

Text To Speech Synthesis for Afaan Oromoo Language Using Deep Learning Approach

SORESSA BEYENE
Ambo University, Hachalu Hundessa Campus Institute of Technology Department of Computer Science

Raja.K (PhD)
Ambo University, Hachalu Hundessa Campus Institute of Technology Department of Computer Science

Abstract
Text to speech synthesis (TTS) which generate input texts is generate to the speech from texts. TTS is very important in aiding impaired people, in teaching and learning process. But, to implemented TTS have a lot of challenging such as text processing, time to phoneme mapping and acoustic modeling for Afaan Oromoo language. So, Afaan Oromoo language mostly required to text to speech synthesis for development of this language. The application of Natural Language Processing is provide that input texts pair speech to generate the desired result outputs of speech in waveforms from prepared text corpus. The normalized text was used for linguistic features are extracted by using Festival toolkit for Afaan Oromoo TTS. The labeled texts are done using Festival toolkit, and generated the utterances of texts from scheme file parameters. The Festival toolkit is used for texts normalized in linguistic extraction from label phoneme alignment to match with speech corpus in trains and tests. The forced alignment is done by HTK toolkit for prepared environment, checked data extracting features within timestamps of state level alignment for acoustic feature extracted. So, this study focus on TTS approach deep learning model based on BLSTM-RNN for Afaan Oromoo language. The RNN model used from a given input feature sequence to extracted duration model and acoustic model. The implementation is done in BLSTM-based on RNN using pytorch library on jupyter notebook, create duration model and generated speech samples from trained acoustic model. We have prepared 1000 texts corpus their matching text transcription from Afaan Oromoo speech corpus by a female speaker dependent for training 700 sentences and tests 300 sentences from dataset domains. In this study, two evaluation techniques used. Frist, the Mean Opinion Score (MOS) evaluation technique is used for intelligibility and naturalness in TTS. The second is Mel Cepstral Distortion (MCD) which is highly used for objective evaluation in model approach for TTS. So, the performance of this model was measured and quality of synthesized speech is assessed in terms of intelligibility and naturalness which results are 3.77 and 3.76 respectively. The total average processed using objective evaluation technique the speech corpus on 16 kHz standards is generated by MCD BLSTM-based on RNN is 3.89 and merlin wave generated is 3.71 correspondingly.
Keywords: Text To Speech Synthesis, Mel Cepstral Distortion (MCD), Mean Opinion Square (MOS), Bidirectional Long Short Term Memory Recurrent Neural Network (BLSTM-RNN)
DOI: 10.7176/NMMC/101-02
Publication date: April 30th 2022

1. Introduction

Text-to-speech (TTS) means input texts is to generate the audio and used for in communication, the sound hear to human. Natural language processing (NLP) is a field which employs computational techniques for learning, understanding and producing human language properties at the intersection of computer science, artificial intelligence and computational linguistics. It is used for both generating human readable information from computer system and converting human language into more formal structures that a computer can understand. The Text-To-Speech (TTS) synthesizer is a computer-based system that able to read any text aloud, whether it is directly introduced in the computer. Among 83 languages which are registered in the country Afaan Oromo is a Cushitic language that has the greatest number of speakers Ethiopia. Moreover, Afaan Oromoo has 60 million speaker as a mother thong and as second language. The speech is formed from phonemes and combined together to form words in Natural Processing Language. The Natural Language Processing study human language learns with the natural sound and they speak to communication throughout their life. Humans also learns easy and efficient mode of communication with machines. So, the Natural Language Processing accepts the input texts pair speech corpus and able to generated speech output after text analysis method. This text analysis contains text normalization, sentence segmentation, tokenization and non-standard words like abbreviations into full word covert (Trilla, 2009).

The method used to develop a text to speech system in concatenative synthesis is based on speech signal processing of natural speech databases and speech signal processing able to perform speech in waveform. In such a manner that, appropriate speech units are concatenated to construct the required speech. The segmental database is built to show the main phoneme extract features of a language from the concatenative synthesis

recorded audio. The method used to set the phonemes is built diphones units, representing the phoneme to phoneme junctures. No uniform units are also used diphones, syllables and words in this concatenative methods for speech synthesis.

The aims to generate a mapping between the textual diphones and their equivalent speech units. Each diphone is represented by two characters, consequently producing speech unit of that diphone. Syllables are any words with the exception of abbreviations and acronyms were considered in system written words considering rule based constrictive depend on language in consonant and vowels.

This syllables is used for small database select unit in TTS systems. The words Systems that simply concatenate isolated words or parts of sentences, are only applicable when a limited vocabulary is required typically a few words and the sentences to be pronounced respect a very restricted structure. The synthesizer of the speech segments, and performs some signal processing to smooth unit transitions and to match predefined prosodic schemes. The direct pitch synchronous waveform processing is one of the most simple and popular synthesis algorithms.

The multi-pulse excitation linear predictive Coding system produces synthetic speech that is more natural sounding than the classical linear predictive coder. In the multi-pulse excitation linear predictive coding system, the excitation signal is modeled with a few pulses per frame of speech regardless of whether the frame is voiced or unvoiced. The quality of the synthesized speech improves with the number of pulses used per frame. Pulses are computed by minimizing the weighted square error between the original speech and the synthetic speech.

The Digital Signal Processing (DSP) turns NLP representation into an output signal (Tilahun, 1993). The advantage of DSP module are the controlling the duration time and frequency (aperiodicity) of the vocal folds so that the output signal matches the input requirements show speech signal processing.

Generating (converting) text to speech encompasses both natural language processing and digital signal processing (Morka.M, 2003). The application of Natural Language Processing (NLP) is to produce a phoneme translation of the text reading with the required intonation and rhythm (Alula, 2010). Text analysis is the responsible for analysis of input text into soundable texts. To achieve this, it organizes the input texts into control the lists of words and proposes all possible part of speech categories for each word taken individually on the basis of words, and then considers words in their context of the recorded and written. Phonetic analysis is purpose for the finding of the phonetic translation of the incoming text. This work can be organize in different ways dictionary based and rule-based strategies (context-based).

The speech is greatly affected by accents occur at stressed syllables and form characteristic words in the pitch tones. The transition periods between syllables place of produce and found to be dependent on the nature of articulation of boundary sound units. The component are responsible for generate the acoustic sequence required to synthesize the input text by finding the pronunciation of individual words in the input text. The style of pronunciation was influenced by the gender, physical state, and emotional state and focused on the speakers. The prosody features depend on many aspects like the speaker characteristics gender, emotions and meaning of the sentence (Samuel, 2007).

Speech is the most efficient and natural way to communicate with each other. Speech is the agreement and common understood of communication between human being. When human read text as the rule based of the phonology, with native language (mother language) speech, the person hears the individual words and sounds. Every speech not converted to standard written words or texts. So, speech can be written using letter to sound format the words. But, this is not true if the person hearing the speech is not familiar with the language.

The conversation of text to speech are several method .The development of society and economic system since prehistory time has been paralleled by a growth in man's dependence upon technology. Speech enabled interfaces are desirable because they promise hands-free, natural, and ubiquitous access to the interacting device (Solomon, 2005). However, it is one of the least supported and least researched languages in the world. He remarkable works some contributed doing on text to speech synthesis for Afaan Oromo languages (Solomon, 2005).

Amharic Text-To-Speech Speech Synthesis System stated phonetic once analysis done, the final block of NLP to prosody generation, which is responsible for finding correct intonation, stress, and duration from written text in prosodic features. The prosodic features are features that appear when to input sounds together in connected speech (Alula, 2010). It is advanced in prosodic features as successful communication depends on intonation, stress and rhythm as on the correct pronunciation of sounds. Intonation, stress and rhythm are prosodic features. The rule-based methods use manually produced rules, extracted from utterances structures.

Afaan Oromoo speech synthesis system was developed on a hidden Markov model method (Wosho, 2020). The HMM model stated for the neighbor rule and able to processes limited datasets. The researcher not stated and mentioned acoustic feature and linguistic feature in statistical parameter used for text to speech synthesis based on Hidden Markov Model (HMM).

In NLP several methods have been used in the phone duration model working like linear regression models are based on the assumption that among the features which affect the segmental duration there is linear

independency. The models used to predictions linear regression with small amount of training data not model the dependency among the features extracted. On the other hand, decision tree models and in particular classification and regression tree models, which are based on binary splitting of the feature space, can represent the dependencies among the features. The phone duration model, where the segment duration prediction was based on a sum of factors and their product terms that affect the noise duration (Yang, 2014).It can effectively extracted the hidden internal structures of data and use more powerful modeling capabilities to characterize the data.

