American Journal of Embedded Systems and Applications
2022; 9(1): 6-12
http://www.sciencepublishinggroup.com/j/ajesa
doi: 10.11648/j.ajesa.20220901.12
ISSN: 2376-6069 (Print); ISSN: 2376-6085 (Online)

Hidden Markov Model Based Text to Speech Synthesis for Afan Oromo

Kumera Chala Chemeda

Department of Information Technology, Mettu University, Mettu, Ethiopia

Email address:
kumerachala.au@gmail.com

To cite this article:
Kumera Chala Chemeda. Hidden Markov Model Based Text to Speech Synthesis for Afan Oromo. *American Journal of Embedded Systems and Applications*. Vol. 9, No. 1, 2022, pp. 6-12. doi: 10.11648/j.ajesa.20220901.12

Received: January 31, 2022; **Accepted**: March 7, 2022; **Published**: March 15, 2022

Abstract: The goal of a text to speech synthesis system is to produce comprehensible and natural sounding speech. For years, synthesized speech has been a difficult task. Various strategies have been developed and deployed to overcome these obstacles. Though hidden markov models are used to create voice synthesizers for foreign languages, they are not suitable for Afan Oromo since the language's unique peculiarities are not considered. As a result, Hidden Markov Model-based text-to-speech synthesis for Afan Oromo is performed in this study. Text to speech synthesis using a hidden markov model comprises two parts: training and synthesis. Preparing the training dataset, constructing utterance structure, language modeling, generating labeled text, feature extraction, and training the model are the primary operations in the training section. Models are chosen according to the text to be synthesized in the synthesis section, and then speech parameters are constructed from them. Finally, the speech parameters are used to make synthetic speech. A corpus of 10,112 sentences was utilized to train the model, yielding a total of 527 sentences. A total of 20 words are utilized to test the system's performance and are not part of the training dataset. The mean opinion score evaluation technique is employed in this study. The average opinion score for intelligibility and naturalness was found to be 4.04 and 3.51, respectively. The synthesis system is categorized as good in terms of intelligibility and fair in terms of naturalness, according to the mean opinion score results. The results are promising, and further study directions are suggested to increase the system's performance.

Keywords: Text to Speech Synthesis, Hidden Markov Model, Language Modeling, Afan Oromo

1. Introduction

As you know, each individual need that the computer system must act like human being and showed to be user friendly [12]. Many researchers have imagined machines infiltrating every aspect of human life. Speech and spoken words have long played a significant part in people's individual and collective life. In everyday life, speech is the most important method of communication. As a result, researchers and academics have been working hard to make the machine speak natural language and make the job easier. Converting text to speech is a complex process the brain exactly controls the articulatory system at higher speed and teller gets acoustic response of his/her communication via hearing organs. To be able to mimic speech production artificially, not only the articulatory system but also the mechanism of the brain should be understood [16]. Many researchers have worked in the field of text to speech synthesis systems throughout the last few decades [4].

Natural language processing is a science that deals with the interpretation of human (natural) language. It comes from computer science because the target devices for such processing are computers or other processing units. Natural language processing is highly beneficial in the creation of a written text from an input voice sound (text to speech synthesis) and the development of speech from an input text (speech recognition) [13].

Converting text to speech Synthesis is the creation of human sound by humans. A text-to-speech technology turns written material into spoken words. Speech synthesizer is a system used for converting text to speech and also it is implemented in hardware and software products. As a result, speech synthesis can be used in spoken dialog systems, applications for the blind and visually handicapped, communications applications, and hands-free apps [18].

The non-natural creation of sound has a long history. The

first mechanical speech production system built by Farkas Kempelen in 1791 [23].

Text to Speech synthesis can be synthesized mostly by three approaches: These are: Articulatory synthesis, Concatenative synthesis and Formant synthesis [6].

Computational methods for producing speech based on models of the human vocal tract and the articulation procedures that occur there are known as articulatory synthesis. The form of the vocal tract can be controlled in a variety of ways, as can the location of the speech articulators such as the tongue, jaw, and lips. The flow of air is digitally simulated using a simulation of the vocal tract to produce speech. Because of the complexity of human articulation organs this technique is very difficult to implement [10].

