Text to Speech Synthesizer for Afaan Oromo Using Hidden Markov Model

Muhidin Kedir Wosho
Lecturer at Mizan-Tepi University
Email: muhidinkedir@gmail.com

***Abstract*—** This study explores the application of natural language processing techniques with text to speech synthesis for the Afaan Oromo language using the "Hidden Markov model" on 600 news datasets that were prepared in collaboration with linguists and experts of Afaan Oromo language. Speech synthesizers are the most essential in helping impaired people, in the teaching and learning process, for telecommunications and industries. The dataset was tested on a hidden Markov model algorithm. The synthesiser has two core components: training and testing phases. In this study, the subjective Mean Opinion Score (MOS) and objective Mel Cepstral Distortion (MCD) evaluation techniques are used. The subjective results obtained using the mean opinion score (MOS) are 4.3 and 4.1 in terms of intelligibility and naturalness of the synthesised speech, respectively. The objective result obtained using the mean opinion score is 6.8 out of 8 that is encouraging.

***Keywords*—** Hidden Markov model, Text to Speech, Afaan Oromoo.

1. INTRODUCTION

Natural language processing (NLP) is a field that works with computational techniques for the purpose of education, understanding, and producing human language content [1]. The goal of NLP is to design and build software that will analyze, understand, and generate languages that humans use naturally [1]. The most important limitation of NLP today is the fact that most NLP resources and systems are available only for high-resource languages (HRLs), such as English, French, Spanish, German, and Chinese. However, many low-resource languages (LRLs) like Bengali, Indonesian, Punjabi, Ethiopia, and Swahili have spoken are written by many people that haven't any such resources or systems available. Text-to-speech synthesis (TTS) is one of among the foremost emerging technologies in speech processing techniques for producing speech signals from a randomly given text so as to transmit information from a machine to a person by voice [2]. The main goal of text-to-speech (TTS) synthesis is to produce natural-sounding speech from arbitrary text [3]. In the context of the Afaan Oromo language, there is a lack of a proper TTS system that considers such as robustness, small footprint, naturalness, intelligibility, cost-effectiveness, and expressivity for the language's speakers can be raised as the main problem behind conducting the research [4]. Afaan Oromo is also the most widely used language in Kenya, Somalia, and Djibouti in addition to Ethiopia. Even though the language is spoken and is a representation language for quite 50 million people, yet a scarcity of active research on the Afaan Oromo text to speech synthesis factors for this work to come up with models that can alleviate these problems.

2. RELATED WORKS

Tokuda et al. [31] develop an HMM speech synthesis system for the English language. The authors have used festival speech tools for text analysis and feature extraction such as contextual factors. For training the model, they have used 524 sentences and the speech signal was sampled at the rate of 16 kHz. The contextual factors have been considered during speech synthesis. However, the authors did not put a quantitative analysis of the result and simply concluded that the provided result of this synthesized speech is good as compared to any other rule-based speech synthesizers, like formant based approach.

Ntsako et al. [5] Develop a highly intelligible and acceptably natural-sounding speech synthesis system for the Xitsonga language using a hidden Markov model (HMM) speech synthesis method. They found the overall system was rated as excellent by 7.7%, good by 38.9%, acceptable by 46%, and poor by 7.7% of the respondents. This means that the authors received an acceptability level of 92.3%. However, this method can synthesize speech on a footprint of only a few megabytes of training speech data. Finally, the synthesized speech is produced from the speech parameters. A overall of five hundred sentences are used for training the model from a corpus having a size of 11,670 sentences and twenty sentences that are not included in the training dataset are used for testing the performance of the system. Generally, even if these related works have a great contribution to the area, speech synthesis on the Afaan Oromo language has not yet thoroughly explored like other abroad languages. Based on the review made and as to the knowledge of

the researcher none of the work has been attempted to design speech synthesizer for Afaan Oromo language with the integration of non-standard words of advantage over the other approach such as needs smaller corpus for training, requires very little memory and can easily be integrated into handheld devices. The other quality of the HMM-based speech synthesis approach is the flexibility to change voice characteristics. Therefore, the Hidden Markov model speech synthesis is the best and the most significant approach to solve the challenges of speech synthesis such as naturalness, intelligibility, cost-effectiveness, expressivity and produce average speech units, smooth to stable, stores statistics rather than waveforms.

3. METHODOLOGY

3.1 Data Collection

To construct the Afaan Oromo speech database, first phonetically balanced sentences are collected. These sentences are collected from news, blogs, stories, political essays, literary works, sports sections, magazines, holy books, proverbs, and newspapers. In the Afaan Oromo language, there are 26 pure phonemes and 7 borrowed phonemes are available. For the evaluation purpose, we have used ten sentences are selected randomly out of the trained dataset.