Deep learning is a part of machine learning which trains the model with large datasets using multiple layers and Feedforward neural networks for single layer. Deep learning that is capable of process in short time duration large dataset of training become important method for text to speech system. The HMM based for speech synthesis method maps linguistic features into probability densities of speech parameters with various decision trees. The deep learning based method directly perform mapping from linguistic features to acoustic features with BLSTM-RNN which have proven extra ordinarily efficient at learning inherent features of data. It is important for readers better understand the development process of these methods used deep learning approach. Deep learning based models approach significant progresses like handwriting recognition machine translation (Sutskever, Vinyals, & Le, 2014). The speech recognition (Graves A. Mohamed, 2013) and speech synthesis (Zen & Alan, 2009).

Recurrent Neural Networks (RNNs) is the also the family of deep learning that are well-suited for pattern classification tasks whose inputs and outputs are sequences, for example tasks such as speech recognition, speech synthesis, named-entity recognition, language modelling, and machine translation (M. S. Al-Radhi, 2017). The Recurrent Neural Network (RNN) method to model speech-like sequential data that represents associations among bordering frames training duration model and acoustic model. It can also practice all the accessible input features to forecast output features at each frame. Particularly, the RNN model is different from the DNN since it operates not only on inputs but also on network internal states that are updated as a function of the entire input history. Training RNN incorporates backpropagation.

The RNN connections are able to mapping the utterance and understanding input datasets for train the acoustic sequence, which is purpose waveforms to show speech in signal processing to generated prediction outputs desired (M. S. Al-Radhi, 2017).

Long short-term memory networks (LSTM) are a class of recurrent networks composed of units with a certain structure to manage better with the vanishing gradient problems during training of recurrent neural network and maintain potential long-distance dependencies (M. S. Al-Radhi, 2017). This focused on linguistics adapted with technology. The Text to Speech is soundable communicate information to the user, where digital audio recordings, for developing a user of speech synthesizes in Natural Language Processes. The performance of evaluation used intelligibility and naturalness encourage to investigate the text to speech synthesizer in Afaan Oromo language. So, this research training datasets are extract the linguistic features for Afaan Oromoo language.

2. Related Work

In this chapter, from presented the review of number of speech synthesizer developed focusing on different approaches. The deep learning approach is one of advanced in Natural Language Process for text to speech synthesis. Deep learning approach is extract the hidden internal structures of data and use more powerful modeling capabilities to characterize the data. Therefore, concerning to Afaan Oromoo language from the previous work done (literature reviewed), text to speech synthesis for Afaan Oromoo language has not still methodically exploring using deep learning approach. The researcher used to the deep learning approach with the BLSTM-based on RNN, text to speech synthesis for Afaan Oromoo language and consideration the model used to synthesize the desire fully context labels (speech, texts) pairs which are phone mapping, duration(linguistic) modeling, acoustic modeling, generated speech for Afaan Oromoo language Table 1 showed below.

Table 1: Table 1:Summary of Related Work on Text to Speech

Author	Title	method	Number of datasets	Result	Limitation
(Morka.M, 2003).	Text-To-Speech System for Afaan Oromo Language	Rhyme Test, approximately	for test 15 words	For person one (Type 1- people) I 43.33% and person 2 83.33 % of the test data	Isolated word utterances are very slow as compared a continuous speech
(Samson, 2011).	Text-to-Speech System for Afaan Oromoo	Concatenative within Residual Excited Linear Predictive Coding (RELPC)	unspecified	Diphone of 75% and triphone 54% respectively	Prosodic method not include. model was unidentified
(Tewodros, 2009)	A TTS synthesizer for Wolaytta language	Festival tools for generated nonsense Praat for record audio	1369 1369 nonsense and the recorded words 841	Performance of the system 78.3%.	The Speaker specific intonation and speaker specific duration not considered Letter-to-Sound Rule
(Alula, 2010)	Amharic TTS system	Hidden Markov Model toolkit	80 standard words and 30 non-standard words (NSWs)	Standard evaluation 76.7% and non-standard 70%, respectively	Lack of prosody analysis building part of speech is crucial
(Kedir, 2020).	Text to Speech Synthesis for Afaan Oromoo	Statistical parameter based on Hidden Markov Model	400 sentences are used for training and 10 sentences for testing	In MOS evaluated intelligibility is 4.3 and naturalness 4.1 of the speech synthesized respectively.	Audio conversation mechanism, spoken language non-standard words like time, acronym are not considered
(Alem F., 2007)	Bangla text-to-speech system	Deep learning based on deep neural networks	1,35000 words	Training 94%, testing 3% and validation 3%.	
(Zen H. S., 2013)	Text-to-Speech Synthesis for English	Bidirectional LSTM based on Recurrent Neural Networks	13,100 speech	Not specified	Not effected generate Acoustic model

3. Research Methodology

This section deals with the Afaan Oromoo text to speech synthesis system methodology within architecture. It explains the whole of design the representation and description of components. The design architecture easily understand the method of a text to speech synthesis for Afaan Oromoo approach to deep learning. The training phase a text corpus is used passes through the text analysis process (tokenization, normalization and linguistic features extraction).The extracted features are used as an input for the duration model, from the speech corpus acoustic features are extracted. The input is used for the acoustic model with the linguistic features and duration information generated by the duration model. Output of the acoustic model is used as input for the Vocoder to generate the final speech. The generation of speech duration model and acoustic features are extract. The extracted features are then used as an input for duration model and acoustic model training. Finally, the speech synthesis is evaluated. So, the architecture of text to speech using BLSTM-based on RNN for the Afaan Oromoo language is illustrated in Figure 1.

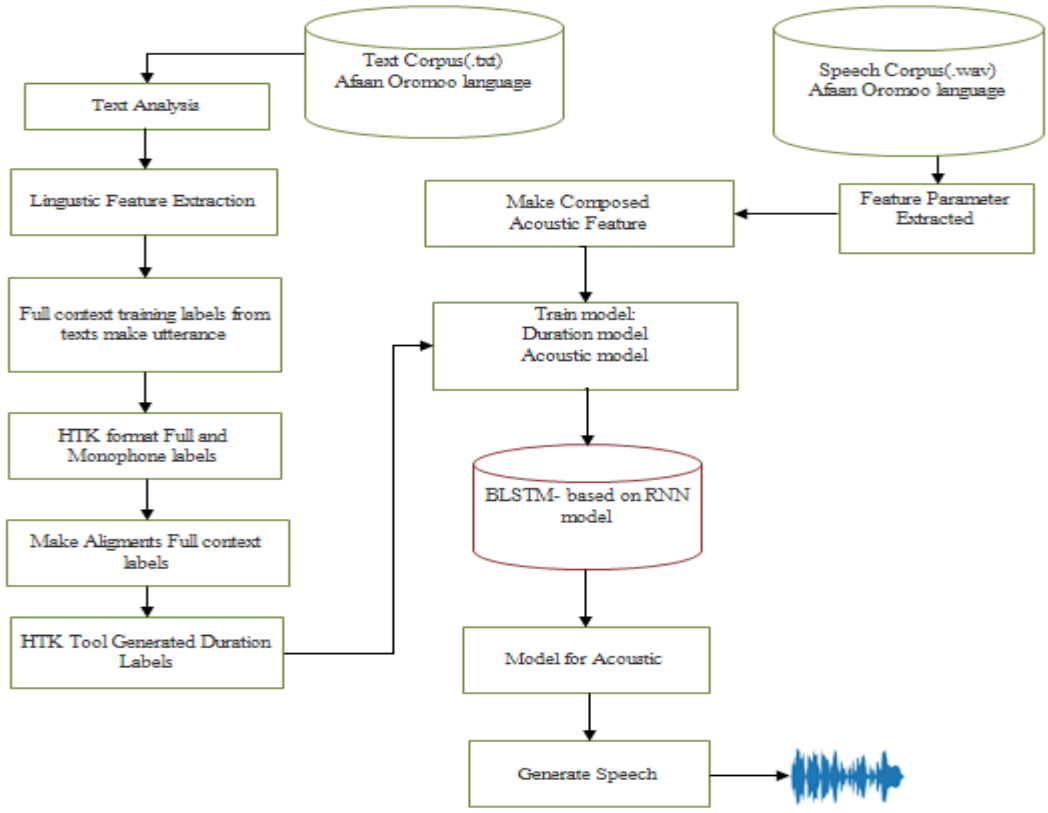

Figure 1: proposed model for Afaan Oromoo Text to Speech

3.1 Data Collection and Preparation

To collect datasets, for achieving the general objective. In is very challenge in text to speech for Afaan Oromoo not archived and no standard datasets. Therefore, some text accesses the raw data source from Afaan Oromoo Wikipedia book, serlugaa Afaan Oromoo book and Oromia broadcast network (OBN) of Afaan Oromoo language from their websites. Thus, the data collection process represents one of the significant stages of this study. The acoustic data has its transcription which allows the training of models and duration data separately from total dataset text corpus and speech corpus total 1000 are label to phoneme alignment. Text corpus consists of total 1000 sentences, 12359 total words and 2043 unique words. The process of text and speech corpus mapping is performed by applying the forced alignment tool used htk toolkit for linguistic and an acoustic model. So, this study focused on thesis scopes to collect datasets (texts and speech corpus) for Afaan Oromoo language. The speech sample is recorded at 16000 Hz mono sound standards and the files are stored in waveform format (.wav). The Festvox system and full style label use to htk toolkit to make it easier to create raw text pair speech signal to suitable machine learning and easily understanding format. The Praat is free open source software that used to record the each text into speech in wave form. For reduce noise using microphone and a normal office computer made up the hardware equipment record the texts for preparation speech corpus in wave format.