Another approach in text to speech synthesis is concatenative synthesis. In this approach pieces of recorded speech are concatenated together to produce natural sounding synthesized speech. Differences between natural changes in speech and the nature of automated methods for separating recorded speech into waveforms, on the other hand, can occasionally cause perceptible difficulties in the final product [7].

The third approach in text to speech synthesis is format synthesis. This synthesis does not use the recorded speech samples at execution time. However, additive synthesis and an acoustic model are used to construct the synthesized speech (physical modelling synthesis). To generate the waveform of an artificial speech various levels of parameters can be changed over time such as f0, voicing and noise.

This method is known as rule-based synthesis. Artificial, robotic-sounding speech is generated using formant synthesis technology. Because it is difficult to estimate the vocal tract model and source parameters, rule-based formant synthesis can generate good speech that sounds odd [11].

In this research work, a text-to-speech synthesis system approach based on HMM is selected. The HMM-based text-to-speech synthesis is also referred to as statistical speech synthesis (SPS). In this system, the frequency spectrum (vocal tract), F0, and duration of speech are demonstrated at the same time by HMMs and the speech waveforms are produced by themselves from HMMs based on the maximum likelihood criterion [25].

The HTS toolkit is used for experimentation purposes. The capacity to synthesize comprehensible and natural sounding speech without requiring a large training corpus drove the choice of HMM-based Text to Speech synthesis over alternative techniques [5].

This method completes the goal by employing HMMs to statistically model speech parameters. Furthermore, when removing the text analysis component, the real-time synthesis engine of HTS, the software used for hidden Markov model-based text to speech synthesis, is quite modest, covering just a few megabytes (MBs). Low memory needs, versatility, and adaptability to speaker voice features and speaking styles were some of the factors that influenced the choice of this text to speech synthesis approach over others. As a result, a text-to-speech synthesis system based on Hidden Markov Model-based text-to-speech synthesis can be implemented on a variety of platforms. Speech synthesizers based on HMMs can sound more natural than formant synthesizers. Furthermore, they are more resistant to changes in speech quality than unit selection systems.

2. Statement of the Problem

Speech is one of the essential forms of communication for human being and it is the basic activity in our day to day life. Currently, many speech synthesis systems have been made possible and successful results were obtained in various application areas for languages such as English, Japanese, Spanish, etc. However, there has been little work on text to speech synthesis for languages spoken in Ethiopia. Some of the studies are carried out on Amharic, Afan Oromo, wolaytta and etc.

Samson [17] tried to develop a text to speech synthesis for Afan Oromo language based on concatenative techniques, where diphone and triphones are the speech units that are focused on.

Argaw and Sebsibe [2] attempted to build a text to speech synthesis system for Afan Oromo Using Unit Selection techniques, but they never consider Handling prosodic issues (intonation, stress, and Duration) and Application of HMM-based speech synthesis method were not applied. The researchers recommend the need to work on improvement of the quality of speech synthesizer for Afan Oromo language.

A speech synthesis system that generates natural sounding and intelligible speech with small resource requirement is essential for many application areas. The objective of a text to speech synthesis system is to generate a human like voice from random text. These systems have been under development for several decades. Recent progress in speech synthesis has produced synthesizers with high intelligibility and naturalness. Traditional concatenative or unit selection based speech synthesis systems which synthesize the speech by joining different length speech-units (like phones, diphones and syllables, etc.) derived from the natural speech, requires large amount of training data to synthesis good quality of speech [9]. However, it is very difficult to collect and store large speech corpora. Furthermore, the quality of synthesis in these systems depend upon the goodness in joining of the natural speech units. To overcome these problems, the researcher used the Hidden markov model based text to speech synthesis system for Afan Oromo Language.

Research questions

This research tries to answer and address the following research questions:

1) What is the overall performance of hidden markov model based text to speech synthesis for Afan Oromo language?
2) Can hidden markov model technique improve the problem in Afan Oromo text to speech synthesis?
3) What are the parameters used to develop speech synthesizer for Afan Oromo language based on HMM techniques?

4) What are the limitations in developed HMM based TTS for Afan Oromo language?