The recording process took place in an office environment with minimal noise. The speech is recorded at a noise-free studio using a PC at Oromia Broad Casting Networks (OBN) Biiro at Adama with female journalists. The speech sample was recorded at 44.1 kHz stereo and the files are stored in waveform format (.wav). The waveform files are normalized and changed to conform to 16 kHz, 16 bit, RIFF format as required by the festvox system and to make it easier to create raw files of small sizes. Praat is the software that is used to record the speech corpus. A regular microphone and a normal office computer made up the hardware equipment used to record the speech corpus.

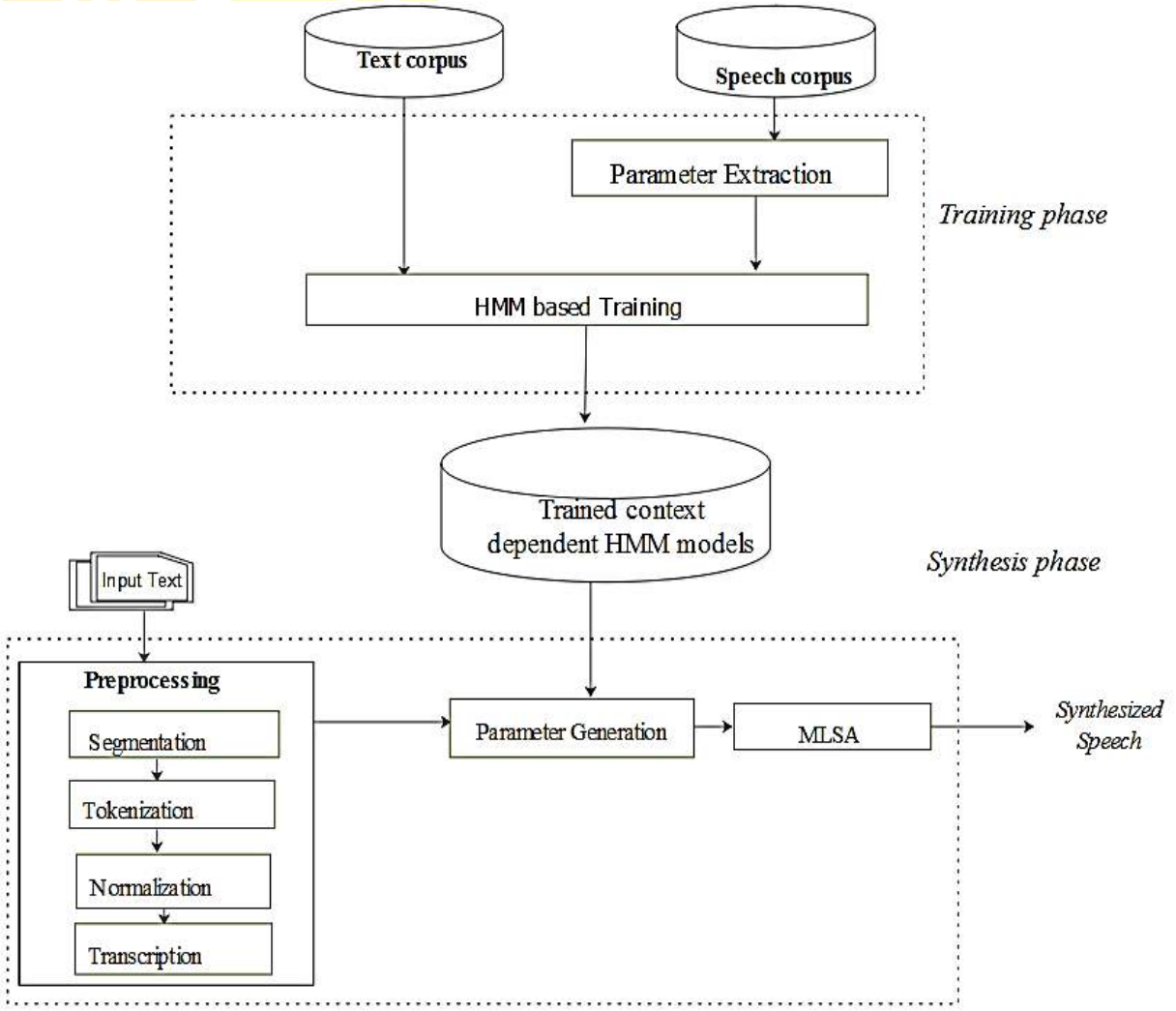

Figure 1: Architecture for TTS for Afaan Oromoo using HMM

3.2 Trained Parameters Models

Defining the structure and complete form of a set of HMMs is the first step towards building a synthesizer. The second step is to assessment the parameters of the HMMs from the data sequences that they are intended to model. Those kinds of activity of the parameter estimation is usually called training. Training the model means estimating the HMMs parameters, which are the mean, the variance, and the transition probabilities based on the utterance structure and the extracted parameters (features). Once the parameters are extracted, training of HMMs is performed with the hidden Markov model toolkit (HTK).

```
Input: Text corpus and speech corpus
Output: Model trained context dependent
HMM and duration
Begin
Input prepared or defined text and speech
corpus
Extract the speech parameters from speech
database
Align the analyzed text with its
parameters
Trained and Cluster based on F0, spectrum
and duration
Generate trained context dependent
parameters
End if
```

Algorithm 4.1: Algorithms for trained the HMM based speech.

PREPROCESSING

Preprocessing is one of the most powerful and complex task in natural language processing module. In this work, the raw text is preprocessed into a context dependent label sequence by a text analyzer. These procedures of raw text preprocessing include four main tasks such as sentence segmentations, tokenization, and normalization and text transcription (G2P) conversation.

Tokenization

Tokenization is the process of segmenting and running texts into words, sentences and phonemes. Tokenization is the process of breaking sequences of sentences into its constituent words. During tokenization, the white space delimiter and special characters called hudha " ' " (diacritical) are the main focus areas where, whenever there exists the mentioned delimiters between characters in a sentence, sequences of characters are broken to produce a meaningful word for a given specific language. The festival and festvox speech tools can understand Afaan Oromoo letter to sound transcription, phone sets, and tokenizes any amount of standard words into single word by looking for the white space. For instance, the sentence 'Tolosan barata cimadha' which means 'Tolosa is a clever student', is tokenized into three tokens: Tolosan, barata and cimadha