The corresponding text files are prepare by manual ways from data sources due to benchmark results dataset is not available for Afaan Oromoo regard to text to speech. The database and experiments contains an audio record by a female speaker dependent automatically parameter extracted was used. These dataset preparation is focused on linguistic and acoustic feature train for text to speech conversion. The linguistic and acoustic features are focused the on original texts translation and recorded audio respectively.

The next steps is to record the speech corpus, the text parallel recording speech is done by a native and professional in linguistics female 25 year old speaker dependent in a quiet room to reduce disturbance. For high-quality recording microphone with noise isolation facility used the personal computer. The recording process is carried out used praat free software which provides tools for sound file recording, end-pointing, rate manipulation, and noise reduction.

3.2 Text Analysis

The text analysis module is used to do tokenization and text normalization and then the normalized text is used

to extract linguistic features and the extracted features are used as an input for the duration model. The durations for each phone are first is predicted using the duration model using the pre-trained model. Then after, the duration model is used to predict the timestamps of each phoneme mapping duration of each phone to be synthesized. The abbreviations and acronyms are not pronounced as they are input or written. The first work for the raw input abbreviation to replaces the abbreviations and acronyms in their expanded as rule in Afaan Oromoo language. The python scripting language used for performs this normalization with sample Afaan Oromoo abbreviations is common known when expanded. Text corpus preparation for linguistic feature extraction for each of wave (.wav) files, it requires text (.txt) files with exactly the same name that contains exactly the text that was recorded in the speech corpus. Text to speech (TTS) synthesis is conversion of text into speech.TTS system consists in text analysis, where the input text is transcribed into linguistic representation. In this TTS part, the input sentence is segmented into tokenize.

The Utterance structures are central when synthesizing speech used the Festival speech synthesis system for text transcribed into phonetics. The relevant properties of the utterance to be synthesized in a way they are specifications of desired utterances. Instead of created an utterance structure by synthesizing some text, we can create an utterance structure from a speaker's natural utterance, taking all properties from the speaker's utterance instead of predicting them from texts. We have an utterance that looks like a synthesized utterance, but which is guaranteed to be a valid natural utterance, with correct phonetic properties.

An utterance structure consists of many items like single words, syllables, phones, phrases and items are connected through several relations. Exactly which relations are present depends on the synthesis method, among other things. But some of them are always present, such as the word relation connecting word items, the syllable relation connecting syllables, the segment relation connecting phones, and the syllables structure relation connecting items.

The requirements for generated utterances structures by the Festival system and provided a script to create utterance structures from text corpus, linking the hierarchical relations by their time information (the script looks at the end times of items in lower relations to decide which items in a higher relation must be their parent). However, these label files need and for punctuation marks. The punctuation is represented as a feature at token level by Festival into utterances. The aligner did not differentiate between words and tokens, instead of word label files created by the aligner are actually tokens in Festival. So in order to have punctuation at the token level in Festival system, we have that information in the word label files. We have able to make use of the features and the scripts for converting to utterances created one utterance per word label file. Unfortunately, the librivox recordings are usually quite long, and it is not possible to cut them into reasonably small pieces for Festival system before running the aligner, because then you would have to sit down and segment the text(.txt) files accordingly. Speech corpus files move all short wave files into the wav directory within /home/soro/merlin/Text2Speech/AO_Speech_Syntesis/soro and phonemic files and the words files to this directory, so everything is at the same place. The raw audio preparation for training is used to htk toolkit audio make feature acoustic extraction Mel generation coefficient (mgc), fundamental log frequency (lf0), and band periodicities (bap).

The made the file was the extracted acoustic feature from composed acoustic features. These made train was generated in scripts. The full context train labeled extracted from text corpus for utterance build. The made file was generated used HTK tools for full context label like monophone mapped and full labeled texts pair speech. The list of made file is generated under folder master label files and model list files. The acoustic features accesses is extracted audio waveform from make features which includes log fundamental frequency (fl0) represented the pitch and Mel generalized cepstral features which displayed spectral parameter of the speech.

3.2.1 Text pre-processing

Text pre-processing is task must be expanded into full words digits and numerals. Another task is to find correct pronunciation for different contexts in the text. The synthesis speech in text analysis of the raw input text into pronounceable words. From the texts contains string punctuation marks to clean like!"#$ %&()*+,-./:;?@[\]^_{|} ~\t\n providing many functionalities in deep learning.

The construct datasets from pertaining linguistic duration and acoustic features because computing features from the label files on-demand are performance, particularly for duration model extract features used to python script and using bash in form (./filename.sh) as following steps:-

Step1. Prepare corpus similar text within speech corpus in this same file name in text and audio filename separately in folders. The purpose for preprocessing text pair speech easily undestanded for machine learning in linguistic (text) analysis.

Step2. From saved file text pair speech automatically created label phoneme alignment and state alignment using method htk tool and festival fronted tools

Step3. Create the duration model and acoustic model

step4. Create training duration model and training acoustic model

step5. Speech synthesis from feature prepared
step6. Test speech synthesized duration and acoustic using the name of voice created.

3.2.2 Normalization

The first iteration in the loop again gets the basename. The second eliminates pauses and the header from the aligner word label file. The third extracts the lines of pause labels from the word label file and stores them in a file with extension pauses. These lines are important because they contain the label times for the pauses, which we don't want to lose. The words with punctuation file by writing a hash sign for the end of the header into it (overwriting a possibly existing file of that name). **T**he speech preprocess was normalization speech record in wave format. The displayed the graph used to matplotlib and librosa display the wave plot attribute for sample rate when reducing noises.

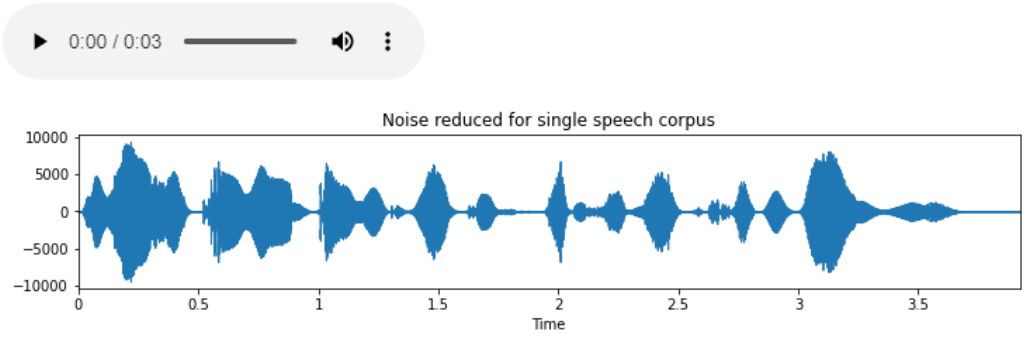

Figure 2: Noise reduced for speech corpus

The Figure 11 displayed linguistic features output the text translation, which typical desired additional resources of pronunciation of the language.

The TTS that translate raw text converted to linguistic used festival tool fronted and WORLD tools the phone alimented for all datasets after label the files text pair speech corpus. So, the linguistic representation text that is generated from festival tool outputs. The HTK scripts transform with the contextual labeled information and structured Festival represent the list of phoneme mapped in utterance information full context labeled style. To create the directory name of folder to copy into template desired the datasets speech pair texts.

3.3 Linguistic Feature Extraction

The forced alignment in state align is created the training labels used to htk tools and phone align created the training labels with Festvox tools. To check how your sentence was synthesized. We have synthesized an utterance and stored it in a variable utterance used to Festival system (set! utt (Say Text "iji waaqayyoo iddoo hundumaa jira, isa hamaa fi isa gaarii ni arga**. "**)) checked which relations are present by (utt.relation names utt) generated the following relations token , word , phrase ,syllable , segment, syllable structure , intonation , target unit source segments and wave.

To train a speech corpus in wave files converted into phone align labels for mapped phone and timestamps for duration of the texts in corpus when recorded. So, choice one convenience and set the option accordingly in global settings configuration file.

The training process for both synthesis systems requires that each sentence of the is obtained from phone align created training labels transcription was match with texts pair speech corpus. The global settings configuration file helped for feature extract in duration mode depends on source kind from waveform, source format wave files, and target rate at 50000 for timestamps and target kind for Mel frequency cepstral coefficients.

The linguistic features alongside timestamps are then used because the input for the BLSTM acoustic model to get corresponding compressed acoustic features, which include parameters Mel generated coefficient (mgc), fundamental logarithm frequency (log F0) and band (BAP) are generated form linguistic source files from label phone aligns text files label (.lab).