3. Related Works

The most generally used and easiest way of evaluating speech quality is the Mean Opinion Score (MOS), which is employed in this study. It can also be used to assess synthetic speech in general. MOS is a five-level measure that ranges from poor (1) to outstanding (5) [19]. For both intelligibility and naturalness criteria of synthesized speech, the assessors give their opinions based on the MOS scale. Related literatures are reviewed that deals about the text to speech synthesis systems developed for local and foreign languages with special focus on those researches done using hidden markov model speech synthesis techniques.

Agazi and Tibebe [1] tried the speech synthesis Entitled with "Unit Selection Based Text- to-Speech Synthesizer for Tigrinya Language" is developed using festival frame work. They constructed a corpus of 13171 words and they are selected from a large corpus. The collected corpus is intended to cover regularly used syllable and context [1]. The results they obtained shows that 38.8% the speech was very good to listen and 58.3% the speech was good and 2.7% the speech is unnatural. Generally, the result was acceptable.

Hailemariam et al. [8] experimented with a "Unit Selection Voice for Amharic Using Festvox" research project. They used a transliteration approach to create a unit selection concatenative speech synthesizer. Festvox, a voice building framework used for developing unit selection voices in a new language, is employed, as stated in the research report. The research's perceptual evaluation used six levels ranging from Excellent (5) to Very Poor (0), yielding a score of 2.9. Finally, the researchers proposed that the number of entries in the corpus used be increased in order to improve quality.

Tewodros [19] undertake a research work on "TTS synthesizer for Wolaytta language" for the first time. Residual LPC was used as a smoothing technique in the TTS, which was based on diphone speech units. Finally, the system's performance was rated at 78 percent, while the naturalness and intelligence of the system were rated at 2.77 and 3.17 on the MOS scale, respectively.

As a result, no previous study has been done to construct a text to speech synthesis utilizing HMM for Afan Oromo languages. This study can be considered a first attempt to apply this technique to the Afan Oromo language. Other languages, such as English, Greek, Swedish, and Malay, make use of the approach.

4. Research Methodology

An HMM-based text to speech synthesis approach is selected for use in this research work. This approach is flexible for adapting to multiple utterer's voice characteristics and styles of speaking. It needs less memory for run time engine and less recorded speech data for training the system [10].

An HMM based speech synthesis software called Hidden markov model-based text to speech synthesis which is given as patch to a Hidden Markov Model Toolkit is used. By this approach intelligible and natural speech can be generated. There are two parts for hidden markov model-based text to speech synthesis. Those are training and synthesis parts.

Data preparation, language modeling, feature extraction (i.e., Parameter Extraction and Spectral Parameter Extraction), utterance structure generation, labeled text generation, and HMM construction are all included in the training phase. The synthesis phase, on the other hand, entails creating labeled text from the text input, extracting speech parameters from HMMs, and then constructing the speech waveform from the speech parameters.

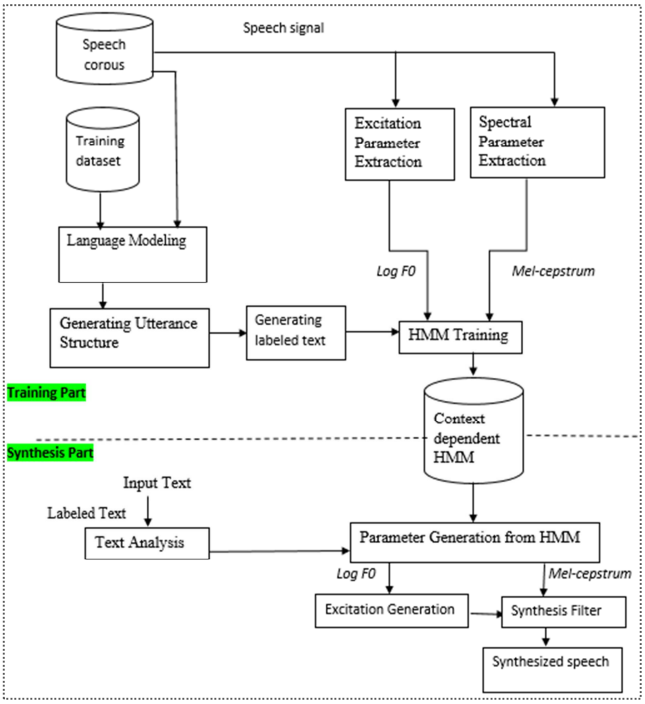

Figure 1. *System architecture of the Afan Oromo Text to speech synthesis using HMM.*