TEXT NORMALIZATION

Text normalization is the process of generating normalized word from the text containing nonstandard words. Afaan Oromo text includes nonstandard words (NSW), which cannot be directly transcribed to phonemes, e.g. numbers, dates, abbreviations, currency, etc. These activities include, the conversion of nonstandard words into their orthographic representation by expansion into their full spoken words. Grapheme to phoneme (G2P) conversion is used for the reason that it is supported in festival system and Afaan Oromo being a phonetic language.

```
Input: input text
Output: transcription of the text
Begin
Input the text
Define phone set and AOLTS module
Preprocessing the text
Generate transcribed text (grapheme to
phoneme)
End if
```

Algorithms 4.2: Algorithms for transcription text.

For instance, the initial probabiity of phoneme KON= 1.0%, KO=0.0%, LA=0.0%, TA= 0.0% found on database respectively. So having this mathematical probability estimation, we choose the first letter that have highest probability KON as initial.

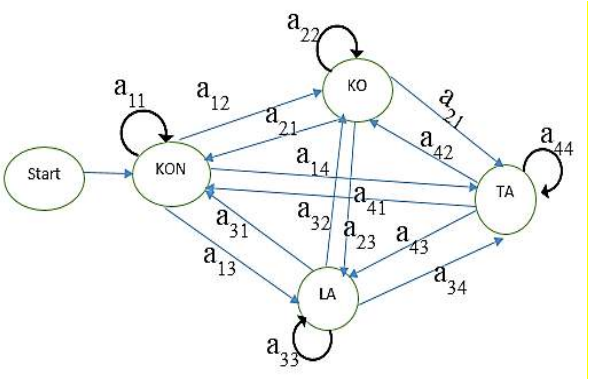

Figure 2. 6: HMM State for the word "KONKOLATA".

Table 1: Shows the example to calculate probability distribution for the word "KONKOLATA".

	KON	KO	LA	TA
KON	0.3	0.9	0.6	0.5
KO	0.1	0.5	0.8	0.4
LA	0.3	0.2	0.7	0.8
TA	0.2	0.5	0.6	0.7

The value of each phoneme are randomly given by the language expert to calculate the probability distribution function to know the corresponding phonemes from the given words using Bayes rule. By using Bayes rule we can calculate the probability of each phoneme concatenation and choose the highest confidence probability as updated HMM state for the next sequence and continues by follow the above step until the full concatenation of grapheme with the language grammatical structure i.e., KONKO, *KONKOLA*, *KONKOLATA*. At training stage, acoustic (real-valued) and linguistic (discrete) features sequences are extracted from the speech waveforms and its transcription respectively [6]. Then the acoustic model is trained to model the conditional distribution of an acoustic feature sequence given a linguistic feature as following formula.

$$\hat{\lambda} = arg\max_{\lambda} P(o|l,W) \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Where O is acoustic feature sequence l is the linguistic features and λ is the acoustic *model*.