Features required to train duration model assign the variable X as linguistic and Y as duration. The variable X as duration source find the linguistic source, add frame features used phone alignment from question paths file dataset sources the declared variables X duration source and Y duration. We saved the features for duration mode duration linguistic feature dimension in variable X duration and duration feature dimension variable Y duration. For the variable x, y to enumerate in zipped code, name split text basename the collected files in variable X duration join with automatically make directory x and y path converted into binary formatted(.bin) this processed used for normalization duration model produce linguistic features.


```
mkdirs: ./data/  speech/X_duration
mkdirs: ./data/  speech/Y_duration
mkdirs: ./data/  speech/X_acoustic
mkdirs: ./data/  speech/Y_acoustic
Duration linguistic feature dim (38, 416)
Duration feature dim (38, 1)
151it [00:00, 5701.14it/s]
Acoustic linguistic feature dim (768, 420)
Acoustic feature dim (474, 190)
151it [00:19,  7.82it/s]
```

Figure 3: Prepare Features for Linguistic Feature Training

3.3.1 Utterance Length by Histogram Visualization

Then, the duration features are taken as an input for the Vocoder to generate speech in waveform.

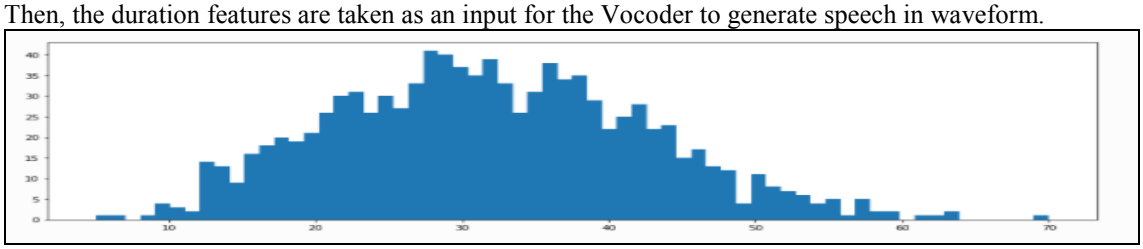

Figure 4: Utterance Lengths Histogram Train and Text Visualization

The Figure 4 is show the final step to synthesized speech waveform from the recover spectrum and can listen to the sound and pysptk speech representation within waveform using on jupyter notebook using librosa audio visualization the sentences show the plotted different duration records**.** The utterance lengths histogram as total number of utterances 144 and total number of frames 9153 sample.

The feature normalization in this linguistic feature extracted is to show the utterance training feature from constructed duration and acoustic datasets in precomputed the input variable within the focus (target) in acoustic datasets. The features load on demand to epoch the utterance length from total number of duration and acoustic training. The linguistic feature is variant in every scale of at everyone dimension. This linguistic clear to visualization when the normalization the feature of linguistic and variance while applying the normalization duration and acoustic features is computed. The linguistic features become to normalization at any one indexed the feature domain between 0.01 and 0.99 values.

3.4 Acoustic Feature Extraction

The process for training the acoustic to extracted feature model was no more difference that for train duration model. Since, configuration files which used to initialize and BLSTM based on RNN model. The utterance lengths histogram as total number of utterances 144 and frame is 128033 sample utterance acoustic train show as Figure 15 below.

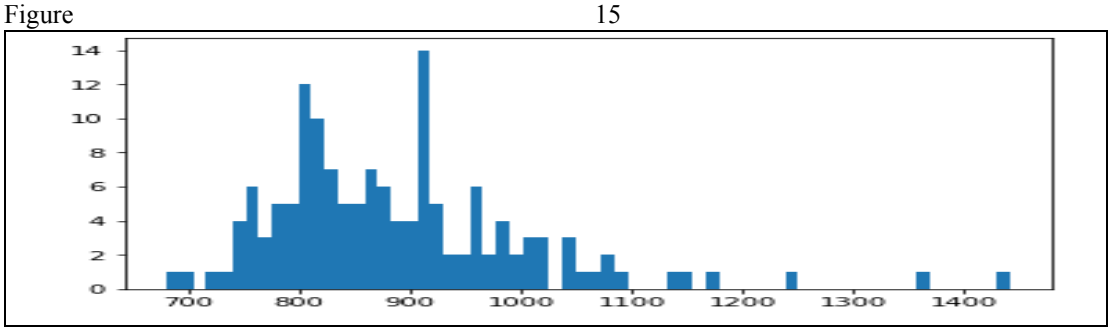

Figure 5: Utterance Lengths Histogram Acoustic Train

The Figure 5 shows, the output total number of utterances lengths acoustic and train. The output visualization histogram in utterances acoustic within binary at 64 computing the total number of frames used to summation in attributes of numpy acoustic training.

3.4.1 Spectrogram

The first spectrograms to generation of speech synthesis to analysis is used python speech parameter toolkit (pysptk) and important used the librosa for representation features. The pysptk was contains have sequential include windowing, mel-generalized cestrum analysis, visualize spectral envelope estimates and F0 estimation. The spectral parameter estimation and visualize its spectral envelop estimate. The first option for an audio features representation is spectrogram. The spectrogram two dimensional tensor is displayed vertical dimension indexes times and horizontal dimension frequency.

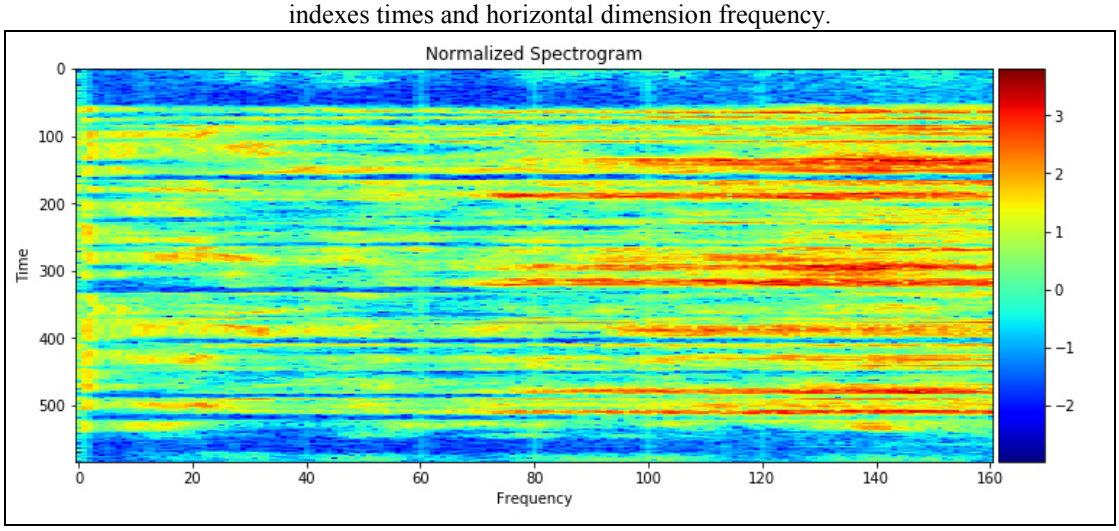

Figure 6: Shape of Spectrogram Feature

Figure 6 shows acoustic features that are the acoustic properties of speech signals for speech analysis. When comparing a spectrogram to a normalized spectrogram, the resolution between the frequencies that matter for speech is higher and there is less redundant information. It emphasizes details in lower frequencies that are critical for speech modeling and de-emphasizes the high frequencies.

3.4.2 Mel-Frequency Cepstral Coefficients (MFCC)

The second of the audio feature representation is MFCCS. The aim of the behind MFCCs features same as spectrogram feature at time window. The MFCCs feature representation feature vectors that characteristics sound within the window. The remember that MFCCs feature is more lower dimension the spectrogram features, which help as acoustic model to avoid overfitting to the training dataset.

The visualization audio feature representation entire in the tensor assume values close to zeros used to MFCC shape. However, deeply get the details of how MFCCs are calculated. This spectrogram displayed using the python speech features python package. The generated spectrogram features from the MFCCs are normalized in was representation. This focused on MFCC features is the same as spectrogram features at each time window and the MFCC feature yields a feature vector that characterizes the sound within the window. So, that the MFCC feature, which helped an acoustic model to avoid overfitting to the training dataset shape use torch library sizes.