4.1. Training Phase

First from the speech corpus, excitation and spectrum parameters were extracted. For the fundamental frequency training purpose various parameters are used. The parameters are context dependent phoneme, model speaker characteristics and speaking styles. To simulate the temporal structure of speech, hidden markov models use state duration densities. Consequently, Hidden markov model-based text to speech synthesis models fundamental frequency and duration in united framework of hidden markov model [14].

4.2. Spectrum Modeling

By using mel-cepstral coefficients we can control the synthesis filter by hidden markov model. As a result, using the MLSA filter, we may re-synthesize the speech straight from the mel-cepstral coefficients [20]. The observation order of f0 is consisting of 1-Dimensional nonstop values and discrete

symbol which denotes “unvoiced”. So, we cannot apply these two components to f0 pattern modeling [15].

4.3. Decision Tree-based Context Clustering

We have used decision tree-based context clustering for identifying various factors that affect the fundamental frequency, duration and spectrum. Phone identity variables, stress-related factors, and locational factors are among the causes. So, it is impossible to precisely guess the model parameters with incomplete training data. Additionally, preparing the speech corpus which contains all appropriate factors is incredible. For fundamental frequency, duration, and spectrum, a decision-tree based context clustering method is used to solve this problem [20]. All contextual factors must be clustered independently, because each factor have its own significant factor. In this scenario, multivariate Gaussian distributions and multi-space probability distributions are used to simulate f0 and spectrum, which are both parts of the state outcome [21].

4.4. Synthesis Phase

Text analysis, speech parameter generation from models, and synthesized speech generation from the obtained parameters are all part of the synthesis process. Initially, randomly provided text to be converted to speech is changed to a context-based label sequence. Next from the label sequence, a hidden markov model is created by concatenating context dependent hidden markov model [20]. State durations of the hidden markov model sentences are determined and the order of fundamental frequency and mel-cepstral coefficient values with pronounced and unspoken decision is determined in this way the output probability for the hidden markov model is increased using the speech parameter generation algorithm [26]. The concealed markov models statistical characteristics force the speech parameter order created in the synthesis phase to be precise. Finally, the Mel Log Spectrum Approximation filter is used to generate speech waveforms directly from the created fundamental frequency and mel-cepstral coefficients [22].

4.5. Speech Corpora

The data has to be gathered for Afan Oromo language from the text data [17]. The contents for Afan Oromo are gathered from the Afan Oromo educational books, Afan Oromo bible, magazines like kallacha oromiyaa and Bariisa. Then after, the compiled text was recoded in noise free environment with sampling rate of 44.1 kHz mono by male speaker. A speech corpus contains the following components: waveform, transcripted text and labeled file. The Afan Oromo speech corpus, which is suitable for HMM-based text to speech synthesis, was not accessible when the thesis was started. Therefore, it is very essential to prepare a speech corpus that composed of phonetically balanced sentences with equal regular phoneme distribution based on Afan Oromo characteristics.

4.6. Normalization

After gathering the speech data recorded by the professional linguist male speaker, the text data was normalized. Normalization was very important for avoiding vagueness while changing the text data to the order of phonemes [24].

4.7. Segmentation and Labeling

The recorded speech corpus is merged in to sentences. Then after, each recorded speech of every sentence has been partitioned in to phonemes and labeled based on its spectrogram [3]. Representations are provided to every partitioned sound in the sentence. So, this outputs in the creation of label files for every speech waveform file.

5. Results and Discussions

Purposive sampling, a non-probability sampling technique, is employed to construct the training dataset in this study. Five hundred twenty-seven sentences were extracted using this method from a corpus of 10,112 sentences. We used 527 recorded sentences in our thesis. In total, 15,628 phonemes are included in the training dataset.

