ICTCS 2020). *Lecture Notes in Networks and Systems*. 2021;190. Available from: https://doi.org/10.1007/978-981-16-0882-7_84.
16) Madhfar MAH, Qamar AM. Effective Deep Learning Models for Automatic Diacritization of Arabic Text. *IEEE Access*. 2021;9:273–288. Available from: https://dx.doi.org/10.1109/access.2020.3041676. doi:10.1109/access.2020.3041676.
17) Tanberk S, Dagli V, Gurkan MK. Deep Learning for Videoconferencing: A Brief Examination of Speech to Text and Speech Synthesis. *2021 6th International Conference on Computer Science and Engineering (UBMK)*. 2021;p. 506–511. doi:10.1109/UBMK52708.2021.9558954.
18) Mahar SA. Prosody Generation Using Back Propagation Neural Networks for Sindhi Speech Processing Applications. *Indian Journal of Science and Technology*. 2020;13(2):218–228. Available from: https://dx.doi.org/10.17485/ijst/2020/v13i02/149356. doi:10.17485/ijst/2020/v13i02/149356.