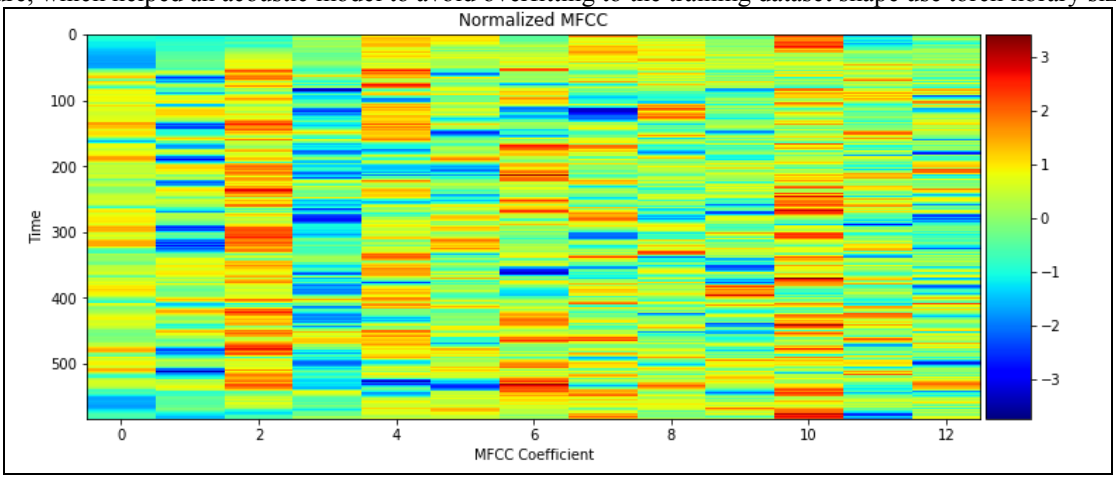

Figure 7: Audio Feature Representation in MFCCS

The Figure 7 show the results of the data in sampling rate in Mel frequency cepstral coefficients at 13 plot in figure size (14, 4) is used to librosa display speech show the MFCC within time rates.

Based on their semantic interpretation audio features are classified into physical and perceptual features. Perceptual features are basic features that are perceived by human listeners such as loudness, pitch, rhythm, and

timber. In contrast, physical features are properties that present mathematical, statistical, and physical properties of audio signals. The most commonly used acoustic features in speech synthesis are the Mel Frequency Cepstral Coefficients. The acoustic features are often used as a low level audio representation to bridge the synthesizer and the Vocoder in the backend of a TTS system. In this study, we used Mel Frequency Cepstral Coefficients and linear scale frequency spectrograms as an intermediate acoustic feature representation. To be specific, the output of the feature predictor and the conditional input of the Vocoder are sequences of Mel and linear frequency spectrograms. The frame period at five and the trimming zero frames of the spectrogram is small power to make good visualize using pyworld (python world).

3.5 BLSTM based on RNNs Model

The defined model used to bidirectional long short term memory based on recurrent neural network variable declare RNN models for duration and acoustic within bidirectional long short term memory activation using pytorch for simple implementation.

```
models = {}
for ty in ["duration", "acoustic"]:
    models[ty] = MyRNN(X[ty]["train"][0].shape[-1],
                           hidden_size, Y[typ]["train"][0].shape[-1],
                           num_hidden_layers, bidirectional=True)
    print("Model for {}\n".format(ty), models[ty])
```

```
('Model for duration\n', MyRNN(
  (lstm): LSTM(416, 256, num_layers=5, batch_first=True, bidirectional=True)
  (hidden2out): Linear(in_features=512, out_features=1, bias=True)
))
('Model for acoustic\n', MyRNN(
  (lstm): LSTM(420, 256, num_layers=5, batch_first=True, bidirectional=True)
  (hidden2out): Linear(in_features=512, out_features=1, bias=True)
))
```

Figure 8: BLSTM based on RNNs Model

3.5.1 Training Duration Model

Specifically, to create the directory name database for the processed dataset and experiment using on Linux operating system. The experiment directory include the label texts which contains alignments of phones generated from linguistic (duration) and the states alignment generated from speech corpus. Additionally, the experiment directory contains the extracted speech record features in folder linguistic (duration) and acoustic model. For a little soundness check from the two alignments in labels (time to phoneme) files for the text pair speech have different lines of pronounced in each characters in words. In everyone line of files label represents a single alignments in time to phonemes. The desired for a given sound files in its state alignment have more lines than its phoneme label files. In this one phoneme made up of multiple states in the machinery use for generate alignments. To look details the acoustic model used to do forced alignment to generated the labels, but the difference between numbers of lines in the alignments files are around the expected numbers of the text files '**cafeen teesse'** which label name A0_00202.lab (A= Afaan Oromoo from listed ID generated) have total phonemes twenty (20) and two hundred state level in acoustic (audio) file. This gives to know how many acoustic in label state alignment have phonemes is two hundred divided for twenty (200/200) result ten (10) states per phoneme. The model duration training was nicely. It kept decreasing training and test loss over time. But there is one thing the test loss was stopped decreasing and started to increase. This means that the network had started to overfitted to the training set and regularization techniques such dropout and obtain extract of the duration model displayed the finished train datasets text pair speech within number of epoch 25**.**

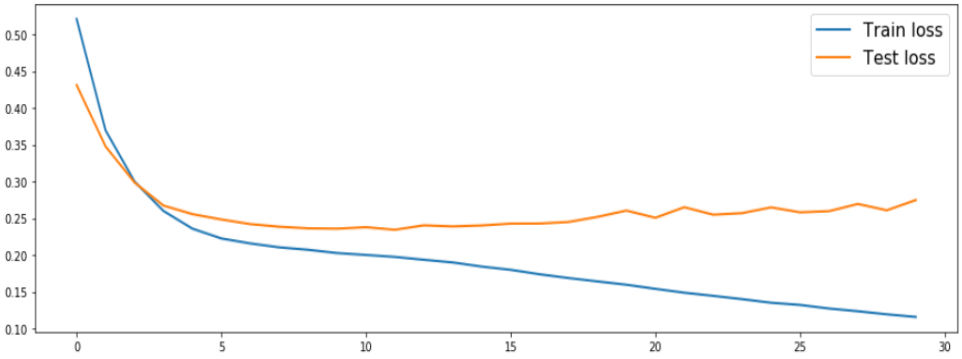

Figure 9: Train Duration Model

From this Figure 9, the number of epoch or iteration increase it displaying different training and test loss in datasets. So, good result displayed.

Table 2: Duration Model using Phone Align and State Align

Labels phone align				Labels state align			
Learning Rate	Valid in RMSE	Correction	Test in RMSE	Learning Rate	Valid in RMSE	Correction	Test in RMSE
0.002	6.777 frames/phone me	CORR 0.633	7.665 frames/phone me	0.002	6.826 frames/phone me	0.624	7.840 frames/phone me

Table 2 describe duration model label phone alignment and state level alignment to demonstration speech pair text train data, test data for train, valid, and test respectively evaluation used to RMSE within learning rate at 0.002 in all data sources 63.3 % in label phone align was one of preferred.

3.5.2 Training Acoustic Model

The train acoustic model is somewhat good to decreasing training and validation loss over time. But there is one thing. The validation loss to be stopped decreasing and started to increase showed. This means that the network had started to overfitted to the training set and regularization techniques such dropout and obtain predictions of the acoustic model displayed finished train datasets texts pair speech at number of epoch 25.

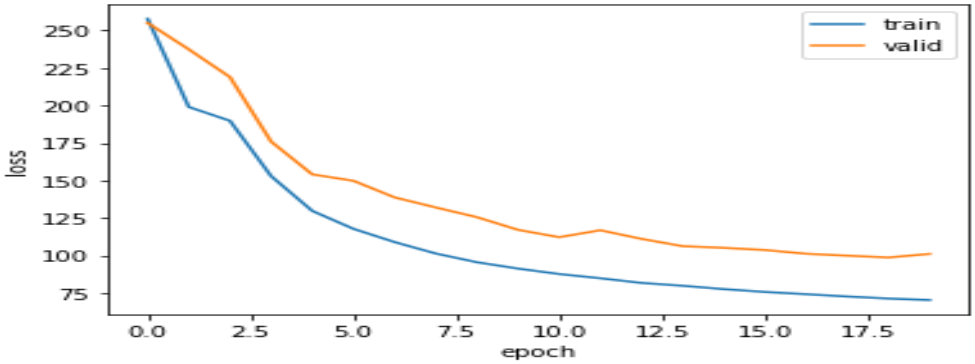

Figure 10: Train Acoustic Model

The challenge from Figure 10 when, we repeat execute programming for training the shape of diagram train and loss are changed. The processing train and validation are parallel increase to control the overfitting models.

Table 3: Acoustic Model Phone Alignment

Labels phone align valid					
Learning Rate	Valid MCD	BAP	F0 in RMSE	CORR	VUV
0.002	6.762 dB	0.246 dB	19.433 Hz	0.538	11.403%
Labels phone test					
Learning Rate	Test MCD	BAP	F0 in RMSE	CORR	VUV
0.002	6.704 dB	0.262 dB	15.264 Hz	0.700	8.907%

The Table 3 describe acoustic model label phone alignment average voice model single female audio records from A0_00202 to A0_01000 within 9153 utterances and trained datasets evaluation results. From this label state align test correct or accuracy 70% at learning rate was 0.002.