Table 1. *Distribution of Afan Oromo phonemes in the training dataset.*

Afan Oromo Phonemes	Frequency of occurrence
A	3565
B	457
C	138
Ch	218
D	520
Dh	250
E	897
F	293
G	280
H	479
I	1448
J	152
K	389
L	389
M	570
N	872
Ny	306
O	738
P	26
Ph	28
Q	150
R	619
S	554
Sh	409
T	661
U	828
V	5
W	144
X	40
Y	196
Z	7
Total	15,628

The speeches were recorded in office where minimal noise

was available. Then we recorded a voice corpus using Praat software. The speeches were recorded in mono at 44.1 kHz, and the files were saved in wav format and this file format was changed in to 16 kHz RIFF because the Resource Interchange file format was needed by Festvox system.

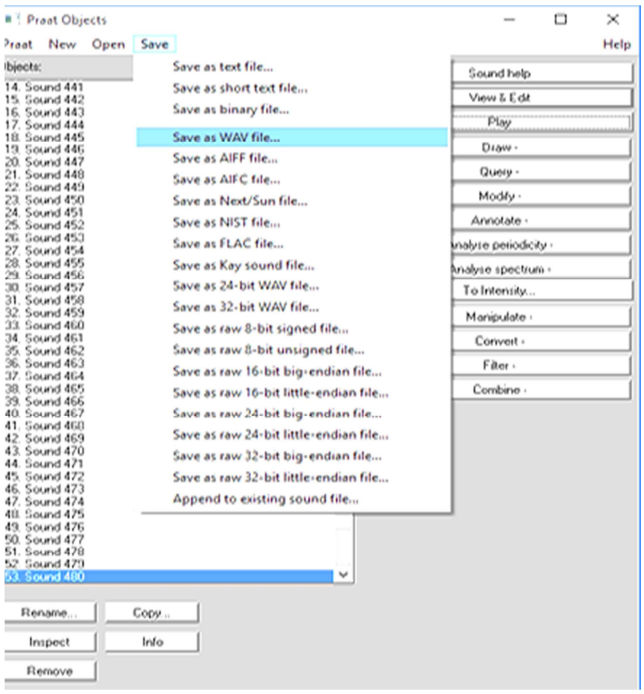

Figure 2. *Recording speech by using Praat.*

For example, the sentence “sangaa diimaan bitame” is labeled as following.

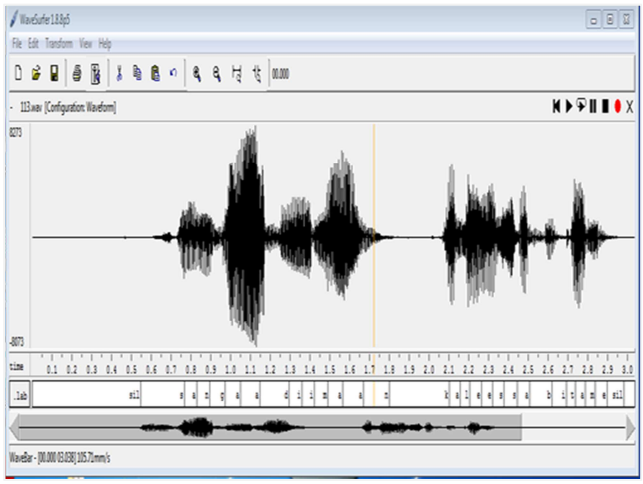

Figure 3. *A speech waves form corresponding to its labels.*

We trained the designed Text to speech synthesis system by using 527 sentences recorded professional linguist male speaker. For evaluating the text to speech synthesis system based on the mean opinion score technique, we select 20 sentences that were not included in the training data. We used evaluators from several backgrounds to test the system. The mean opinion score scales the quality of speech into 5 levels as shown below in the table.

Table 1. *Mean opinion score level.*

Score	Quality	Description of Quality
5	Excellent	No difference with natural
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and Slightly Annoying
2	Poor	Annoying but not objectionable
1	Bad	Very Annoying and objectionable

The system was evaluated by a total of twenty native speakers of the language. There were fourteen males and six females among them.