In equation (8), the fundamental problem that needs to be solved in corpus-based speech synthesis, i.e., finding the most likely speech parameters **O** for a given word sequence W using the training data **O**, and the corresponding word sequence W. We can solve this maximization problem by using Bayesian speech parameter generation algorithms. Bayesian Approach: Bayesian learning is used to estimate posterior distribution of model parameters from prior distributions and training data [7, 6]. Bayesian learning techniques applied to HMM- based speech synthesis to determine the observation (hidden) O as following formula:

Bayes Rule

$$P\left(\frac{A}{B}\right) = \frac{P(B|A) \,.\, P(A)}{P(B)} = \frac{P(B|A) \,.\, P(A)}{\sum_i P(B/Ai)P(Ai)} \ldots\ldots\ldots (9)$$

At synthesis at text to be synthesized is first converted to corresponding linguistic feature sequences is predicted from the trained acoustic model [6] as follows

$$\hat{O} = arg\max_{0} P(o|l,\ \hat{\lambda}) \ldots\ldots\ldots\ldots\ldots (10)$$

4. RESULTS AND EVALUATION

The artificial speech is evaluated depending on its naturalness and intelligibility using subjective (MOS) and Mel cepstral distortion (MCD) as objective terms. MOS is a technique that indicates their assessments on a scale of bad (1) to excellent (5). Six hundred sentences are used for training and ten sentences for testing. Five men and five females' native speakers of the language are randomly selected to evaluate the system.

The participants are allowed to listen to the recorded voice samples before they check the developed text to the speech synthesizer. Subsequently, each participant plays the sample voice to check the quality of the voice. However, in order to make the test effort easy and understandable by the participants, a questionnaire is delivered beforehand to familiarize them with it. Then, after listening to the synthetic speech from the system, the participants are requested to fill in marks on the questionnaire properly. Finally, according to the mean opinion score, the mean and standard deviation cumulative results are calculated as per the respondents' responses. Table 2 shows the result.

Table 2: The Average MOS result of Afaan Oromo Speech Synthesizer

Evaluators	Intelligibility	Naturalness
Females average score result	4.6	3.9
Males average score result	4.2	4.3
Average	4.3	4.1

Subjective evaluation is expensive and time consuming. An objective evaluation would offer an alternative for assessing synthetic speech. This is known as mel cepstral distortion (MCD). The system correlates the effective characteristics of the natural sound with the synthetic sound. Our system works on tenfold cross validation (9/10) rule which is widely used and an effective training method for the system. Accordingly, using mel cepstral distortion (MCD) the result obtained is 6.8, which is very encouraging.

5. CONCLUSION

Text-to-Speech (TTS) synthesis can convert an arbitrary input text to intelligible and natural-sounding speech so as to transmit information from a machine to a person. It can be used as message readers, teaching assistants, tools to aid in communication, and learning for the handicapped and visually challenged people. During developing Afaan Oromo speech synthesis, the system

involved collecting text, preprocessing the text, preparing phonetically balanced sentences, recording the sentences, preparing an annotated speech database, and design a prototype. The mean opinion score evaluation technique was used to test the performance of the system. For this study, six hundred sentences are used for training, and out of the trained sentences ten arbitrary sentences used for testing. According to training and testing our system, we obtained the result 4.1 and 4.3 out of 5 scores in terms of naturalness and intelligibility respectively. A tenfold threshold method is used for training and testing of the prototype.

REFERENCES

[1] C. Manning and J. H. Christopher, "advanced in natural language processing," in in proceeding of the 52nd annual meeting of the association for computational liticsnguis, stanford, 2014.

[2] T. Masuko, "HMM-Based Speech Synthesis and Its Applications," Tokyo, November, 2002.

[3] S. R. Mache, M. R. Baheti and C. N. Mahende, "Review on Text-To-Speech Synthesizer," International Journal of Advanced Research in Computer and Communication Engineering, vol. Vol. 4, no. Issue 8, p. 54, August 2015.

[4] S. T. Ofgaa, "Concatenative Text-To-Speech System for Afaan Oromo Language," Addis ababa, Ethiopia, may,2011.

[5] B. Ntsako, "Text- To-Speech Synthesis System for Xitsonga using Hidden Markov Models," June, 2012.

[6] N. K. a. T. T. Tokuda K, "Speech Synthesis Based on Hidden Markov Models," in Edinburg Research Explorer, Japan, januart 2013.

[7] S. K, "HMMs as Generative Models of Speech," Workshop on Text-to-Speech (TTS) Synthesis, 16-18 June 2014.