Table 4: Acoustic Model State Alignment

Labels state align valid					
Learning Rate	Valid MCD	BAP	F0 in RMSE	CORR	VUV
0.002	6.559 dB	0.242 dB	19.573 Hz	0.529	11.655%
Labels state align test					
Learning Rate	Test MCD	BAP	F0 in RMSE	CORR	VUV
0.002	6.586 dB	0.259 dB	15.309 Hz	0.701	8.821%

Table 4 show average speech model single female speech records from A0_00202 to A0_01000 (A is stands Afaan) within 9153 utterances and data distribution train, valid and test respectively evaluation results. From this label state align test correct or accuracy 70.1% at learning rate was 0.002. The displayed label state level alignment for tested the labeled path by using jupyter notebook within different package install in python scripts.

3.6 Generated of Speech Samples

The acoustic training model text, audio pair, which is typically needed to train acoustic models and load wave file to gain text file. The generated speech from the label phoneme alignment for objective test displayed used to

BLSTM base on RNN.

Figure 11: Sample Generated Speech Use Model

This Figure 11 shows the sample generated speech from text pair speech corpus. From this generated speech it is important to download and adjusted within playback speed listen for regulate for many application of TTS like teaching style and another.

(A0_0001 "iji waaqayyoo iddoo hundumaa jira, isa hamaa fi isa gaarii ni arga")
(A0_00208 "fardi dammaqe")
(A0_00209 "waraabeessi boolla keessa dhokate")
(A0_00211 "qe'een isaanii onte")
(A0_00212 "sareen dutte")

3.6.1 Speech in Waveforms

The load the speech was generated into with SciPy input and output wave file packages. Speech synthesis using the spectrogram, and Mel-Frequency Cepstral Coefficients to extract acoustic feature. So, we have converted the extracted acoustic feature using jupyter notebook within pysptk and pyworld in displayed to waveform showed as Figure 12 below.

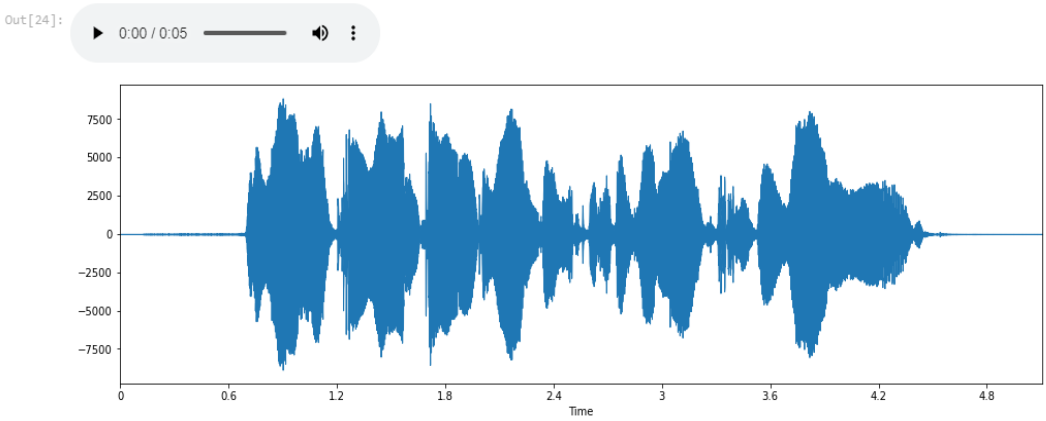

Figure 12 : Generation of Speech in Waveform

The Figure 12 is described, generated the speech in waveforms sample from test label to display librosa in wave plot and pyworld speech synthesis method. This is removed part of silent and silent periodicity in frame of fundamental frequency (f0). The librosa display in attribute wave plot is generated speech within frequency rates in waveforms.

4.1 Evaluation of Text to Speech Synthesized for Afaan Oromoo

Different ways to evaluate the text-to-speech systems occurs. The process model evaluation performance to perform used two ways for TTS Afaan Oromoo language the subjective and objective evaluation.

4.1.2 Subjective Evaluation

The subjective evaluation way to evaluating to listen the tests audio and the listener tackles the speech quality and naturalness.

A total of 13 native Afaan Oromoo speakers and reader randomly selected to evaluated speech quality, the generated speech is given for evaluator that focus on the intelligibility and naturalness of the synthesized speech text pair speech given as Figure was showed. The reason we have chosen from total datasets only thirteen (13) sentences for subjective evaluation to checking the understanding and check the outputs also difficult as the whole datasets tested.

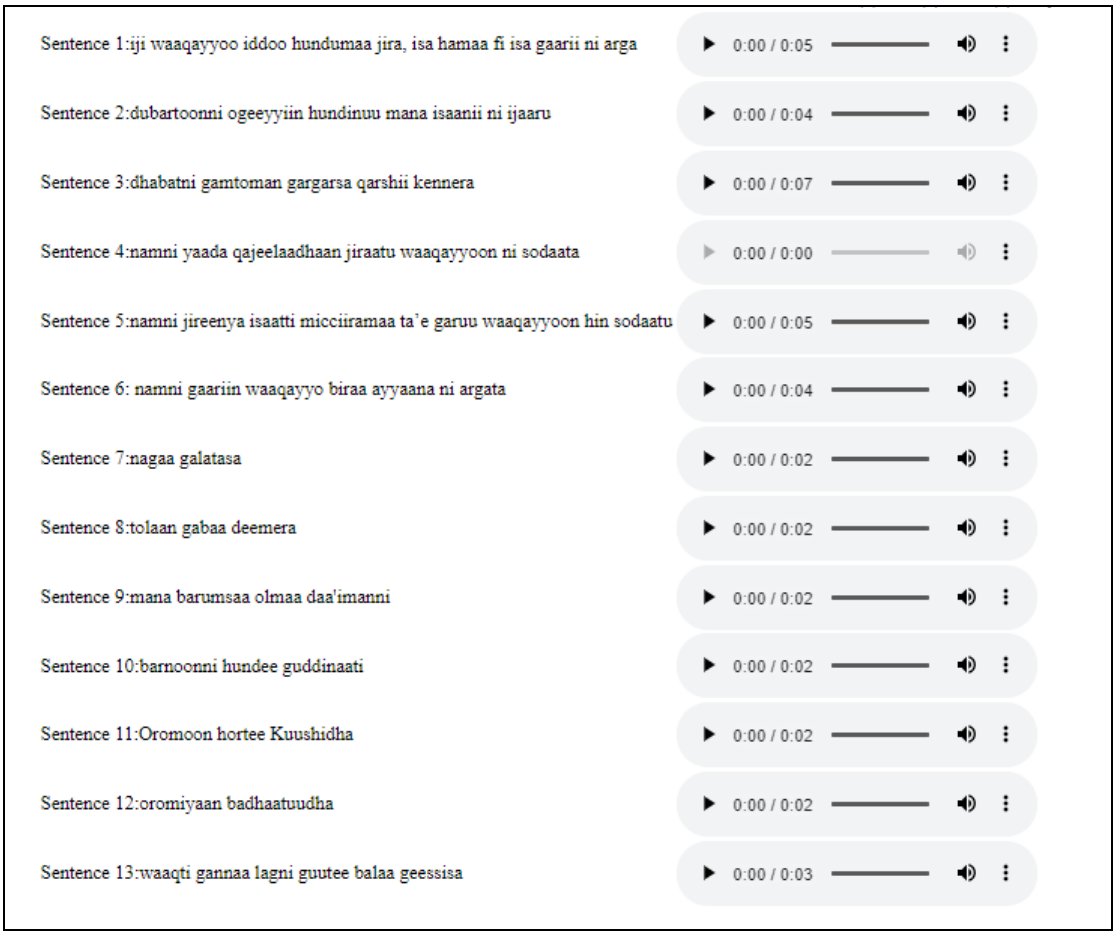

Figure 13:Text pair Speech for Subjective Test

This Figure 13 show the first question is targeting in measuring the intelligibility of the synthesized speech and the second is aimed at measuring whether the synthesized speech is human like or not.

The mean opinion score (MOS) test is chosen for this evaluation, which allowed us to score and compare the global quality of TTS systems with respect to naturalness and intelligibility thirteen (13) sentences for tested by native Afaan Oromoo speaker and professional have in Afaan Oromoo language for Naturalness from each evaluators represented as participants p1-p13 and sentence one(s).