Table 2. *Average MOS scores of Afan Oromo text to speech synthesis.*

Test data (sentences)	Intelligibility	Naturalness
1	3.857143	3.285714
2	3.809524	3.047619
3	3.761905	2.952381
4	3.714286	3.333333
5	3.857143	3.380952
6	3.761905	3.285714
7	4.047619	3.285714
8	3.857143	3.476190
9	3.904762	3.333333
10	3.952381	3.571429
11	4.047619	3.380952
12	4.047619	3.619048
13	4.095238	3.666667
14	4.047619	3.714286
15	4.095238	3.666667
16	4.333333	3.714286
17	4.380952	3.666667
18	4.476190	3.904762
19	4.285714	3.857143
20	4.476190	3.952381
Average score	4.040476	3.511905

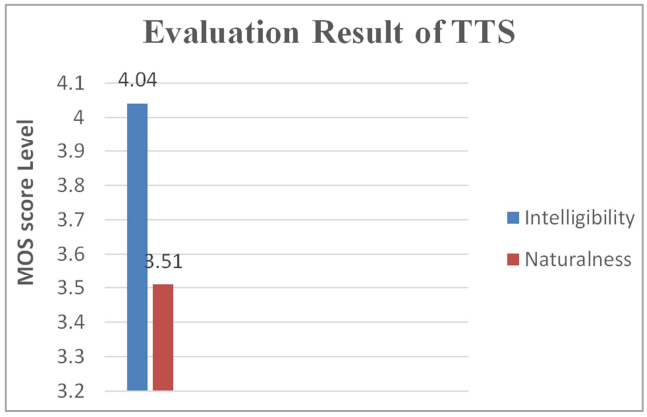

Figure 4. *Diagrammatic view of the MOS TTS.*

The system's overall intelligibility for the twenty Afan Oromo sentences was found to be 4.04, indicating that the synthesizer is 'Good' on the MOS scale. The synthesizer's overall naturalness was determined to be 3.51, indicating that it is 'fair' according to the MOS test scale. The clarity of the voice and the quality of the speech are critical for any text to speech synthesis system.

6. Conclusions

This study demonstrated the Afan Oromo TTS system, which generates speech that is both understandable and natural. The text to speech synthesis system was created using an HMM-based technique. However, the HMM technique is the best technique for the text to speech synthesis that includes opportunity for implementing text to speech synthesis system with minimum recorded speech corpus including data preparation difficulties. The researcher has demonstrated that utilizing HMM, it is possible to construct a text-to-speech synthesis system for Afan Oromo.

In most of application areas, the intelligibility and naturalness of text to speech synthesis system have reached suitable level but, to achieve more natural sounding speech, further effort and improvements are needed in the prosodic, text preprocessing, and pronunciation sectors. The model was trained using the utterance structure created by festival and Festvox, as well as the parameters derived from the raw wave data.

The speech parameters such as fundamental frequency and spectrum are described in hidden markov model training and these parameters are classified by decision tree-based context clustering. Then, context information is produced by text analysis sequence in synthesis part and the system guess hidden markov model order by the decision tree.

Finally, a vocoder uses the speech parameter order, which is constructed based on the estimated models, to synthesize speech. Essentially, the text to be synthesized was presumed to have been transcribed. It indicates that all of the pre-processing is completed before it is sent into the synthesis mechanism. As a result, the trained model generates synthesized speech based on the labeled input text.

The system's performance was evaluated using the Mean Opinion Score method. According to the MOS test, the system's performance in creating understandable speech is good, and the overall outcome of the synthesized speech in terms of naturalness is fair. The outcome appears promising, but it still has to be improved in terms of intelligibility and naturalness.

Author Contributions

Dr. Teklu Urgessa has contributed to conceptualization, methodological framework, data curation and supervision of the entire research process and draft preparation.

Kumera Chala has contributed in experimentation, analysis; validation and original research report writing.

Dr. T. GopiKrishina writing review, editing, visualization and communication with the team members to deliver the final draft research.