Table 5: Speech Quality for Naturalness from Evalutors

Evaluators	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	Average
p1	3	4	4	5	4	4	5	3	3	4	5	5	4	4.076923
p2	2	4	4	5	3	4	5	2	4	3	4	5	3	3.692308
p3	4	3	3	5	4	4	3	3	5	5	2	3	2	3.538462
p4	5	3	4	4	5	3	4	4	4	3	4	2	3	3.692308
p5	4	2	3	4	4	4	4	4	4	2	2	3	5	3.461538
p6	5	4	5	5	5	5	5	4	4	4	5	4	2	4.384615
p7	3	3	3	4	4	4	3	4	5	3	4	5	2	3.615385
p8	5	5	5	4	4	4	2	2	3	4	5	5	5	4.076923
p9	4	4	4	3	3	4	2	2	3	3	2	3	4	3.153846
p10	3	4	4	4	5	5	4	4	2	3	5	5	5	4.076923
p11	4	5	1	2	5	3	4	5	5	5	4	2	3	3.692308
p12	3	3	4	2	5	4	4	5	5	5	5	4	4	4.076923
p13	4	4	3	3	3	3	3	4	5	4	4	4	1	3.461538
Total Average														3.769231

The participants are allowed to listen to the audio recorded samples before they check the developed text to speech synthesizer. Subsequently, each participant plays the sample record speech to check the quality of the speech. After listening the synthetic speech from the system, the participants are requested to fill marks on the

properly given text pair speech for intelligibility.

Table 6: Speech Quality for Intelligibility from Evaluators

Evaluators	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	Average
p1	3	4	4	5	4	4	5	3	3	4	5	5	4	4.076923
p2	2	4	4	5	3	4	5	2	4	3	4	5	3	3.692308
p3	4	3	3	5	4	4	3	3	5	5	2	3	2	3.538462
p4	5	3	4	4	5	3	4	4	4	3	4	2	3	3.692308
p5	4	2	3	4	4	4	4	4	4	2	2	3	5	3.461538
p6	5	4	5	5	5	5	5	4	4	4	5	4	2	4.384615
p7	3	3	3	4	4	4	3	4	5	3	4	5	2	3.615385
p8	5	5	5	4	4	4	2	2	3	4	5	5	5	4.076923
p9	4	4	4	3	3	4	2	2	3	3	2	3	4	3.153846
p10	3	4	4	4	5	5	4	4	2	3	5	5	5	4.076923
p11	4	5	1	2	5	3	4	5	5	5	4	2	3	3.692308
p12	3	3	4	2	5	4	4	5	5	5	5	4	4	4.076923
p13	4	4	3	3	3	3	3	4	5	4	4	4	2	3.538462
Total Average calculated by MOS														3.775148

Finally, according to mean opinion score, the mean results are calculated as per the respondents' responses.

Table 7: Average MOS Result of Afaan Oromoo Speech Synthesizer

Evaluators	Intelligibility	Naturalness
average score results	3.77	3.76

The results obtained from Table 7, the subjective evaluation are collected for the 13 test sentences. For all the test sessions, results is organized with their set of scores rated by each subject. In order to select the appropriate test to analyses the data, the first step was to apply a normality test. The results revealed that data was normally distributed. Hence, a parametric test was used. The sentence was ranks given by evaluators for naturalness and intelligibility by using calculate the MOS results. The intelligibility is good result values 37.7% was responded from evaluators.

4.1.3 Objective Evaluation

Objective evaluation where measurement of speech quality performance approximated by applying appropriate speech in waveforms. One method using Mel cepstral distortion for evaluating objective and it helping the difference between speech syntheses. This difference shows how the reproduced speech is related to the natural one and not extract like natural speech production.

Table 8: MCD Objective Test

Generate	BLSTM-based on RNN	Merlin _AO_Speech_Syntesis
average score results	3.89	3.71

Table 8 shows the MCD evaluation and test audio pair texts within BLSTM based on RNN and Merlin Afaan Oromoo speech synthesis female speech corpus analysis used to speech signal toolkit and pyworld on jupyter notebook. The total average process of MCD at Mel cepstral coefficients, the recoding audio were 16 kHz to generate the MCD BLSTM-based on RNN is 3.89 and merlin wave generated is around 3.71 respectively.

5. Discussion, Conclusion and Recommendation

5.1 Discussion

The results Discussion from experiment overall and all model as tested on the spectrogram features and MFCC the model training for features synthesis. The output of the using Bidirectional LSTM based on RNN model convert text into speech. Although, the results were achieved on an overfitted model, with longer training and hyperparameters optimization, to achieve desired results on validation data within developed system. At the first, the generated MFCC is perhaps the most important result obtained from the Recurrent Neural Network part of the system.

Using the parameters is learnt by the model, generate MFCC is predicted sequences model from speech source. The converted of the MFCC input a shape of coefficients and sequence length. The Mel generated coefficients are predicted with a source of speech, the effectiveness given in modeling sequential data. Using RNN for acoustic model training to take the time sequence of audio features as input. At each time step, including each of the 33 Phoneset in the Afaan Oromoo language hudha (apostrophe) was removed as punctuations in preprocessed.

The output of the RNN at each time step is a vector of probabilities the entries five short vowels, twenty for

consonant and seven double phoneme (Qubee dacha) where the entry total phoneme in Afaan oromoo language in numbers 33.The characters (Phone) are mapped to indices in the estimate numerical form (binary) and look at the char mapping in the files.

The predictable of RNN model is that they are only able to make use of previous context. In acoustic model, where whole utterances are transcribed at once, predict the feature context. The BLSTM-based on RNN, to complete the bidirectional rnn model function in sample models and specifying the bidirectional covering. The train the deep learning used BLSTM based RNN as specified in input to softmax. Then, the model was finished training, saved in the path for training duration model and training acoustic model to visualize binary formats.

5.2 Conclusion

Text-to-speech synthesis (TTS) which means input texts is generate to the audio from text. In Afaan Oromoo language mostly required to text to speech synthesis for development of this language. So as to transmit information. In this work, a first attempt is investigated speech synthesis for Afaan Oromoo language using deep learning approach based on BLSTM- based on RNN model. The filename is long, but it contains some very important information in activation functions used in hidden layers type is TANH, hidden layer size was 1024 and number of hidden layers were five in numbers. The dimensionality of hidden layers is 1024 number of input nodes (the dimensionality of labels phonemes in model) was 416 and number of output nodes (the dimensionality of acoustic features to predict) is five .Datasets for train file number seven hundred valid files three and tests file three. The buffer size of each block of data to buffer size is 200000 the model file name feed forward 6tanh and the learning rate used was 0.002. From this automatically created in python training and the feature extracted prepared using python scripts.

The purpose of prepared the python feature for extracted duration model and acoustic model preprocessed converted into binary formatted. The dimension of vector was created from parameter generated for visualization utterances length for linguistic features and acoustic feature training and test.

From duration and acoustic plotted visual utterances generated total utterance number and frames. From this utterance plotted we prepared normalization in X max, X min, and Y mean. The Y variable and Y scale for both duration and acoustic train utterance length(utt_lengeth).We Used to the pytorch dataset for generated acoustic model in waveforms and using Recurrent Neural Network within bidirectional activation true in LSTM within vector(416,256) the model for duration and model for acoustic was generated.

At the end from label state alignment and label phone alignment used sample generated speech and user can download the generated speech from Jupyter notebook on python scripting. It can be used as message readers, teaching assistants, tools to aid in communication and learning for the handicapped and impaired challenged people. During developing Afaan Oromoo speech synthesis, the system involved collecting text, preprocessing the text, preparing phonetically balanced sentences, recording the sentences, preparing annotated speech database, and design a prototype. In training first the text and speech corpus are manually prepared for processing. Mel-cepstral coefficients parameters are obtained from speech data sources used Mel-cepstral analysis like pyworld and pysptk. Then using automatic htk toolkit and front end used to festival toolkit context labeler in state level alignment and label phone alignment. The text corpus and speech parameters are align to generate linguistic (utterance) feature. However, every feature of Afaan Oromoo language was not considered because it needs a lot of time and deep linguistic way of creation of Afaan Oromoo phonemes are considered. The Mean opinion score evaluation technique was used to subjective test the performance of synthesized speech from the evaluator of native speech to listen the audio recorded within their test. The thirteen sentences used for testing used for subjective test, the result is 3.76 and 3.77 out of 5 score in terms of naturalness and intelligibility respectively.

5.3 Contribution of This Study

- ✓ Development of audio recording tools that facilitate the audio recording process by providing the text transcription to be recorded with its aligned audio file name
- ✓ We have developed a phoneme map for generated the linguistic features for Afaan Oromoo language.
- ✓ We have prepared own corpus which can be used for text to speech synthesis to create sound with good naturalness and intelligibility
- ✓ We have addressed the issues of duration and acoustic modeling using BLSTM based on RNN for Afaan Oromoo.

5.4 Recommendation and Future Works

In this study, Afaan Oromoo deep learning based on BLSTM- based on RNN speech synthesizer was developed. The bandmat a periodicities form package used independency bandmat. The bandmat kind of feature extracted from the audio, and one file for every audio file from the prepared text pair speech corpus.

Label files (time-to-phone alignments) these files are the label files which contain alignments for phoneme

alignments or state alignments audio files from the datasets.

The log-fundamental frequencies (lf0) is the log-fundamental frequency files feature file extracted from our audio files in datasets.

Mel-generalized cepstral coefficients (mgc) these files contain the generalized cepstral coefficients for our audio files in our datasets and move next feature file type. In these files contain the generalized cepstral coefficients for our audio files in our data set and script file for filenames to create the acoustic model and duration model.

The test speech synthesis the sound generated (utterance and prompt utterance, train sentences, file identification list). The deep learning model used to Bidirectional LSTM-based RNNs. Using pytorch, simple to demonistrated.The BLSTM is undestanded the previous input data or gate data used and RNN predicts output feature sequence given an input feature sequence. The Training acoustic model and duration model generated the audio from text converted to label (time to phoneme each sentence) using on python command line or jupyter notebook within Control Process Unit (CPU) device in torch package configuration.

In the future will be use deep neural network model (DNN) hybrid within HMM techniques statistical speech synthesis parameters unit selection synthesis speech.

In the future, we will try to investigate deep bidirectional long short-term memory recurrent neural network (DBLSTM-RNN) with an even details structure with a larger corpus and considering all the abbreviated words, numbers, to provide a high quality of speech synthesized for language. Another task that needs future work is effectively acoustic feature extraction used to WaveRNN model within Tacotron speech synthesis considered in developing within NVIDIA Nsight HUD Launcher 5.4 driver (CUDA toolkit) for fast speed processing large datasets.

References

A. Balyan, S. S. (2013). Speech Synthesis. A Review," International Journal of Engineering Research & Technology (IJERT), pp. 57-75.

Alem F., K. N. (2007). “Text To Speech for Bangla Language using Festival” . BRSC University, Bangladesh.

Allen Jonathan, H. M. (1987). From Text to Speech: The MITalk system. Cambridge University Press.

Alula. (2010). A generalized approach for Amharic Text To Speech system. Addis Abeba.

Barnwell, A. V. (1995, Jul). A mixed excitation LPC vocoder model for low bit rate speech coding. in IEEE Transactions on Speech and Audio Processing, vol. 3, pp. 242-250.

Bluche, T. N. (2013). Tandem HMM with Convolutional Neural Network for Handwritten Word Recognition. In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP2013),. Vancouver, BC, Canada.

Cassia Valentin. (2013). Intelligibility Enhancement Of Synthetic Speech In Noise. Ph. D. Dissertation. University Of Edinburgh, Germany.

Charpentier, M. a. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication.

Christian Kratzenstein. (1779). the Danish scientist working at the Russian Academy of Sciences, built the first talking machine. Danish, Russian Academy.

Dutoit, T. (1997). A Short Introduction to Text-to-Speech. Dordrecht, Boston, London.: Kluwer Academic Publisher.

Figueiredo, A. I. (2006). Automatically Estimating the Input Parameters of Formant-Based Speech Synthesizers.

Flanagan, J. (1965). Speech analysis, synthesis, and perception. Springer, Berlin.

Goubanova, O. T. (2000). Using Bayesian belief networks for model duration in text-to -speech systems. In Proc. of ICSLP ICSLP-20 00, Beijing, China.

Graves A. Mohamed, A. G. (2013). Speech recognition with deep recurrent neural networks.In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (Vol. 38th). Vancouver, BC, Canada,.

Graves, A. (2012.). Supervised Sequence Labelling with Recurrent Neural Networks;. Springer: Berlin, Germany,.

Graves, A., Jaitly, N., & Mohamed, A. (8–12 December 2013;). Hybrid speech recognition with Deep Bidirectional LSTM. In Proceedings of the 2013 IEEEWorkshop on Automatic Speech Recognition and Understanding,. Olomouc, Czech Republic,.

Holmes, J. H. (2003). “Speech Synthesis and Recognition” e,. Taylor and Francies New Fetter Lan, London ECAP 4EE.

I. H. Witten, E. F. (2012). Practical Machine Learning Tools and Techniques of Data Mining .

Javidan, R. (2010). Concatenative Synthesis ofPersian Language Based on Word, Diphone and Triphone Data-bases. Persian.

Klatt, D. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of

the Acoustic Society of America.
Klatt, D. (1980,). Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am., 67, 971–995.
Kubichek, R. (1993). Mel-cepstral distance measure for objective speech quality assesment. In Proceedings of IEEE Pacific Rim Conference on Communications, vol. volume 1, pp. pages 125–128.
Lazaridis, A. Z. (, 2007). Segmental duration modeling for Greek speech synthesis. In Proc.of IEEE ICTAI ICTAI-2007, Patras, Greece,.
Lemmetty, S. (1999). Review of speech synthesis technology. Helsinki University of Technology. From http://www.acoustics.hut.fi/~slemmett/dippa/%5Cnhttp://www.acoustics.hut.fi/public
Lloret, M. (1997). “Oromo Phonology” ,Phonologies of Asia and Africa . Winona Lake, Ind. Eisenbrauns.
M. S. Al-Radhi, T. G. (2017). "Deep Recurrent Neural Networks in Speech Synthesis Using a Continuous Vocoder. n International Conference on Speech and Computer.
Melba, G. (1988). Introduction of Oromo people and Oromia.
Morise.M, F. Y. (2016). WORLD A vocoder-based high quality speech synthesis system for real-time applications in IEICE Transactions on Information and Systems.
Morka, H. (20013). Afaan Oromo TTS system. Addis Abeba.
Morka.M. (2003). Text-To-Speech System for Afan Oromo Masters of Thesis. Addis Ababa University.
Moulines, E., & Charpentier, F. (1990,). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphone. Speech Commun.
Naslund, P. (2018). Artificial Neural Networks in Swedish Speech Synthesis Master in Computer Science.
Ofgaa, S. T. (,May, 2011). Concatenative Text-To-Speech System for Afaan Oromo Language .
Rabiner L.R and Juang, B. (1993). Fundamentals of Speech Recognition. Englewood Cliff. New Jersey: Prentice Hall, Inc.
Rashad, M. E.-B. (2010). An overview of text-tospeech synthesis techniques.
Rodman, R. D. (1999). Computer Speech Technology. A. rtech House, Inc., London.
Samson, T. O. (2011). Concatenative Text-To-Speech System for Afaan Oromo Language. Addis Ababa Universisty, Ethiopia.
Samuel, T. (2007). “Natural Sounding Text-ToSpeech Synthesis, based on Syllable-Like Units”, Master of Science . India.
Sangramsing N. Kayte, D. G. (2015). The Marathi Text-To-Speech Synthesizer Based On Artificial Neural Networks. International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue,. From www.irjet.net
Solomon, T. (2005). Automatic Speech Recognition for Amharic. Doctoral Dissertation. Hamburg University,Germany.
Sproat, R. B. (2001). “Normalization of Non-standard Words, Computer Speech and Language” (Vols. Vol. 15,).
Sutskever, I. V. (2014). Sequence to sequence learning with neural networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, ,. Montreal, QB, Canada.
Sutskever, I., Vinyals, O., & Le, Q. (2014). Sequence-to-Sequence Neural Network Models for Grapheme-to-Phoneme Conversion. ,Lake Tahoe, NV, USA.
Takeda, K. S. (1989 .). On sentence sentence-level factors governing segmental duration in Japanese. Journal of the Acoustical Society of America.
Tesfaye. (2004). Diphone Based Text To Speech System for Tigrigna Language. Addis Abeba.
Tewodros. (2009). Text To Speech synthesizer for Wolaytta language . Addi Abeba.
Tilahun, G. (1993). Qubee Afaan Oromoo: Reasons for Choosing the Latin Script for Developing an Oromo Alphabet t.The Journal of OromoStudies.
Van Santen, J. (1992). Contextual effects on vowel durations. Speech Communication.
Wosho, K. M. (2020). Text to Speech Synthesizer for Afaan Oromoo using Statistical Parametric Speech Synthesis. Addis Ababa, Ethiopia.
Xu, S. (2007). Study on HMM-Based Chinese Speech Synthesis; Beijing University of Posts and Telecommunications:. Beijing, China.
Y. Fan, Y. Q. (2014). TTS Synthesis with Bidirectional LSTM. Asia, Beijing, China.
Yamagishi et al. (2004). speaker-independent and speaker adaptability which give special emphasis on voice characteristics such as speaker individualities, speaking styles, and emotions.
Yang, J. (2014). Deep learning theory and its application in speech recognition.Commun. Countermeas.
Yao, K., & Zweig, G. (2015). Sequence-to-Sequence Neural Network Models for Grapheme-to-Phoneme Conversion. In Proceedings of the Annual Conference of the International Speech Communication Association, Dresden,. Germany.
Yoshimura, T. (2002). Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based Text-To-Speech systems. Nagoya Institute of Technology: PhD dissertation.
Yoshimura, T. T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech

synthesis. In Proceedings of the Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99). Budapest, Hungary,.

Zen, H. S. (2013). Deep learning method speech sysnthesis. Vancouver, BC, Canada.

Zen, H. T., & Alan, W. (2009). Statistical parametric speech synthesis. Speech Commun.