Acknowledgements

First and foremost, I give thanks to Almighty God for providing me with grace, love, patience, health, wisdom, and the capacity to walk through all of the challenges and hurdles that have arisen throughout this era of my life, as well as the bravery to complete this task. Next, I'd want to convey my gratitude to my adviser, Dr. Teklu Urgessa, for his leadership and assistance during this study research. I would also thank my co-Advisor Mr. Jabessa Daba who provided me with the necessary assistance during the thesis work. All of the respondents who took the time to fill out the evaluation questionnaire are also to be thanked.

References

[1] Agazi, K., & Tibebe, B. (2012). Unit Selection Based Text-to-Speech Synthesizer for Tigrinya Language. Hilcoe Journal of Computer Science and Technology, 1 (1), 13–21.

[2] Argaw, K., & Sebsibe, H. (2014). Syllabification Design and TTS System for Afan Oromo Using Unit Selection. Hilcoe Journal of Computer Science and Technology, 1.

[3] Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, *568* (7753), 493-498.

[4] Balint, P. (2013). Hidden Markov-model based text-to-speech synthesis (Doctoral dissertation, Budapest University of Technology and Economics).

[5] Baloyi, N. (2012). A text-to-speech synthesis system for Xitsonga using hidden Markov models (Doctoral dissertation, University of Limpopo)

[6] Bansal, D. (2012). Punjabi Speech Synthesis System using HTK. International Journal of Information Sciences and Techniques, 2 (4), 58–70.

[7] Campbell, N. (2005). Developments in corpus-based speech synthesis: Approaching natural conversational speech. *IEICE transactions on information and systems*, *88* (3), 376-383.

[8] Hailemariam, S., Kishore, S. P., Kumar, R., Black, A. W., & Sangal, R. (2004). Unit selection voice for Amharic using festvox. ISCA, 103–107.

[9] Hunt, A. J., & Black, W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In 1996 IEEE International Conference on Acoustics, Speech and Signal Processing Conference Proceedings. 1, 373-376.

[10] Ipsic, S. M.-I. And I. (2006). Croatian HMM-based Speech Synthesis. Journal of Computing and Information Technology, 4 (June 2014), 307–313.

[11] Klatt, D. H. (1987). Review of text to speech conversion for English. The Journal of the Acoustical Society of America, 82 (3), 737-793.

[12] Lieberman, H., & Selker, T. (2000). Out of context: Computer systems that adapt to, and learn from, context. *IBM systems journal*, *39* (3.4), 617-632.

[13] Maninder, S., & Verma, K. G. (2013). Text to speech synthesis for numerals into Punjabi language (Doctoral dissertation).

[14] Panda, S. P., Nayak, A. K., & Rai, S. C. (2020). A survey on speech synthesis techniques in Indian languages. *Multimedia Systems*, *26*, 453-478.

[15] Patil, S. P., & Lahudkar, S. L. (2019). Hidden-Markov-model based statistical parametric speech synthesis for Marathi with optimal number of hidden states. *International Journal of Speech Technology*, *22* (1), 93-98.

[16] Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. Journal of neurolinguistics, 25 (5), 382-407.

[17] Samson, T. (2011). Concatenative Text-To-Speech System for Afan Oromo Language (Doctoral dissertation, Addis Ababa University).

[18] Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

[19] Tewodros, A. (2009). Text-to-speech synthesizer for wolaytta language (Doctoral dissertation, Addis Ababa University).

[20] Tokuda, K., Zen, H., & Black, A. W. (2002, September). An HMM-based speech synthesis system applied to English. In *IEEE Speech Synthesis Workshop* (pp. 227-230).

[21] Tokuda, K., Masuko, T., Miyazaki, N., & Kobayashi, T. (2002). Multi-space probability distribution HMM. *IEICE TRANSACTIONS on Information and Systems*, *85* (3), 455-464.

[22] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000, June). Speech parameter generation algorithms for HMM-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)* (Vol. 3, pp. 1315-1318). IEEE.

[23] Traunmuller, H. (1997). Wolfgang von Kempelen's speaking machine and its successors.

[24] Yasuda, Y., Wang, X., & Yamagishi, J. (2021). Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech Synthesis. Computer Speech & Language, 67, 101183.

[25] Yoshimura, T. (2002). Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems. *PhD diss, Nagoya Institute of Technology*.

[26] Zen, H., & Toda, T. (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005.